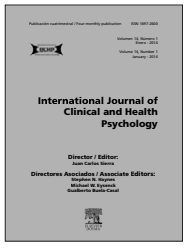
International Journal of Clinical and Health Psychology (2014) **14**, 48–57

International Journal of Clinical and Health Psychology

www.elsevier.es/ijchp



ORIGINAL ARTICLE

Performance evaluation of recent information criteria for selecting multilevel models in Behavioral and Social Sciences

Guillermo Vallejo[a,*], Ellián Tuero-Herrero[a], José Carlos Núñez[a], Pedro Rosário[b]

[a] *Universidad de Oviedo, Spain*
[b] *Universidade do Minho, Portugal*

Received April 13, 2013; accepted July 11, 2013

KEYWORDS
Model selection;
Multilevel models;
Information criteria;
Monte-Carlo study

Abstract This study was designed to find the best strategy for selecting the correct multilevel model among several alternatives taking into account variables such as intraclass correlation, number of groups (m), group size (n), or others as parameter values and intercept-slope covariance. First, we examine this question in a simulation study and second, to illustrate the behavior of the criteria and to explore the generalizability of the findings, a previously published educational dataset is analyzed. The results showed that none of the selection criteria behaved correctly under all the conditions or was consistently better than the others. The intraclass correlation somewhat affects the performance of all selection criteria, but the extent of this influence is relatively minor compared to sample size, parameter values, and correlation between random effects. A large number of groups appears more important than a large number of individuals per group in selecting the best model ($m \geq 50$ and $n \geq 20$ is suggested). Finally, model selection tools such as Akaike's Information Criterion (AIC) or the conditional AIC are recommend when it is assumed that random effects are correlated, whereas use of the Schwarz's Bayesian Information Criterion or the consistent AIC are advantageous for uncorrelated random effects.

PALABRAS CLAVE
Selección de modelos;
Modelos multinivel;
Criterios de información;
Estudio Monte-Carlo

Resumen Se considera el problema de seleccionar el mejor modelo multinivel entre varios modelos candidatos, teniendo en cuenta las variables siguientes: correlación intraclase, número de grupos (m), tamaño del grupo (n), valor de los parámetros y covarianza intercepto-pendiente. Primero se analiza la cuestión reseñada mediante simulación Monte-Carlo, después se utiliza un conjunto de datos previamente publicados para ilustrar el comportamiento de los criterios y explorar su posible generalización. Los resultados mostraron que ningún criterio de selección se comportó correctamente en todas las condiciones, ni fue consistentemente mejor que otro.

*Corresponding author at: Departamento de Psicología, Plaza Feijóo, 33003. Oviedo, Spain.
E-mail address: gvallejo@uniovi.es (G. Vallejo).

También se observó que la correlación intraclase afectaba al rendimiento de los criterios, pero su influencia era más pequeña que la ejercida por el tamaño de muestra, valor de los parámetros y correlación entre los efectos aleatorios. Con respecto al impacto del tamaño de muestra, destacar la importancia de contar con más grupos que participantes dentro del grupo (se sugiere $m \geq 50$ y $n \geq 20$). Finalmente, se recomienda usar el Criterio de Información de Akaike (AIC) o el AIC condicional cuando se asumen efectos aleatorios independientes y el Criterio de Información Bayesiano de Schwarz o el AIC consistente cuando se asumen dependientes.

Longitudinal and hierarchically clustered data are very common in behavioral and social 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 Dettmers, Trautwein, Lüdtke, Kunter, & Baumert, 2010; Imel, Hubbard, Rutter, & Simon, 2013; Núñez, Rosário, Vallejo, & González-Pienda, 2013; Sobral, Villar, Gómez-Fraguela, Romero, & Luengo, 2013). 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 utilize multilevel analysis techniques because they incorporate random effects into the model to accommodate the possible intra-cluster or intra-individual correlation (e.g., Gibbons, Hedeker, & DuToit, 2010).

In fitting multilevel data one is required to choose a set of candidate models, a statistical modeling technique, and a tool to find a working model that provides a closest approximation to the unknown truth than competing alternatives. As noted by several authors (e.g., Hamaker, Van Hattum, Kuiper, & Hoijtink, 2011; Sterba & Peck, 2012), the debate has focused on what should be the proper model selection strategy to compare the adequacy of different models, rather than simply evaluating the fit of a single model in isolation. Thus, before fitting multilevel models, on the basis of well-developed theory, researchers must clearly specify a set of theoretical models that may be appropriate for a given dataset. These ideas are expressed first as verbal hypotheses and then as mathematical equations that specify how the data were generated. A model comparison approach is finally implemented to help evaluate to what extent the data support the selected model and associated hypotheses. Here, it is important to note that the venerable method of null hypothesis testing is like a piece of the overall model-building process.

Rationale for the use of multilevel analysis

In clinical and medical settings, health psychologists often compare different treatment approaches conducted at several clusters (i.e., clinics, hospitals or mental health units), in which both patients and therapists have specific characteristics. For example, patients are enrolled from each clinic and randomly assigned to one of the treatment conditions. In this case, patients are nested within clinics, but clinics are crossed with treatment because patients within each clinic are randomized to each treatment. Another different type of design is one where patients are nested within a clinic, but clinics are randomized to treatments, so that patients from any clinic receive the same treatment. In this design, clinics are nested within treatment but, obviously, cannot be crossed. An additional level can easily be incorporated in the above mentioned two-level designs if patients in each clinic are measured repeatedly across time. Such designs are often referred to as multi-site clinical trial and cluster randomized trials, respectively.

A non-ignorable issue for designs like these is that, in addition to correlation produced by repeated measurements made on different patients is usually inappropriate, patients within the same clinic have similar characteristics, leading to erroneous conclusions when traditional analyses are used. The assumption of independence may be maintained by using group means. However, inferences about individuals based on aggregate data analysis can be biased. Multilevel analysis incorporates both levels in the model so that no choice needs to be made between an individual-level analysis and an aggregate group-level analysis. For this reason, to accommodate the possible clustering effect, hierarchical or multilevel analysis techniques have become the method of choice (Gibbons et al., 2010).

A key aspect of multilevel modeling is to specify a model that includes appropriate random effects, i.e. choice of a particular model within a set of candidate models. Because in many practical applications it is not straightforward to determine the correct multilevel model, different criteria selection procedures currently available in software packages (such as R/Splus, SPSS/PASW, STATA or SAS) are considered for inclusion or exclusion of random effects and to evaluate the goodness of fit of the final model to the data.

Model selection procedures in multilevel analysis

Since various decades ago, null hypothesis significance testing has been the dominant approach to statistical inference. This approach is appropriate for assessing univariate causality and for interpreting data that arise in the context of controlled experiments in which the role of specific hypotheses is well-defined. In non-experimental settings including longitudinal surveys and program evaluation, in contrast, researchers typically utilize significance tests to compare alternative models for observed data or to assess multivariate patterns of causality.

It is this application that is better served by procedures specifically designed for comparison among models, such as model selection criteria, which provide researchers with flexible analytic tools for these types of data (see Burnham, Anderson, & Huyvaert, 2011, for more discussion).

The two most commonly used model selection procedures are likelihood ratio tests (LRTs) and information criteria (IC). Other available tools to such ends (e.g., model averaging, predictive methods and graphical techniques) are used less frequently in the multilevel field. As noted by Johnson and Omland (2004), LRTs are often used hierarchically in a manner analogous to forward selection (backward elimination) in multiple regression, where the analyst starts with an empty (full) model and adds (removes) terms as LRTs indicate a significant improvement in fit. This approach has three primary drawbacks. First, the LRT statistic is typically restricted to comparing pairs of nested models from among the candidate set. Second, in some cases, it can lead to selecting different models depending on the order in which the models are compared. Third, it cannot be used for evaluating the support in the data for each of the models that is examined (e.g., see Hamaker et al., 2011, for details).

To overcome the above limitations, IC-based model selection tools have been recommended, and Akaike's IC (AIC), Hurvich and Tsai's corrected AIC (AICC), Bozdogan's consistent AIC (CAIC), and Schwarz's Bayesian IC (BIC) have been the most commonly used to differentiate between candidate models. The Deviance Information Criterion (DIC) proposed by Spiegelhalter, Best, Carlin, and Van der Linde (2002) is also a method routinely used for Bayesian model comparison. Since Spiegelhalter et al. (2002), different constructions of the DIC have been introduced for selection of models with missing data (e.g., Best, Mason, & Richardson, 2012). However, the appropriate use of the selection criteria in multilevel modeling is a topic of ongoing discussion. Vaida and Blanchard (2005), for instance, pointed out that for analyzing multilevel data, one has to decide whether the substantive questions of interest refer to the clusters (random effects) or to the general population (fixed effects). These authors explicitly elucidated that, when the researchers' focus is on clusters instead of on population, the marginal AIC-type criteria may be unfit, and suggested their conditional counterparts (referred to hereafter as c-AIC). As a consequence, one has to decide on the likelihood (marginal vs. conditional) and correct number of parameters for the penalty term (specification vs. estimation) to use. Several authors provide extensions of the conditional AIC-type criteria in the multilevel field (Greven & Kneib, 2010; Srivastava & Kubokawa, 2010).

Recent studies have extensively evaluated the performance of likelihood-based criteria in the selection of nested and non-nested repeated measures models (e.g., Gurka, 2006; Vallejo, Arnau, Bono, Fernández, & Tuero-Herrero, 2010; Vallejo, Ato, & Valdés, 2008; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011). Performance of the criteria was evaluated under three different scenarios: (a) with respect to their ability to select the correct mean model given a particular covariance structure, (b) with respect to their ability to select the correct covariance structure when the mean model is known, and (c) with respect to their ability to simultaneously select the correct mean and covariance structure. Except for very parsimonious covariance structures and large sample sizes, none of the criteria behaved well in all considered cases. It is also interesting to note that whereas BIC-type criteria performed more accurately than AIC-type criteria in Gurka's (2006) study, they did not perform more accurately than AIC-type criteria or the Hannan-Quinn Criterion (HQC) in Vallejo et al.'s (2008, 2010, 2011) studies.

In addition to the appropriateness of existing likelihood-based model selection criteria, it is natural to ask: should one use Maximum Likelihood (ML) or restricted ML (REML)? It has been argued that REML-based criteria are not appropriate for selecting the fixed effects of the multilevel model, whereas ML-based criteria are appropriate for selecting both fixed and random effects (e.g., Hox, 2010; Verbeke & Molenberghs, 2009). However, Gurka (2006) and Vallejo et al. (2011) found conflicting results in terms of selecting the best multilevel growth curve model, showing that the criteria performed better or equally well under REML estimation compared to ML estimation when choosing the proper mean and covariance structure simultaneously. Thus, more work still needs to be done to understand the role of IC for fitting multilevel models.

Study aim

This paper investigates two issues. First, we examine the question of model selection in a simulation study. Despite the very different theoretical motivations, the goal is the same: to rank models. To our knowledge, there is a lack of evidence that the IC associated with the cluster focus (i.e., c-AIC and DIC) perform well for model selection, as no in-depth numerical study or other additional comparative procedures have been conducted. Here, we are concerned with the c-AIC (Vaida & Blanchard, 2005) and DIC (Spiegelhalter et al., 2002) because they may be obtained using standard statistical packages (e.g., MlwiN, Mplus, SAS, WinBUGS). For purposes of comparison, we also evaluated the behavior of the IC based on the population focus (i.e., AIC, BIC, AICC, CAIC, and HQC). Second, to illustrate the behavior of the criteria and to explore the generalizability of the findings, a previously published dataset is analyzed in the empirical study section.

Method

The article was prepared following the recommendations of Hartley (2012). The causal-comparative design that forms a basis for simulation study is taken from Núñez, Vallejo, Rosário, Tuero-Herrero, and Valle (in press). This study focused on the relationship between contextual variables and students' Biology achievement (*BA*). To contribute to explaining the stated objective, *BA* 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 (*LA*), prior domain knowledge (*PD*), class absence (*CA*), homework completion (*HC*), students' gender

(*SG*), study time (*ST*), and parents' educational level (*PE*). In addition to the teaching approaches (*TA*) *per se*, other explanatory variables included in level-2 were teachers' experience (*TE*), class size (*CS*), and teachers' gender (*TG*).

True data-generating model

In the data-generating process, only the first three explanatory variables at level-1 and the first two explanatory variables at level-2 were included. The model used to simulate the data becomes, at level-1:

$$BA_{ij}=b_{0j}+b_{1j}LA_{ij}+b_2PD_{ij}+b_3CA_{ij}+e_{ij},$$

and at level-2:

$$b_{0j}=\gamma_{00}+\gamma_{01}TA_j+\gamma_{02}TE_j+u_{0j},$$
$$b_{1j}=\gamma_{10}+\gamma_{11}TA_j+\gamma_{12}TE_j+u_{1j},$$

Consistent with common practice in multilevel modeling, we assume that the student-level residuals, e_{1j} have a normal distribution with mean zero and variancei σ^2. We also assume that the class-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}. Level-1 regression coefficients with subscript *j* (i.e., b_{0j} and b_{1j}) are random coefficients that varied across the classes and were treated as dependent variables in the level-2 equations; those without subscript *j* are fixed coefficients. In our example, it is predicted that classes with low intercept (b_{0j}) will have lower academic achievement, on average, than those with high intercept. Similarly, differences in the slope coefficient (b_{1j}) indicate that the relationship between *LA* and *BA* varies randomly from class to class.

Combining the class-level model and the student-level model yields the model with cross-level interactions

$$BA_{ij}=\gamma_{00}+\gamma_{01}TA_j+\gamma_{02}TE_j+\gamma_{10}LA_{ij}+\gamma_{11}LA_{ij}TA_j+\gamma_{12}LA_{ij}TE_j$$
$$+\gamma_{20}PD_{ij}+\gamma_{30}CA_{ij}+u_{0j}+u_{1j}LA_{ij}+e_{1j}, \quad (M_1)$$

which illustrates that the *BA* may be viewed as a function of the overall intercept (γ_{00}), the main effect of teacher's *TA*(γ_{01}), the main effect of teacher's *TE*(γ_{02}), the main effect of student's *LA*(γ_{10}), the main effect of student's *PD*(γ_{20}), the main effect of student's *CA*(γ_{30}), and cross-level interactions involving *TE* with *LA*(γ_{12}), and *TA* with *LA*(γ_{11}), plus a random error: $u_{0j}+u_{1j}LA_{ij}+e_{1j}$. The variable e_{1j} varies over student within a class, however, the variables u_{0j} and u_{1j} are constant for students within classes but vary across classes. The interaction terms appears in the model as consequence of modeling the varying regression slope b_{1j} of student level variable *LA* with the class level variables *TA* and *TE*. Interactions are typically moderators. For example, *TA* and *TE* act as moderator variables for the relationship between *BA* and *LA*.

In order to assess the performance of the different IC in choosing the best model, ten candidate models were fit for each generated dataset. The candidate models were misspecified by incorrectly adding or removing a parameter from the true model (i.e., M_1) described above. For the simple model set (i.e., slope-intercept correlation was set to zero), the nine models were misspecified as follows: (M_2) by dropping $LA_{ij}TA_j$ from the model; (M_3) by dropping $LA_{ij}TE_j$ from the model; (M_4) by dropping u_{ij} from the model; (M_5) by including an interaction between PD_{ij} and CA_{ij}, (M_6) by including an interaction between LA_{ij} and PD_{ij}, (M_7) by including an slope u_{1j}-intercept u_{0j} correlation; (M_8) by including an interaction between LA_j and TE_j, (M_9) by dropping PD_{ij} and including an interaction between LA_{ij} and CA_{ij} (M_{10}) by dropping $LA_{ij}TA_j$ and including an slope-intercept correlation.

Study variables

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

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 (Maas & Hox, 2004). In small size clusters (e.g., therapy groups), however, ICCs above .3 can be found.
2) *Number of groups* (*m*). As multilevel analysis is affected by sample size at the group level, the performance of the criteria was investigated using three different sizes: 30, 60, and 90. For accurate estimates, 100 or more groups would be advisable; however, 50 groups is a frequently occurring number in educational research, and 10 is the smallest required number of clusters (Snijders & Bosker, 2012).
3) *Group size* (*n*). Within each group, we will use sample sizes of 10, 20, and 30, which represent fairly small to moderate to large total sample sizes. The size of the groups is based on the literature and on practice (Maas & Hox, 2004; Núñez et al., in press).
4) *Parameter values*. The regression coefficients are specified as follows: 1 for the intercept, and .5 or 1 for all regression slopes. This represents moderate to large effect sizes.
5) *Intercept-slope covariance*. Because the statistical inference in multilevel modeling has been shown to be sensitive to correlated random effects, slope-intercept correlation was set to 0, .2, and .4.

Information criteria for model selection

In this study, all criteria considered include two basic elements. One term measures the goodness of fit (deviance) of a model, and the other is a penalty for model complexity (Lee & Ghosh, 2009). Below is a brief description of the IC based on the cluster focus (i.e., c-AIC and DIC) that are the object of the present study. The details of the IC based on the above-mentioned population focus are presented in Vallejo et al. (2011), which are summarized in Table 1.

Conditional Akaike's Information Criteria

The conditional AIC is similar in form to the marginal AIC; however, these focuses have different likelihood functions

Table 1 Formulas for commonly used information criteria.

Criteria	ML-estimation	REML-estimation
AIC	$\hat{d}_{\mathrm{ML}}+2\,s$	$\hat{d}_{REML}+2\,s^*$
AICC	$\hat{d}_{\mathrm{ML}}+2\,s[(N)/(N–s–1)]$	$\hat{d}_{REML}+2\,s^*\,[(N–p)/(N–p–s^*–1)]$
BIC_N	$\hat{d}_{\mathrm{ML}}+s\log\,(N)$	$\hat{d}_{REML}+s^*\,\log(N–p)$
BIC_m	$\hat{d}_{\mathrm{ML}}+s\log\,(m)$	$\hat{d}_{REML}+s^*\,log(m)$
$CAIC_N$	$\hat{d}_{\mathrm{ML}}+[s\log\,(N)+1]$	$\hat{d}_{REML}+s^*\,[\log(N–p)+1]$
$CAIC_m$	$\hat{d}_{\mathrm{ML}}+[s\log\,(m)+1]$	$\hat{d}_{REML}+s^*\,[\log(m)+1]$
HQC	$\hat{d}_{\mathrm{ML}}+2\,s\log[\log(m)]$	$\hat{d}_{REML}+s^*\,\log[\log(m)]$

ML = maximum likelihood; REML = restricted maximum likelihood.
Note. $s = p + q$ and $s^* = q$, with p and q representing the dimension of mean and covariance structures; deviance (d) is minus 2 times the log-likelihood function at convergence; N is the total number of observations; m is the total number of clusters.

and a different number of parameters. The c-AIC in "smaller-is-better" form is defined as

$$\mathrm{cAIC} = \hat{d}+2\mathrm{Sc},$$

where the deviance ($\hat{d}$) is minus 2 times the conditional log-likelihood function at convergence, and s_c is the effective number of parameters of the candidate model defined in Vaida and Blanchard (2005). When REML estimation is used, $\hat{d}$ is replaced by the maximized conditional REML log-likelihood. To obtain $\hat{d}$ and s_c, which are needed to compute the c-AIC, we use Proc GLIMMIX and a SAS/IML module that encapsulates the function hatTrace from *lme*R, respectively.

Deviance Information Criterion

The DIC is a generalization of AIC (Table 1) to a Bayesian setting (Spiegelhalter et al., 2002), where s is replaced by the Bayesian equivalent, namely p_D, and the goodness of fit in the first term is replaced by a Bayesian estimate (e.g., posterior mean). The DIC in "smaller-is-better" form is defined as:

$$\mathrm{DIC} = \mathrm{D}(\bar{\mathbf{0}})+2P\mathrm{D},$$

where $\bar{\Phi}=(\gamma',\mathrm{u}',\sigma)'$, $\mathrm{D}(\bar{\Phi})=-2\log L(\mathrm{y}/\bar{\Phi})$ is the deviance of the model evaluated at the means of the posterior distributions of the parameters, and $\mathrm{p}_D=\overline{\mathrm{D}(\Phi)}-\mathrm{D}(\bar{\Phi})$ is the effective number of parameters. SAS Version 9.3 (SAS Institute, 2011) PROC MCMC calculates DIC taking $\overline{\mathrm{D}(\Phi)}$ to be the posterior mean of $-2\log L(\mathrm{y}/\,\bar{\Phi})$, and evaluating $\mathrm{D}(\bar{\Phi})$ as –2 times the log likelihood at the posterior mean of the stochastic nodes. Each model was run for 10,000 iterations, with an additional 5,000 iterations for burn-in. To confirm the convergence of the Markov chains, we used the Geweke diagnostic test. If the chain failed to converge, the model was re-run using the same data and the convergence was re-checked. The convergence of the MCMC chains was generally very good, and less than 10% of the simulations needed to be refitted using more MCMC samples. The number of Markov chain iterations was increased to 50,000.

Procedure

For each previous condition, we generated 1,000 simulated datasets using the RANNOR random number generator in SAS version 9.3, and the number of times that each criterion chose the correct model was recorded. The first-level variance component (i.e., σ^2) was set to 1. The second-level variance components (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 corresponding covariances (i.e., τ_{01}) were set to 0.022, .044, .086, and .172, yielding slope-intercept correlations of 0, .2, and .4, respectively. The fixed values for the observations on the explanatory variables 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.

Results

Simulation study

We first present the percentage of times, averaged across the total sample size, that the correct multilevel model was chosen by the IC when the random effects were assumed to be independent. We then consider results from correlated random effects. In order to conserve space, individual success rates are not tabled but are available from the authors upon request. For comparison, we also considered two variations of the penalty term when computing the consistent BIC and CAIC under ML and REML estimation, respectively. Specifically, the corrections were based on the total sample size ($N = m \bullet n$) as used by SPSS and the total number of clusters (m) as used by SAS.

Uncorrelated random effects

The average percentages of successes are shown in Table 2. They are summarized as follows:

Table 2 Average percentage of correct choices by type of criterion when the random effects were uncorrelated (maximum likelihood-estimation/ restricted maximum likelihood-estimation).

	$AIC_{(SPSS/SAS)}$	$c\text{-}AIC_{(SAS+R)}$	$AICC_{(SPSS/SAS)}$	$BICN_{(SPSS)}$	$BICm_{(SAS)}$	$CAICN_{(SPSS)}$	$CAICm_{(SAS)}$	$HQC_{(SAS)}$	$DIC_{(SAS)}$
PM=0.5/ICC=.1	37/ 69	35/ 46	37/ 69	34/ 74	44/ 75	29/ 73	41/ 75	44/ 75	34
PM=1.0/ICC=.1	53/ 74	49/ 66	54/ 75	72/ 80	73/ 82	69/ 79	74/ 82	72/ 82	49
PM=0.5/ICC=.3	24/ 67	28/ 40	24/ 67	12/ 79	27/ 76	10/ 79	18/ 77	25/ 73	26
PM=1.0/ICC=.3	50/ 76	46/ 65	51/ 77	64/ 89	68/ 86	59/ 84	41/ 75	67/ 85	45

Note. ICC = Intraclass correlation; PM = Parameter magnitude.

1) The performance of likelihood-based selection criteria was much better under REML than under ML estimation. On average, the success rates were 41 and 72% for the AIC, 41 and 55% for the c-AIC, 42 and 72% for the AICC, 46 and 81% for the BIC_N, 51 and 80% for the BIC_m, 42 and 80% for the $CAIC_N$, 47 and 79% for the $CAIC_m$ and 52 and 79% for the HQC under ML and REML, respectively. Interestingly, the DIC only correctly selected the true model in just over 38% of the examined cases.
2) The ability of IC to select the correct model was substantially affected by sample sizes (i.e., *m* and *n*) and parameter magnitude. It must be noted that a large *m* appears more important than a large *n*. With respect to the number of groups (*m*), the average success rate was 45% for $m = 30$, 59% for $m = 60$, and 68% for $m = 90$. With respect to the group size (*n*), the average success rate was 47% for $n = 10$, 62% for $n = 20$, and 64% for $n = 30$. Thus, having larger groups (over 20) does not improve performance very much. It was also easier to distinguish between models in the high parameter magnitude condition than in the low parameter magnitude condition, regardless of the method of estimation used. Still, whereas the average difference between the two magnitudes was about 30 percentage points under ML, it never exceeded 10 percentage points under REML. With respect to the DIC, the average difference was on the order of 16 percentage points. Further, the IC generally performed better when the ICC value was low than when the ICC value was higher. However, under REML estimation, ICC influence was totally irrelevant.
3) The consistent IC (BIC, CAIC, and HQC) outperformed their efficient counterparts (AIC, c-AIC, and AICC), regardless of the manipulated variables. Furthermore, when comparing the consistent IC based on *N* and the consistent IC based on *m*, the latter led to a considerably larger percentage of correct decisions.

Correlated random effects

The pattern of results showed in Table 3 is qualitatively similar for the two levels of slope-intercept correlation manipulated. For this reason, the average percentages of successes are described jointly, and summarized as follows:

1) The likelihood-based IC generally performed better when computed under REML than when computed under ML. On average, the success rates were 47 and 54% for the AIC, 68 and 67% for the c-AIC, 46 and 53% for the AICC, 14 and 20% for the BIC_N, 29 and 37% for the BIC_m, 11 and 16% for the $CAIC_N$, 23 and 30% for the $CAIC_m$, and 39 and 47% for the HQC under ML and REML, respectively. The average success rate for selecting the true model was 39% for DIC.
2) All evaluated selection criteria performed slightly better at the highest level of ICC, and performed substantially better at the highest level of slope-intercept correlation and in the conditions with the larger sample sizes (i.e., *m* and *n*). It was also easier to distinguish among candidate models in the high parameter magnitude condition than in the low parameter magnitude condition. The average difference between the two magnitudes was about 14 percentage points under ML, 6 percentage points under REML, and 4 percentage points under DIC.
3) Contrary to what occurred with level-2 uncorrelated residuals, the efficient IC (AIC, c-AIC, and AICC) outperformed their consistent counterparts (BIC, CAIC, and HQC). Thus, for the efficient IC it is easier to distinguish among competing models when the data-generating model is complex than when the data-generating model is simple, and vice versa for the consistent IC.

Empirical study

In presenting the data-driven selection method, we return to the study conducted by Núñez et al. (in press). As noted in the Method section, the purpose of this study was to determine how contextual and characteristic factors predicted high school students' *BA*. Based on 988 students in 57 classrooms, the true data-generating process will be approximated using the SAS procedures MIXED and MCMC. For consistency with the simulation study, we want to fit the relationship between BA and the first three explanatory variables at level-1 (i.e., LA, PD and CA) and the first two explanatory variables at level-2 (i.e., TA and TE). A SAS program (available from the first author upon request) was used to evaluate the performance of different criteria.

In order to avoid complete enumeration of all possible models, we will use a four-step modeling strategy for selecting the best model by computing IC. In the first step, we formulate a model with all student-level predictors

Table 3 Average percentage of correct choices by type of criterion when the random effects were correlated (maximum likelihood-estimation/ restricted maximum likelihood-estimation).

	AIC	cAIC	AICC	BICN	BICm	CAICN	CAICm	HQC	DIC
PM=0.5/ICC=.1/τu_{01}=.2	26/ 32	55/ 53	26/ 32	03/ 05	11/ 15	02/ 03	07/ 10	19/ 24	20
PM=1.0/ICC=.1/τu_{01}=.2	36/ 35	57/ 58	35/ 35	06/ 06	20/ 18	04/ 04	13/ 13	28/ 29	30
PM=0.5/ICC=.3/τu_{01}=.2	25/ 36	61/ 57	24/ 36	03/ 07	09/ 19	03/ 05	06/ 14	17/ 28	22
PM=1.0/ICC=.3/τu_{01}=.2	39/ 42	71/ 73	38/ 41	08/ 08	21/ 21	05/ 06	15/ 17	32/ 33	29
PM=0.5/ICC=.1/τu_{01}=.4	53/ 63	67/ 63	53/ 63	14/ 27	33/ 47	10/ 21	25/ 40	45/ 57	42
PM=1.0/ICC=.1/τu_{01}=.4	70/ 70	73/ 71	69/ 70	31/ 29	54/ 51	25/ 23	45/ 42	64/ 63	62
PM=0.5/ICC=.3/τu_{01}=.4	49/ 70	77/ 70	48/ 70	10/ 36	27/ 57	07/ 31	19/ 49	39/ 66	46
PM=1.0/ICC=.3/τu_{01}=.4	75/ 80	87/ 86	74/ 79	36/ 43	59/ 64	29/ 36	50/ 57	69/ 74	63

Note. See the note in Table 2. τu_{01} is the u_{0j}-u_{1j} correlation.

fixed. At this step, the intercept is assumed to vary across the classes, but the slopes are held constant. In the second step, we add class-level predictors to the model fit at the student level. The third step assesses whether any of the slopes of any of the student-level predictors has a significant variance component across classes, using the mean structure from the second step. Finally, in the fourth step, we add cross-level interactions between class variables and those student variables that had significant random slopes. In the absence of a strong theory, at each step, we use a data-driven strategy to move toward a simpler structure by dropping predictors or (co)variances that do not appear to be related to the criterion variable. For simplicity, the results presented here include only the last step of the iterative model-building process. For more details of the data-driven strategy from this example, see Núñez et al. (in press, Section multilevel analysis).

To illustrate the performance of the evaluated criteria, a set of candidate models was fit to the data reported by Núñez et al. (in press), including the multilevel model used to simulate the data (M_1). The set of candidate models consisted of ten models each having the same fixed and random effects as defined in the Section true data-generating model. The results obtained are presented in Table 4.

As can be seen, the M_1 is selected by AIC (ML/ REML), c-AIC (ML/ REML), AICC (ML/ REML), BIC_N (REML), BIC_m (ML/ REML), $CAIC_N$ (REML), $CAIC_m$ (REML), and HQC (ML/ REML); while the M_4 is selected by BIC_N (ML), $CAIC_N$ (ML), and $CAIC_m$ (ML). Based on the DIC we conclude that the M_7 is preferred. Further analysis of the models selected by the examined IC facilitates the interpretation process. The results for these three models obtained with the SAS procedures MIXED and MCMC are given in Table 5. Looking over the summary of results for fixed and random effects, one notices that selecting M_1 is the most reasonable course of action. For instance, the result from MCMC for the DIC favor M_7; however, the posterior mean for the slope-intercept covariance (i.e.,τ_{01}), is –0.182, and its 95% credibility interval lies between –1.191 and .352. At $\tau_{01}=0$, M_7 reduces to M_1, the second best model chosen by DIC (Table 4). A similar conclusion can be drawn for the IC that led to selecting the M_4 instead of M_1. Consequently, the superiority of efficient criteria compared with ML-based consistent criteria is consistent with the results obtained in our Monte Carlo simulations.

Finally, we highlight that one aspect of the use of model selection criteria becomes evident from this example. The approach is not restricted to nested models and enables multiple models to be compared simultaneously. Note that while M_4 is nested under both M_2 and M_3, the latter two are not nested. Moreover, competing models can be compared to one another to determine the relative support in the observed data for each model.

Discussion and recommendations

Although illness and health (physical and mental) occur in a social context, past research on their determinants often characterized by individualization (i.e., explain the results of individuals solely in terms of variables related to individual). However, as noted at the beginning of this work, the focus of research has changed substantially, increasingly turning to the analysis of the effects at different levels. In this sense, multilevel analysis has been used to examine the effects of group-level variables and individual-level on the outcomes of individuals. While such analysis has been widely used in education, currently is being used more and more frequently in the medical field, health psychology, social psychology, as well as interdisciplinary areas. This growth was fueled, in part, by the resurgence of interest in the ecological and contextual potential determinants of physical and mental health of individuals. In this sense, the idea that the behavior of individuals can be influenced by its context is key in social sciences and health.

However, after several decades of using this methodology, there are still methodological and applications issues that

Table 4 Values obtained by fitting each of the models in the candidate set to the real data example (maximum likelihood-estimation/restricted maximum likelihood-estimation).

	Criterion								
Model	AIC	c-AIC	AICC	BIC_N	BIC_m	$CAIC_N$	$CAIC_m$	HQC	DIC
M_1	**5009.7**/ **5002.4**	**5090.5**/ **5103.6**	**5009.8**/ **5002.4**	5069.9/ **5018.9**	**5032.2**/ **5008.6**	5080.9/ **5021.9**	5043.2/ **5011.6**	**5018.4**/ **5004.8**	4976.0
M_2	5014.9/ 5010.3	5099.3/ 5112.9	5015.0/ 5010.3	5069.7/ 5026.7	5035.3/ 5016.4	5079.7/ 5029.7	5045.3/ 5019.4	5022.8/ 5012.7	4980.1
M_3	5012.8/ 5007.9	5092.1/ 5104.5	5012.9/ 5008.0	5067.5/ 5024.4	5033.2/ 5014.1	5077.5/ 5027.4	5043.2/ 5017.1	5020.7/ 5010.3	4977.6
M_4	5011.9/ 5018.2	5098.4/ 5113.3	5012.0/ 5018.2	**5066.6**/ 5029.1	5032.3/ 5022.2	**5076.6**/ 5031.1	**5042.3**/ 5024.2	5019.8/ 5019.8	4988.8
M_5	5011.3/ 5007.2	5092.0/ 5110.4	5011.4/ 5007.2	5077.0/ 5023.6	5035.8/ 5013.3	5089.0/ 5026.6	5047.8/ 5016.3	5020.8/ 5009.6	4979.0
M_6	5011.7/ 5003.0	5092.3/ 5106.5	5011.9/ 5003.0	5077.4/ 5019.4	5036.2/ 5009.1	5089.4/ 5022.4	5048.2/ 5012.1	5021.2/ 5005.4	4979.1
M_7	5010.6/ 5002.7	5101.2/ 5114.8	5010.8/ 5002.7	5076.3/ 5024.6	5035.1/ 5010.9	5088.3/ 5028.6	5047.1/ 5014.9	5020.1/ 5005.9	**4975.8**
M_8	5010.4/ 5006.6	5092.5/ 5107.7	5010.5/ 5006.6	5070.6/ 5023.0	5032.8/ 5012.7	5081.6/ 5026.0	5043.8/ 5015.7	5019.1/ 5009.0	4977.0
M_9	5011.5/ 5006.5	5092.2/ 5109.5	5011.7/ 5006.5	5077.2/ 5022.9	5036.0/ 5012.6	5089.2/ 5025.9	5048.0/ 5015.6	5021.0/ 5008.9	4978.0
M_{10}	5012.8/ 5007.4	5108.5/ 5122.2	5012.9/ 5007.5	5073.0/ 5029.3	5035.2/ 5015.6	5084.0/ 5033.3	5046.2/ 5019.6	5021.5/ 5010.6	4976.8

Note. Bold values indicate which of the ten models is preferred by the criterion.

need to be addressed. The main of this study was to provide numerical evidence of the appropriateness of IC in selecting the best multilevel model when using ML/REML and MCMC methods. The study also examines a previously published dataset to illustrate the behavior of the criteria and to explore the generalizability of the findings.

Simulation results showed that none of the criteria behaved correctly under all the conditions nor was any consistently better than the others. We found that if the criteria are rank-ordered by mean success rates, rank order from low to high was DIC (39%), CAIC (42%), BIC (45%), HQC (54%), AICC (54%), AIC (55%), and c-AIC (58%). One question that might be brought to attention from the summarized results is whether or not the computational effort required by criteria associated with the cluster focus (i.e., c-AIC and DIC) justifies the ends. In this study, the basic version of AIC proposed originally by Vaida and Blanchard (2005), which seems to be used in practice (Greven & Kneib, 2010), performed better than its most direct competitors, except for uncorrelated random effects with small sample sizes at the group level. However, the lack of an automated option for computing the c-AIC in the major commercial software packages could be a major obstacle for implementing this criterion in substantive research. The DIC proposed for Bayesian inference by Spiegelhalter et al. (2002) did not perform as well as the remaining criteria examined.

Beyond this, the simulation study covered in this paper revealed that the intraclass correlation somewhat affects the performance of all criteria, but the extent of this influence is relatively minor compared to sample size, parameter values, and correlation between random effects. With regard to the sample size, our results reveal that, in general, a large number of groups appears more important than a large number of individuals per group in selecting the best multilevel model. These results differ to some extent from the numerical results reported by Vallejo et al. (2011) and Wang and Schaalje (2009). They concluded that criteria performed better for larger numbers of subjects and performed much better for designs in which the number of repeated measurements was large. Hence, sample size requirements to distinguish between competing models seem to depend on type of data (i.e., clustered or longitudinal data). For clustered data, one should focus on obtaining more groups than subjects within each group, whereas for longitudinal data, one should focus on obtaining more measurements per subject than on trying to gather more subjects. For clustered longitudinal data, one should perhaps target both issues. To date, this has not been proven definitively.

Over and above that, we also found that the efficient criteria (AIC, c-AIC, and AICC) performed better overall when the random effects were correlated, whereas the consistent criteria (BIC, CAIC, and HQC) seem to be advantageous when the random effects were uncorrelated. Similarly, Vallejo et al. (2010, 2011) note the tendency of AIC-type criteria to perform better than BIC-type criteria when the covariance patterns used to generate the data were more complex. Furthermore, with regard to discrepancies in the formulas involving the penalty term of the criteria, at least for the BIC and CAIC, m is suggested

Table 5 Summary of results from analyses of real data example for three models of interest (standard error in parenthesis and 95%credible intervals in square brackets).

Proc MIXED	M_1		M_4		M_1	
	Estimate(SE)	**Pr>| t|**	**Estimate(SE)**	**Pr>| t|**	**Estimate(SE)**	**Pr>| t|**
Fixed-effects						
γ_{00}(Intercept)	10.553(.477)	<.0001	10.519(.551)	<.0001	10.568(.567)	.0001
γ_{10}(LA)	2.157(.601)	.0008	2.169(.547)	<.0001	2.186(.652)	.0016
γ_{20}(PD)	0.766(.181)	<.0001	0.760(.182)	<.0001	0.746(.183)	<.0001
γ_{30}(CA)	–0.123(.024)	<.0001	–0.126(.024)	<.0001	–0.121(.024)	<.0001
γ_{01}(TA)	0.790(.453)	.0814	0.793(.488)	.1046	0.796(.500)	.1120
γ_{02}(TE)	–0.423(.461)	.3599	–0.336(.489)	.4930	–0.429(.505)	.3952
γ_{11}(LATA)	–1.605(.590)	.0067			–1.671(.640)	.0117
γ_{12} (LA#TE)	1.256(.553)	.0234			1.256(.600)	.0368
	Estimate(SE)	*Pr>Z*	*Estimate(SE)*	*Pr>Z*	*Estimate(SE)*	*Pr>Z*
Random-effects						
τ_{00}(Intercept)	0.712(.289)	.0068	.987(.283)	.0002	1.029 (.493)	.0186
τ_{01}(Inter-slope cov)					–0.461 (.503)	.3594
τ_{11}(Slope)	0.667(.399)	.0476			1.173 (.719)	.0514
σ^2(Residual)	8.471(.398)	<.0001	8.605(.398)	<.0001	8.402 (.399)	<.0001

Proc MCMC	M_1			M_4			M_1		
	Mean	**Posterior interval**		**Mean**	**Posterior interval**		**Mean**	**Posterior interval**	
*Parameter**									
γ_{00}	10.548	[9.482	11.559]	10.501	[9.340	11.617]	10.549	[9.523	11.577]
γ_{10}	2.166	[1.010	3.349]	2.177	[1.102	3.284]	2.176	[1.050	3.322]
γ_{20}	0.767	[0.410	1.143]	0.709	[0.385	1.130]	0.752	[0.406	1.102]
γ_{30}	–0.123	[–0.171	–0.076]	–0.126	[–0.175	–0.077]	–0.122	[–0.171	–0.074]
γ_{01}	0.788	[–0.156	1.750]	0.816	[–0.157	1.806]	0.777	[–0.153	1.705]
γ_{02}	–0.406	[–1.319	0.517]	0.338	[–1.339	0.666]	–0.389	[–1.343	0.567]
γ_{11}	–1.594	[–2.769	–0.459]				–1.578	[–2.742	–0.430]
γ_{12}	1.224	[0.160	2.325]				1.227	[0.111	2.305]
τ_{00}	0.880	[0.308	1.676]	1.138	[0.616	1.917]	0.819	[0.312	2.001]
τ_{01}							–0.182	[–1.191	0.352]
τ_{11}	0.666	[0.020	1.748]				0.914	[0.170	2.276]
σ^2	8.543	[7.787	9.385]	8.667	[7.915	9.500]	8.543	[7.784	9.358]

* Based on assuming uninformative priors.

in the correction rather than *N*. As indicated above, sample size in SAS when computing the BIC and CAIC is equal to *m,* whereas sample size in SPSS is equal to *N* under ML and REML, respectively. It should also be noted that, despite having been argued that REML-based information criteria are not appropriate for selection of fixed effects of the multilevel model, in many cases, performance of the criteria was better using REML rather than ML estimation. Again, this result is consistent with the findings of Gurka (2006) and Vallejo et al. (2011).

Finally, we should like to add four brief comments. First and foremost, the current study reinforces the importance of explicitly considering the sample sizes for designing multilevel studies. Researchers interested in carrying out studies that have sufficient power to detect the model closest to the true data generating process should avoid using small sample sizes whenever possible. The results of this simulation study clearly indicate that under REML estimation the consistent criteria (BIC, CAIC, and HQC) selecting the correct model around 83% of the time for moderate sample sizes (using $m = 60$ and $n = 20$) and uncorrelated random effects, while their efficient counterparts (AIC, AICC, and c-AIC) selecting the proper model over 78%of the time for correlated random effects. Thus, in order to reach a rate of correct model selection around 80%, the rule of thumb $m \geq 50$ and $N/m \geq 20$ per group is suggested. Second, for random effects assumed not to be correlated, which is generally unlikely, we recommend using either of the consistent criteria; whereas for the correlated random effects, we recommend using either of the efficient criteria. In addition, in the calculation of BIC and CAIC we recommend using *m* in combination with REML estimation. Third, researchers should be cautioned that the DIC performs less accurately than the

remaining criteria. And fourth, of course, the results are limited to the conditions examined in our study, though we sense that they may be generalizable to a wide variety of commonly encountered situations.

Funding

We gratefully thank the Editor and the anonymous reviewers for the constructive comments that led to substantial improvements in the manuscript. This paper was prepared with support from the Spanish Ministry of Science and Innovation (Ref: PSI-2011-23395 & EDU-2010-16231).

References

Best, N., Mason, A., & Richardson, S. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, *7*, 109-146.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*, 23-35.

Dettmers, S., Trautwein, U., Lüdtke, O., Kunter, M., & Baumert, J. (2010). Homework works if homework quality is high: Using multilevel modeling to predict the development of achievement in mathematics. *Journal of Educational Psychology, 102*, 467-482.

Gibbons, R.D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, *6*, 79-107.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 1-17.

Gurka, M.J. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60*, 19-26.

Hamaker, E. L., Van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. J. Hox, & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 231-255). New York: Taylor & Francis.

Hartley, J. (2012). New ways of making academic articles easier to read. *International Journal of Clinical and Health Psychology, 12,* 143-160.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications (2th. 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.

Johnson, J.B., & Omland, K.S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution*, *19*, 101-108.

Lee, H., & Ghosh, S. K. (2009). Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation*, *79*, 93-106.

Maas, C. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427-440.

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*. DOI:http://dx.doi.org/10.1387/RevPsicodidact.7127

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2th. ed.)*. Thousand Oaks, C.A.: Sage.

Sobral, J., Villar, P., Gómez-Fraguela, J. A., Romero, E., & Luengo, M. A. (2013). Interactive effects of personality and separation as acculturation style on adolescent antisocial behaviour. *International Journal of Clinical and Health Psychology, 13,* 25-31

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 Series B*, *64*, 583-640.

Srivastava M. S., & Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, *101*, 1970-1980.

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods*, *17*, 582-599.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*, 351-370.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data (3th. ed.)*. 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., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology*: *Journal of Research Methods for the Behavioral and Social Sciences, 4,* 10-21.

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

Wang, J., & Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics - Simulation and Computation, 38*, 788-801.