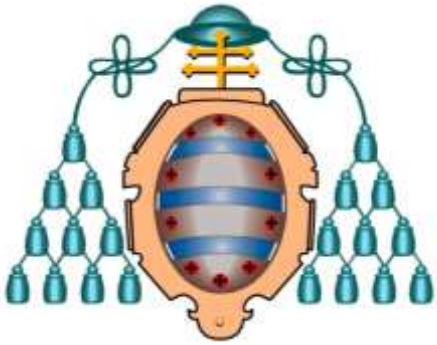
Universidad de Oviedo



Departamento de Psicología

Programa de Doctorado en Psicología

Tesis Doctoral

Selección de Modelos Multinivel en la Investigación en el Campo de la Educación

Ellián Tuero Herrero

La presente Tesis Doctoral ha sido realizada gracias al apoyo económico de las siguientes ayudas concedidas a *Doña Ellián Tuero Herrero* durante su Formación Predoctoral:

- Beca *"Programa Severo Ochoa"* concedida por **FICYT** (Fundación para el Fomento en Asturias de la Investigación Científica Aplicada y la Tecnología) con Referencia BP09-006;
- Beca *"Programa FPI"* concedida por **MICCIN** (Ministerio de Ciencia e Innovación para el Desarrollo de la Formación del Personal Investigador) con Referencia BES-2012-056754 asociada al Proyecto PSI2011-23395.

A Pablo

Agradecimientos

Este trabajo es el producto de un largo proceso de investigación en el que han colaborado muchas personas. Ellas no aparecen en la portada de este documento, pero bien se merecían un hueco por lo importantes que han sido para mí en este tiempo. La culminación de esta tesis no habría sido posible sin ellas. Este espacio de agradecimiento intenta mostrarles mi gratitud por ello.

- Quisiera comenzar por la persona que primero me brindó la oportunidad de trabajar codo con codo en el complejísimo mundo de la Universidad. Mi más sincero agradecimiento al *Dr. Guillermo Vallejo*, por la inmejorable dirección de esta tesis. No tengo palabras para expresar lo muchísimo que le debo... Su dedicación plena, su exigencia y su rigor, han constituido para mí la mejor escuela en la que instruirme para investigar. Su incondicional apoyo en los progresos y sus alentadores y sabios consejos en los difíciles momentos han sido los impulsos que me han ayudado a continuar. Gracias Guillermo por haber depositado, desde el principio, la confianza en mí.

- De igual modo quisiera agradecer al *Dr. José Carlos Núñez* su calurosa acogida en su grupo investigador. Su ayuda, su orientación y todas sus enseñanzas en el ámbito de la educación y en la vida en general, han sido para mí capitales en mi formación. Ha sido todo un lujo haberle tenido como mentor a lo largo de estos años. Gracias Carlos por ser un brillante docente, un excelente investigador y mejor persona.

- Igualmente quisiera hacer un reconocimiento especial a la *Dra. Paula Fernández* porque siempre tuvo una palabra sabia para hacerme entender que *"si se puede"*. Le estaré siempre inmensamente agradecida por sus vitales lecciones. Gracias por nuestros retos superados y nuestros sueños forjados. Ha sido un verdadero placer trabajar a su lado y espero siga siéndolo por mucho tiempo.

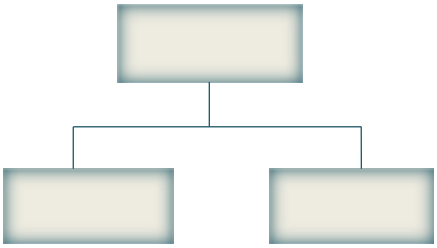
- Agradecer al *Dr. Pedro Rosário* su amable recibimiento durante mi formación en Portugal. Su disponibilidad y su tolerancia siempre a las preguntas ha fomentado mi inquietud por la mejora de los trabajos con calidad. Gracias por haberme

guiado y aconsejado sobre cómo realizar excelente investigación educativa en centros escolares.

- Quienes se merecen muchas y buenas palabras son todas las personas que forman el equipo de investigación con el que he tenido el gusto de trabajar estos años. A los *Doctores: Rebeca Cerezo*, *Celestino Rodríguez*, *David Álvarez* y especialmente a la doctora *Alejandra Dobarro*. A los *becarios: Natalia, Miguel*, *Marisol* y *Trinidad*. Mencionaré aquí con especial cariño a *Estrella*. A ella me unen más que lazos académicos, y con sus generosos consejos me he sentido acompañada todo el periodo de mi doctorado. Todos ellos son para mí un claro ejemplo de trabajo y profesionalidad. ¡Gracias por todo lo que me habéis enseñado y ayudado!

- Contar con el apoyo de mis *compañeros doctorandos* me ha dado fuerzas y ánimos. Gracias por ser el motor de mi motivación y descubrirme el apasionante mundo de la docencia y de la investigación.

- Finalmente, quisiera dar las gracias a mi *familia*. A mis padres *Magdalena* y *Alejandro*, por haberme ayudado siempre en los momentos más difíciles y estar conmigo en los felices. Vuestra confianza en mí, vuestro respeto y comprensión, han sido los ingredientes clave para ir hacia adelante. Gracias por haberme dado todas las oportunidades que me han permitido llegar hasta aquí. Gracias porque todo lo dais y todo lo merecéis. A *Pablo* el horizonte de mi ilusión. Gracias por hacer que lo difícil sea fácil, y lo fácil interesante. Agradezco que seas parte de mí, y parte de este trabajo.



Cuando descubres que las jerarquías existen,

tiendes a verlas por todas partes...

Kreft, De Leeuw & Kim

Índice

LISTA DE TRABAJOS ORIGINALES

Esta Tesis presentada para obtener el grado de Doctor de la Universidad de Oviedo, es el resultado de cuatro estudios llevados a cabo durante cuatro años en el Departamento de Psicología de la Universidad de Oviedo, en las áreas de Metodología de las Ciencias del Comportamiento y de Psicología Evolutiva y de la Educación. Durante dicha etapa obtuve el Diploma de Estudios Avanzados (DEA) en el Programa de Doctorado en Psicología de la Universidad de Oviedo. Los artículos que se muestran en esta Tesis han sido publicados en Revistas Científicas de reconocido prestigio.

Artículo I:

Vallejo, G., Arnau, J., Bono, R., Fernández, P., & Tuero-Herrero, E. (2010). Nested Model Selection for Longitudinal Data Using Information Criteria and the Conditional Adjustment Strategy. *Psicothema, 22*(2), 323-333.

Artículo II:

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero-Herrero, E. (2011). Selecting the Best Unbalanced Repeated Measures Model. *Behavior Research Methods, 43*(3), 18-36. doi: 10.3758/s13428-010-0040-1.

Artículo III:

Vallejo, G., Tuero-Herrero, E., Núñez, J. C., & Rosário, P. (en prensa). Performance Evaluation of Recent Information Criteria for Selecting Multilevel Models in Behavioral and Social Sciences. *International Journal of Clinical and Health Psychology.*

Artículo IV:

Núñez, J. C., Vallejo, G., Rosário, P., Tuero-Herrero, E., & Valle, A. (en prensa). Student, Teacher, and School Context Variables Predicting Academic Achievement in Biology: Analysis from A Multilevel Perspective. *Journal of Psychodidactics*. doi: 10.1387/RevPsicodidact.7127.

RESUMEN

Actualmente, el modelado lineal jerárquico o multinivel constituye un área de investigación muy activa en diversas áreas de las Ciencias Sociales, del Comportamiento y de la Salud, debido a que estas técnicas permiten abordar cuestiones cuyo análisis estadístico resulta problemático con los modelos tradicionales. En el ámbito de la Psicología de la Educación, este enfoque analítico permite la construcción empírica de modelos ajustados a los datos, que hacen posible describir, explicar, predecir y poner a prueba hipótesis acerca de la relación entre los estudiantes y el medio escolar en el que operan. Esta circunstancia enfrenta al investigador al reto de considerar la estructura jerárquica, multinivel o anidada en que, bien de modo natural o bien como consecuencia del diseño de estudio, están organizados los datos, y a medir variables en los distintos niveles de agregación de las unidades de estudio (Vallejo, et al., 2008). Así las cosas, y en vista de la necesidad de obtener modelos con gran capacidad de descripción, explicación y predicción, la investigación metodológica se centra ahora en el desarrollo de métodos de estimación y herramientas de evaluación para la selección de los modelos más parsimoniosos adecuados a cada situación (Ojeda & Velasco, 2012). Realizar la selección del modelo óptimo resulta decisivo para interpretar adecuadamente los datos. Así pues, es crucial para un investigador en el campo educativo resolver las siguientes cuestiones: *¿Cuál es el mejor modelo de todas las alternativas formuladas?* *¿Tiene sentido seleccionar un modelo en función del uso posterior que se vaya a dar al mismo?,* y es que cuando para una misma evidencia muestral existen múltiples modelos candidatos, surge la problemática de la selección (Vallejo et al., 2010, 2011b).

Recientemente se han realizado numerosas investigaciones cuyo objetivo principal ha sido localizar el mejor modelo multinivel bajo distintos escenarios. Por ejemplo, los trabajos de Gurka (2006) y Wang y Schaalje (2009) se centraron en seleccionar el mejor modelo de medias, dada una particular estructura de varianzas y covarianzas; los estudios de Ferron, Dailey y Yi (2002), y Vallejo, Ato y Valdés (2008) mostraron la capacidad de seleccionar el modelo correcto de covarianza cuando la estructura de medias del modelo era conocida; y la capacidad de seleccionar simultáneamente la correcta estructura de medias y de covarianza en los modelos, también fue examinada (Gurka, 2006).

Sin embargo, a pesar de la gran variedad de estudios existentes basados en la estrategia de comparación de modelos en distintos escenarios y con diferentes enfoques para evaluar la bondad de ajuste del modelo final a los datos, actualmente no hay un consenso sobre lo que constituye la herramienta más adecuada para elegir el mejor modelo multinivel. No obstante, en términos generales se puede expresar que todas las investigaciones convergen en que, independientemente del criterio de selección utilizado, la elección del modelo óptimo prospera cuando el tamaño de muestra aumenta en los diseños transversales, cuando aumentan el tamaño de muestra y el número de medidas repetidas en los diseños longitudinales, y en ambos escenarios, cuando la complejidad del modelo disminuye. Por todo ello es necesario el estudio pormenorizado de la fase de ajuste del proceso de modelado estadístico multinivel y de los criterios existentes para dicho ajuste. En este sentido, el objetivo central de esta Tesis es evaluar el desempeño de una amplia variedad de estrategias de bondad de ajuste para elegir el modelo que mejor

se aproxima al verdadero proceso generador de los datos en distintas condiciones de estudio.

A tal fin se llevan a cabo 4 investigaciones. En tres de ellas se pone a prueba el rendimiento de diversos Criterios de Información implementados en el módulo PROC MIXED del programa SAS. Dos de estas investigaciones yacen sobre diseños longitudinales y dos sobre diseños transversales. Brevemente, y en ese mismo orden, en el primer estudio se compara el desempeño del test de ajuste condicional LRT y varias versiones de los Criterios de Información (AIC, AICC, BIC, CAIC, y HQIC) para seleccionar estructuras de medias y de covarianzas anidadas, asumiendo conocido el verdadero proceso generador de los datos. En el segundo trabajo, se examina el rendimiento de los Criterios de Información (AIC, AICC, BIC, CAIC, y HQIC) para seleccionar simultáneamente estructuras de medias y de covarianzas no anidadas, cuando se incumplen los supuestos distribucionales del modelo y existe un alto porcentaje de datos faltantes en los diseños estudiados. La tercera investigación consiste en encontrar la mejor estrategia (AIC, AICC, CAIC, HQIC, cAIC -AIC condicional- y DIC -Criterio de Información de la Desvianza Bayesiano) para seleccionar el modelo multinivel que mejor se aproxima al proceso generador de los datos en un escenario donde se manipulan el número de grupos, el tamaño de los grupos y la correlación intra-clase, entre otras variables. Por último, en el cuarto trabajo se realiza un estudio cross sectional expost facto en el Campo de la Educación, cuyos resultados son analizados desde la propia perspectiva multinivel. Los resultados encontrados, publicados en 4 artículos científicos, se resumen de la manera que a continuación sigue.

En el primer artículo (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2010), se pone de manifiesto que el desempeño del criterio de ajuste condicional LRT basado en el estimador de máxima verosimilitud completa MV es superior al resto de criterios examinados. Los Criterios de Información Eficientes (AIC y AICC) funcionan mejor cuando los patrones de covarianza estudiados son complejos y peor cuando son simples. Al contrario sucede para los Criterios de Información Consistentes (BIC, CAIC y HQIC), que rinden mejor cuando los patrones de covarianza son simples y peor cuando son complejos.

En el segundo artículo (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011b), se corrobora lo encontrado en el estudio previo, esto es, que para seleccionar el modelo correcto los Criterios de Información Eficientes se comportan mejor cuando los patrones de covarianza utilizados para generar los datos son más complejos y viceversa para los Criterios de Información Consistentes. A su vez, el desempeño de éstos últimos (BIC, CAIC y HQIC) es mayor cuando se basan en el número total de individuos (nivel 2) que en el número total de observaciones (nivel 1). Los resultados también indican que dado un conjunto de datos con valores perdidos, los Criterios de Información Eficientes (AIC y AICC) se ven más afectados que los Criterios de Información Consistentes a la falta de normalidad.

En el tercer artículo (Vallejo, Tuero-Herrero, Núñez, & Rosário, en prensa), ninguno de los criterios de selección (marginales-condicionales y clásicos-bayesianos) se comporta adecuadamente en todas las condiciones ni es consistentemente mejor que los otros. Con respecto al tamaño de muestra a la hora de seleccionar el mejor modelo, es más importante un gran número de grupos (NG) que un gran tamaño de grupos (TG) (sugiriéndose $NG \geq 50$ y $N/NG \geq 20$, siendo $N = NG \times TG$). La correlación intra-clase afecta al rendimiento de los criterios de selección, pero la magnitud de esa influencia es relativamente menor en comparación con los valores de los parámetros y con la correlación de los efectos aleatorios. Al igual que en las investigaciones anteriores, los Criterios de Información Eficientes funcionan mejor para los efectos aleatorios

correlacionados (patrones de covarianza más complejos) y los Criterios de Información Consistentes para los efectos no correlacionados (patrones de covarianza más simples). Globalmente, el criterio que tiene un mejor desempeño es el AIC condicional (cAIC) seguido del AIC marginal (AIC).

El cuarto artículo (Núñez, Vallejo, Rosário, Tuero-Herrero, & Valle, en prensa), constituye una aplicación de análisis multinivel a un estudio de campo. Los resultados ponen en relieve la importancia de la interacción entre cómo enseñan los profesores (variable del nivel profesor, nivel 2), cómo aprenden los estudiantes, y el rendimiento académico obtenido (variables del nivel estudiante, nivel 1). También se constata que la mayor parte de la variabilidad en el rendimiento de la asignatura estudiada (Biología) está asociada con variables tomadas a nivel estudiante (el 85,6%), mientras que las variables tomadas a nivel de clase únicamente aportan un 14,4% de la misma. Uno de los aportes más relevantes de este último trabajo es la aplicación del conocimiento que ofrecen los Modelos Multinivel a la evaluación de los estudios en el Campo de la Educación. Se aporta conocimiento de esta técnica de análisis en dos sentidos. De una parte, se conceptualiza la técnica describiéndose de manera rigurosa y detallada. Y de otra, se explica el proceso de modelado estadístico multinivel de forma precisa. Ambas partes se revelan como una contribución pedagógica indispensable para las exigencias actuales en la investigación educativa.

Finalmente los resultados obtenidos en estos estudios se discuten planteándose nuevas líneas de trabajo que no se encuentran muy alejadas de las investigaciones que aquí se presentan.

La realización de los estudios que conforman esta Tesis ha sido posible gracias a las directrices, comentarios y sugerencias constructivas de los doctores *Guillermo Vallejo Seco* y *José Carlos Núñez Pérez*, así como al aporte económico de las becas *"Severo Ochoa"* concedida por FICYT con Referencia BP09-006 y a la *"FPI"* concedida por MICCIN con Referencia BES-2012-056754 asociada al Proyecto PSI2011-23395.

ABSTRACT

Linear, hierarchical, or multilevel modeling is currently a very active research area in the diverse spheres of Social Sciences, Behavior Sciences, and Health because these techniques allow one to address issues that are very problematic to analyze statistically with the traditional models. In the sphere of Educational Psychology, this analytical approach allows the empirical construction of models fitted to the data that enable the description, explanation, prediction, and testing of hypotheses about the relationship between students and the school setting in which they perform. This circumstance brings researchers face to face with the challenge of the hierarchical, multilevel, or nested structure in which the data are organized, either naturally or as a consequence of the study design, and to measure variables at the different levels of aggregated units of study (Vallejo, et al., 2008). Thus, and in view of the need to achieve models with a great capacity to describe, explain, and predict, methodological research is now focusing on the development of estimation methods and assessment tools to select the most parsimonious models that are appropriate for each situation (Ojeda & Velasco, 2012). Selecting the best model is decisive to be able to interpret the data adequately. Thus, it is crucial for a researcher in the educational field to answer the following questions: *Out of all the alternatives formulated, which is the best model? Does it make sense to select a model depending on the subsequent use that will be made of it?,* because when there are multiple candidate models for the same sample evidence, the problem of selection emerges (Vallejo et al., 2010, 2011b).

Numerous investigations have been carried out with the main goal of locating the best multilevel model in different scenarios. For example, the works of Gurka (2006) and Wang and Schaalje (2009) were focused on selecting the best measurement model, given a particular variance-covariance structure; the studies of Ferron, Dailey, and Yi (2002), and Vallejo, Ato, and Valdés (2008) showed the capacity to select the correct covariance model when the mean structure of the model was known; the capacity to simultaneously select the correct mean and covariance structures in the models was also examined (Gurka, 2006).

However, in spite of the great variety of studies based on the strategy of comparing models in different scenarios and with different approaches to assess the goodness of fit of the final model to the data, there is currently no consensus about which is the most adequate tool to choose the best multilevel model. Nevertheless, in general terms, it can be stated that all the research coincides in that, independently of the selection criterion employed, the choice of the best model prospers when sample size increases in cross-sectional designs, when sample size and the number of repeated measures increase in longitudinal designs, and in both scenarios, when the complexity of the model decreases. Hence, it is necessary to study in detail the phase of fitting the process of multilevel statistic modeling and the existing criteria for that adjustment. In this sense, the central goal of this Thesis is to assess the performance of a broad variety of goodness-of-fit strategies in order to choose the model that is closest to the true process of data generation in different study conditions.

For this purpose, four investigations were carried out. In three of them, the performances of various Information Criteria implemented in the PROC MIXED module of the SAS program were tested. Two of these investigations are based on longitudinal designs and two on cross-sectional designs. Briefly, and in the same order, the first study compares the performance of the LRT conditional fit test and various versions of Information Criteria (AIC, AICC, BIC, CAIC, and HQIC) in the selection of nested mean and covariance structures, assuming that the true process of data generation is known. The second work examines the performance of the Information Criteria (AIC, AICC, BIC, CAIC, and HQIC) in the simultaneous selection of nonnested mean and covariance structures when the distributional assumptions of the model are not met and there is a high percentage of missing data in the designs studied. The third investigation seeks the best strategy (AIC, AICC, CAIC, HQIC, cAIC —-conditional AIC— and DIC —Bayesian Deviance Information Criterion) to select the multilevel model that is closest to the data generating process in a scenario in which the number of groups, the size of the groups, and the intra-class correlation, among other variables, are manipulated. Lastly, in the fourth work, a cross-sectional expost facto study is carried out in the field of Education, the results of which are analyzed from a multilevel perspective. The results found, published in four scientific articles, are summarized as follows.

The first article (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2010) shows that the performance of the LRT conditional fit criterion based on the full maximum likelihood estimator MV is superior to the rest of the criteria examined. The Efficient Information Criteria (AIC and AICC) perform better when the covariance patterns studied are complex, and worse when they are simple. The opposite occurs for the Consistent information Criteria (BIC, CAIC and HQIC), which perform better when the covariance patterns are simple, and worse when they are complex.

In the second article (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011b), the findings of the previous study are corroborated; that is, in the selection of the correct model, the Efficient Information Criteria perform better when the covariance patterns used to generate the data are more complex, and vice versa for the Consistent Information Criteria. In turn, the performance of the latter (BIC, CAIC and HQIC) is superior when they are based on the total number of individuals (level 2) rather than on the total number of observations (level 1). The results also indicate that, given a set of data with missing values, the Efficient Information Criteria (AIC and AICC) are more affected by the lack of normality than the Consistent Information Criteria.

In the third article (Vallejo, Tuero-Herrero, Núñez, & Rosário, in press), none of the selection criteria (marginal-conditional and classic-Bayesian) perform adequately in all conditions and none of them is consistently better than the others. With regard to sample size when selecting the best model, a large number of groups (NG) is more important than the large size of groups (SG) (suggesting $NG \geq 50$ and $N/NG \geq 20$, with $N = NG \times SG$). The intra-class correlation affects the performance of the selection criteria, but the magnitude of this influence is relatively low in comparison with the values of the parameters and with the correlation of the random effects. As in the previous investigations, the Efficient Information Criteria perform better for correlated random effects (more complex covariance patterns), and the Consistent Information Criteria perform better for the noncorrelated effects (simpler covariance patterns). Globally, the conditional AIC (cAIC) is the criterion that performs the best, followed by the marginal AIC (AIC).

The fourth article (Núñez, Vallejo, Rosário, Tuero-Herrero, & Valle, in press) involves the application of multilevel analysis to a field study. The results show the importance of the interaction between the way teachers teach (teacher level variable, level 2), the way students learn, and the academic achievement obtained (student level

variables, level 1). The greater part of the variability in performance of the subject matter studied (Biology) is seen to be associated with variables from the student level (85.6%), whereas the variables from class level only contribute 14.4% of the variability. One of the most relevant contributions of this last work is the application of the knowledge provided by Multilevel Models to the assessment of studies in the field of Education. Knowledge of this analysis technique is provided in two ways. On the one hand, the technique is conceptualized, described rigorously and in detail. On the other hand, the process of multilevel statistical modeling is explained precisely. Both parts are revealed as an essential pedagogical contribution to the current demands in educational research.

Lastly, the results obtained in these studies are discussed, proposing new lines of work that are not very different from the investigations presented herein.

Planificación de la Investigación

Todo investigador debe
pensar antes de realizar su estudio

ESTADO DEL ARTE DE LA INVESTIGACIÓN EN EDUCACIÓN

1.1. La Investigación de la Eficacia Escolar

La meta de cualquier investigador que se precie es emitir un informe que genere múltiples líneas de investigación tanto teóricas como prácticas, como metodológicas. El Informe Coleman (Coleman et al., 1966), del que se cumplen ya 47 años este año 2013 puede presumir de ello. Dicho informe, encargado por el gobierno norteamericano a una comisión de expertos, tuvo como finalidad conocer la importancia de las variables escolares sobre los resultados de los estudiantes. A partir de los datos de medio millón de alumnos, el informe mostró que las variables de la escuela no explicaban más allá del 10% de la varianza en los resultados de los estudiantes, siendo éstos determinados principalmente por el origen socio-cultural de los alumnos (Fernández & González, 1997). A partir de este estudio, se desató una enorme polémica sobre el peso que ejercían los factores escolares en la calidad de la educación. Arrancan así, desde ese momento, miles de investigaciones en todo el mundo agrupadas bajo una misma corriente: la investigación sobre la Eficacia Escolar (School Effectiveness). Este movimiento teórico-práctico ha llegado a convertirse hoy por hoy, sin duda alguna, en la línea de investigación pedagógica que más influencia ha tenido en la mejora de la educación (Townsend, 2007).

La denominada Eficacia Escolar, tiene sus orígenes principalmente en el Reino Unido y en Estados Unidos, donde reconocidos estudios realizados durante los años sesenta y setenta pusieron de relieve la escasa influencia de la escuela sobre los resultados educativos, en comparación con otros aspectos de los estudiantes, tales como las aptitudes, la etnia y el estatus socio-económico (Coleman et al., 1966; Jencks et al., 1972). Estos estudios adolecían de un buen número de limitaciones metodológicas y la investigación posterior intentó dar realce a la existencia de efectos escolares significativos, aun reconociendo la gran influencia del contexto socio-económico y cultural de los estudiantes (Edmonds, 1979; Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988). En este

sentido, se intentan dar respuestas a los siguientes interrogantes: *¿Qué determina la efectividad de la escuela? ¿Pueden las escuelas ser efectivas?*

Estas investigaciones tienen, de este modo, un interés creciente por elaborar una teoría comprensiva capaz de integrar todos aquellos elementos que ayuden a que un centro escolar sea eficaz (Townsend, 2007). De esta forma, presentan dos objetivos prioritarios, por un lado la estimación de la magnitud de los efectos escolares y el análisis de sus propiedades científicas; y por otro, el estudio de las características escolares, de aula y de contexto que caracterizan una escuela eficaz, independientemente del enfoque metodológico utilizado (Teddile & Reynolds, 2000). Así, se concibe el efecto escolar como la capacidad que tienen los centros educativos para influir en los resultados del alumnado. En otras palabras, el efecto escolar es el porcentaje de varianza del rendimiento del estudiante debido a los factores de proceso de la escuela.

1.2. Etapas en el Movimiento de las Escuelas Eficaces

La Eficacia Escolar, es una corriente heterogénea repleta de estudios con distintos énfasis y áreas de investigación (Báez, 1994). Atendiendo al orden histórico de este movimiento, sus principales exponentes suelen referirse a cuatro grandes etapas en la historia de estas investigaciones.

Primera Etapa

El revulsivo inicial que impulsó la aparición de este movimiento fue, como se acaba de mencionar, la publicación del Informe Coleman (Coleman et al., 1966), el cual abordó la cuestión de la Desigualdad de Oportunidades en la educación. La conclusión a la que se llegó con este informe fue que la escuela tenía poco o ningún efecto sobre el éxito académico del alumno una vez controladas las variables familiares, de forma que los diferentes modos de organización y funcionamiento de las escuelas y de actuación docente tenían escasa incidencia en los resultados académicos. El modelo utilizado en este informe era el denominado de "caja negra", en el que se tienen en cuenta un conjunto de factores de entrada (tipo de centro, características personales y sociales de los alumnos, etc.) considerados como un todo unitario, y un criterio de salida que son los resultados escolares. Esta concepción pesimista de la labor de la escuela se sirvió del lema "School Doesn´t Matter" (la Escuela No Importa).

Segunda Etapa

La siguiente década (años 1971-1979) estuvo marcada por los estudios centrados en la descripción de las Escuelas Prototípicas o Inusualmente Efectivas (Edmonds, 1979) y las investigaciones se desarrollaron en torno a lo que se denominó como modelo "input-output", mucho más próximo al mundo de la economía de la educación y de los estudios de productividad. La finalidad de estos estudios era relacionar las entradas escolares, inputs (tales como el presupuesto educativo o los recursos didácticos disponibles), con los resultados escolares, outputs (habitualmente los logros académicos de los estudiantes).

Tercera Etapa

Los años siguientes (1980-1986) representaron la consolidación de la investigación sobre la Eficacia Escolar. Esta etapa es considerada como la fase de expansión y vitalidad de los movimientos de investigación de la Eficacia Escolar. Aquí se introduce en el precedente modelo de input-output los procesos (modelo "input-process-output"). Se utilizan grandes muestras, pruebas estandarizadas y técnicas cualitativas para recoger datos de la escuela y del aula. Esta etapa viene marcada por la búsqueda incipiente de modelos de análisis que logran capturar variables de las dinámicas escolares y su influencia sobre los resultados, por ello se hace notoria la mejora de las técnicas estadísticas de investigación (Creemers, 1997).

Cuarta Etapa

En la década de los años noventa, vieron la luz los llamados estudios de segunda generación. Estas investigaciones, denominadas así por Miller (1985), hacen referencia a un conjunto de estudios que no tratan de demostrar lo evidente, sino que han pasado a ocuparse de cuestiones más sustantivas, tales como la Mejora de la Calidad de la Investigación Empírica o el Análisis de los Procesos de Cambio Educativo y Organizativo. Entre estos estudios destacan los trabajos emprendidos para verificar la Eficacia de Programas Específicos de Innovación Curricular (Purkey & Smith, 1983), todos los cuales relacionan los mejores niveles de rendimiento con características como las altas expectativas del profesorado, la flexibilidad de los agrupamientos y las actividades educativas, los sistemas de evaluación, la implicación de directores en el proceso de enseñanza y la participación de los padres en la escuela, entre otros. Éstos y otros estudios se caracterizan por incorporar al modelo de entrada, proceso y producto el

"contexto". De este modo, se comienzan a utilizar modelos estadísticos de análisis multinivel que permiten separar fenómenos inter e intra-escuelas y considerar el efecto de los factores escolares no sólo sobre el rendimiento escolar, sino además sobre las relaciones estructurales dentro de las escuelas y la búsqueda de modelos explicativos consistentes (Cornejo & Redondo, 2007).

1.3. Otras Líneas de Investigación en Eficacia Escolar

Desde entonces, y hasta la fecha, la corriente de la Eficacia Escolar se ha caracterizado por la coordinación con la corriente denominada Mejora de la Escuela (School Improvement). Este movimiento nació como un enfoque distinto, cuyo objetivo era un cambio educativo planificado, mediante el cual los resultados de los estudiantes mejorarían y fortalecerían la capacidad de la escuela para gestionar cambios (Filp, 1984). Liderado por docentes y directivos, busca trasformar la realidad de una escuela para mejorarla. Sus esfuerzos se dirigen a recoger e intercambiar experiencias de innovación, centrándose en los procesos que desarrollan las escuelas que logran mejoras. En este sentido, ha tenido un desarrollo histórico paralelo (Bolívar, 1999). Sin embargo, la suposición de que ambas corrientes no son opuestas fue objeto de algunos estudios, llegando todos ellos a la firme conclusión de que en realidad un enfoque complementa al otro (Mortimore, 1992; Hopkins, 1995).

Así, mientras que la investigación en Eficacia Escolar ha estudiado la calidad y la equidad del funcionamiento de las escuelas para determinar por qué algunas son más eficaces que otras en la consecución de resultados positivos, y qué elementos se encuentran con mayor frecuencia en las escuelas que son más eficaces para todos los estudiantes; la investigación en la Mejora de la Escuela ha centrado su atención en los procesos que desarrollan las escuelas que consiguen poner en marcha un proceso de cambio para optimizar su calidad. Ambas líneas son imprescindibles para mejorar los procesos educativos desde bases científicas. Los estudios de Eficacia Escolar aportan información destacada sobre qué cambiar para educar mejor, mientras que los de Mejora Escolar proporcionan orientaciones sobre cómo llevar a cabo el cambio.

De esta forma, surge la necesidad a principios de los años noventa de la conjunción de ambas corrientes de investigación educativa en un único paradigma teórico-práctico, La Mejora de la Eficacia Escolar (Effective School Improvement) como

han puesto de manifiesto distinto autores (Stoll & Fink, 1996; Thrupp, 1999). Esta nueva corriente de investigación en el campo educativo se está desarrollando de forma plena en Inglaterra, viéndose claramente como un hecho patente en sus políticas nacionales (Reynolds, 2007). Su propósito es aportar conocimiento sobre los procesos de cambio escolar que contribuyen a mejorar el rendimiento de los estudiantes y aplicar ese conocimiento para impulsar procesos de innovación en los centros (Mac Gilchrist, Myers, & Reed, 2004; Gray et al., 1999). Sin embargo, no es la única línea de investigación a la que ha dado lugar el estudio de la Eficacia Escolar. Otra de las líneas de investigación más prometedoras, surgida para mejorar la calidad de la educación, es la investigación sobre Enseñanza Eficaz (Effective Teaching). Su objetivo principal es buscar cuáles son los factores de aula que contribuyen más eficazmente a que los estudiantes aprendan. Hasta el momento los resultados están aportando interesantes elementos para la reflexión sobre cómo tiene que desarrollarse la acción educativa en el aula para lograr criterios de mejora de la calidad y la equidad de la educación (Borich, 2009; Brown, 2009; Good, Wiley, & Florez, 2009; Orlich, Harder, Callahan, Trevisan, & Brown, 2010; Román, 2008).

1.4.De las Deficiencias en el Estudio sobre Eficacia Escolar

La crítica supone un elemento imprescindible en el crecimiento de cualquier ciencia, solamente la discusión dialéctica que supone este juego intelectual de críticas y defensas es capaz de hacer prosperar nuestro conocimiento de la realidad. En este sentido, el movimiento de investigación de Eficacia Escolar ya cuenta con 40 años de historia en los que caben muchas aportaciones, algunos desaciertos y una buena cantidad de críticas (Goldstein, 1980; Preece, 1989). Así, al mismo tiempo que se reconocían las aportaciones del movimiento de las escuelas eficaces, también se acentuaron las limitaciones y deficiencias de índole teórica, del sesgo en la medida del rendimiento a través de los tests estandarizados, y algunas otras, entre las que destacaremos como las más importantes las de carácter metodológico.

En este último aspecto se han venido manifestando dificultades referidas a la falta de claridad de conceptualización de las variables, de control de variables de input, carencia de estudios longitudinales en favor a los transversales, problemas de muestreo, inadecuación de las medidas de las variables (especialmente de los productos), la

utilización masiva de técnicas estadísticas como el análisis de la regresión, entre otras, con exigencias del cumplimiento de supuestos (no comprobados en muchos casos), junto con las limitaciones de la propia técnica para dar respuesta a los objetivos de estas investigaciones de eficacia.

Abundando en la materia de la técnica, el reconocimiento de la incapacidad de los modelos de regresión clásicos para estudiar la relación entre los estudiantes y los diversos contextos en los que se desenvuelven (Andreu, 2011; Vallejo, Arnau & Bono, 2008) ha desembocado en el desarrollo de los Modelos Lineales Jerárquicos o Modelos Multinivel, y en el Análisis de Datos Multinivel, considerado actualmente por investigadores educativos como el modo de estudio más adecuado para la temática que nos ocupa. Esto es así dado que los Modelos Multinivel respetan la organización jerárquica que presentan los datos educativos de forma natural (los estudiantes están agrupados o anidados en aulas, las aulas en centros docentes y los centros en contextos, ya sean distritos escolares, comunidades autónomas, países, u otros). Estos modelos son ampliaciones de los modelos de regresión lineal clásicos, extensiones mediante las cuales se elaboran varios modelos de regresión para cada nivel de análisis (Bickel, 2007; Reise & Duan, 2003). Cada uno de estos submodelos expresa la relación entre las variables dentro de un determinado nivel y especifica cómo las variables de ese nivel influyen en las relaciones que se establecen en otros niveles.

Por lo tanto, los Modelos Multinivel proporcionan una respuesta estadística más realista que los clásicos modelos lineales porque son más sensibles a la "agregación" de los grupos a los contextos y a la "desagregación" de los estudiantes. Así las cosas, estudian la variabilidad individual y de los grupos en los distintos niveles explorando la relación entre las unidades de observación que constituyen la estructura jerárquica (Snijders & Bosker, 1999; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2010; Van der Leeden, 1998a).

PLANTEAMIENTOS METODOLÓGICOS PARA EL ANÁLISIS DE DATOS EN LOS ESTUDIOS EDUCATIVOS

2.1. Técnicas de Análisis de Datos Clásicas en la Investigación Educativa

Tradicionalmente los alumnos en edad escolar fueron evaluados en múltiples variables, unas referentes a ellos mismos (biológicas, de capacidades, de hábitos de estudio...) y otras variables referentes a características de su profesor, clase o escuela. El objetivo final de estas investigaciones era realizar estudios cuantitativos para examinar la influencia en el rendimiento académico que ejercían variables propias del estudiante junto con otras de aula. El siguiente ejemplo permitirá ilustrar de una forma clara la metodología utilizada en los estudios educativos clásicos.

Supóngase que un investigador está interesado en demostrar que existe una relación inversa entre la cantidad de errores que cometen alumnos de 9 años en la lectura de un texto de dificultad media y la experiencia de sus profesores (medida en número de años como profesional docente). Para poner a prueba su hipótesis, este investigador no puede obviar el hecho de que el número de errores en lectura (variable continua) también va a depender de algunas características de los alumnos y tendrá que tenerlas en cuenta (como por ejemplo la nota final en la asignatura de Lengua que dichos estudiantes obtuvieron en el curso anterior). La investigación la realiza en 3 colegios de una determinada población, en 4 clases del mismo curso de cada colegio, donde en cada una hay un profesor distinto (en este caso el efecto de las características del profesor se confunde con el efecto de la clase).

Una posible estrategia de análisis sería no distinguir entre los distintos profesores y, por lo tanto, tampoco entre las clases ni entre los colegios. De esta manera, se considera únicamente al conjunto de estudiantes de todos los colegios, de todas las clases, de todos los profesores y se examina la relación entre las variables que interesan, a saber: número de errores y nota final en la asignatura de Lengua. Dadas entonces las características métricas de las dos variables, el investigador utiliza para el análisis de sus datos un Modelo Clásico de la Regresión Lineal (submodelo del Modelo Lineal General, en adelante MLG). Así, éste adquiriría la siguiente forma:

$$Y_{i=}\beta_0 + \beta_1 X_i + e_i \qquad [1]$$

Donde en [1] Y_i es el número de errores que comete el estudiante i, el coeficiente β_0 (denominado constante o intersección) indica el número de palabras que el modelo pronostica para un estudiante que inicialmente tiene un cero en la nota de Lengua del curso anterior. β_1 (pendiente de la recta de regresión) es el cambio promedio en el número de errores en lectura de los estudiantes que el modelo pronostica por unidad de aumento de nota en la asignatura de Lengua del año anterior. Por tanto, no es baladí, ni el significado de β_0, ni el signo de β_1. Tal vez, pensar que β_0 es el valor que adquiere $\hat{y}$ cuando $X = 0$ no sirva de mucho. En este sentido, quizás interese más conocer cuál es el número de errores que los estudiantes cometen cuando en Lengua alcanzan una nota igual a la media que han alcanzado el conjunto de todos los estudiantes. Si así es, se debe centrar la variable predictora. De este modo β_0 se convierte en la media de la variable dependiente y, que es justamente el valor pronosticado para la puntuación media en la nota de Lengua. El centrado de variables facilita la interpretación de los resultados y la realización de inferencias sobre los parámetros dado que, a menudo, las variables psicológicas son medidas en escalas de intervalo y el origen es difícilmente interpretable (el valor del cero es arbitrario y no indica ausencia de medida). Al centrar la variable predictora sólo cambia el valor de la intersección y, por tanto, su significado; pero no cambian ni el valor de la pendiente ni el valor de los residuales. Finalmente, el término ε_i, (error) representa el residual asociado a cada pronóstico individual (la diferencia existente entre los aciertos que en realidad tienen y los pronosticados por el modelo). Bajo este modelo, se asume que estos errores se distribuyen normalmente con varianza finita (σ_e^2).

Si tal como se acaba de hacer, se elige trabajar sólo a nivel del estudiante con todos los alumnos de todas las clases ignorando las características del contexto (profesor, clase) se puede caer en la denominada *Falacia Atomista*, proponiendo que las mismas asociaciones encontradas a nivel individual se producen a nivel contextual (Alker, 1969; Rose, 1985). En este caso, los valores estimados de la variable dependiente, $\hat{y}$, se representarían mediante una sola recta (la correspondiente a la estimación del modelo [1]).

No obstante, en este estudio el investigador desea examinar la variable profesor (y evitar caer así en la falacia atomista) y decide hacerlo de dos maneras distintas. Una de ellas mediante modelos de regresión simple, realizando un modelo para cada clase (que en este caso se confunde con la variable profesor) y, la segunda, mediante un modelo de regresión múltiple. A continuación se muestra como sería.

Análisis de los Datos mediante la Regresión Lineal Simple

El investigador tendrá tantos modelos como clases (profesores) y podrá conocer qué acontece en la clase A, B, C y D del Colegio 1, como sigue:

$$\begin{aligned} Y_{iA=}&\beta_0 + \beta_1 X_{iA} + e_{iA} \\ Y_{iB=}&\beta_0 + \beta_1 X_{iB} + e_{iB} \\ Y_{iC=}&\beta_0 + \beta_1 X_{iC} + e_{iC} \\ Y_{iD=}&\beta_0 + \beta_1 X_{iD} + e_{iD} \end{aligned} \qquad [2]$$

Y así sucesivamente hasta conocer lo que sucede en el total de las clases de la suma de colegios participantes. En esta situación, pueden suceder varios casos. Puede ser que las clases tengan aproximadamente la misma media inicial o, justamente todo lo contrario, que las medias sean claramente distintas. Puede ocurrir, también, que las pendientes en las cuatro clases sean muy parecidas a las de los cuatro grupos o, puede, que sean mayores (positivas) en las clases A y B que en C y D. Incluso pudieran diferir todas ellas en las medias y en las pendientes. Sin embargo, analizando los datos con la regresión simple no se puede saber si esa variabilidad aparente entre las medias y entre las pendientes es o no estadísticamente significativa. En estos modelos, se cae en el error de asumir que β_0 y β_1 son fijos, exclusivos de cada clase y absolutamente independientes.

Análisis de los datos mediante la Regresión Múltiple

Para distinguir entonces cómo las distintas experiencias de los profesores afectan a los errores cometidos por los estudiantes, se puede considerar un modelo para todos los estudiantes en el que se incluya la experiencia del docente, la variable Z, por lo que el modelo anterior quedaría de la siguiente forma:

$$Y_{ij=}\beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + e_{ij} \qquad [3]$$

Donde en [3], Y_{ij} son los errores del estudiante *i,* cuyo profesor es *j,* y Z_j es la experiencia medida en número de años de dicho profesor. β_2 es el cambio promedio en errores entre dos estudiantes con la misma nota media, pero atendidos por dos profesores que se diferencian en un año de experiencia (el coeficiente asociado a la variable profesor) y e_{ij}

es el término de error residual. Ahora la representación gráfica de este nuevo modelo, más general, es un conjunto de rectas ligado al número de errores, una recta para cada uno de los profesores, pero todas las rectas paralelas, pues la pendiente o efecto de las notas sobre el número de errores es β_1 que no depende de *j*.

El modelo de regresión múltiple de la ecuación [3], aunque tiene en cuenta la variable explicativa del nivel 2, no recoge adecuadamente la estructura jerárquica de los datos (los parámetros β_0 y β_1 se consideran fijos y no permite que puedan variar en función del profesor). Pero el investigador puede hacer posible que las notas medias de los estudiantes dependan de la experiencia del profesor añadiendo un término de interacción $\beta_2 X_{ij} Z_j$ al modelo anterior (interacción entre la experiencia del profesor y los errores en lectura que cometen los estudiantes) del siguiente modo:

$$Y_{ij=}\beta_0 + (\beta_1 + \beta_2 Z_j)X_{ij} + \beta_2 Z_j + e_{ij} \qquad [4]$$

Así, el coeficiente de las notas $(\beta_1 + \beta_2 Z_j)$ ahora sí depende de la experiencia del profesor. La representación gráfica de este nuevo modelo es un conjunto de rectas, una para cada profesor, pero sin la restricción de que deban ser paralelas, pues la pendiente de cada una de ellas ahora depende de la experiencia del profesor que imparte clase a los estudiantes.

En fin, múltiples ecuaciones (tantas como profesores, clases) jamás aportarían el peso que ejerce la experiencia del profesor sobre el número de errores. Sería imposible realizar ninguna predicción para un profesor cualquiera. Pero el ajuste del modelo para cada profesor será bueno, y las asunciones del modelo se cumplirán sin problemas. Una única ecuación aditiva, asumiría que los parámetros β_0 y β_1 son independientes del profesor al que pertenecen los estudiantes (asume que no están asociados a las características de los profesores que dan clase a los alumnos). Una única ecuación no aditiva no permitiría saber de manera correcta y adecuada qué peso ejerce exclusivamente sobre la variable dependiente la nota en Lengua, ni discernir qué peso ejerce por sí sola la experiencia del profesor, ni qué cualidades del profesor son las responsables de que los estudiantes cometan menos errores en la lectura. Además, cuando el investigador analice el ajuste de estos dos últimos modelos se encontrará con la sorpresa de que los errores no son independientes. Lo que está sucediendo es que ninguno de estos modelos propuestos incorpora la estructura natural jerárquica de los datos. Mientras que las notas de los estudiantes en Lengua pertenecen al estudiante (nivel 1), la experiencia del profesor que les da clase pertenece al contexto (nivel 2).

No obstante, el investigador puede buscar otra alternativa para poder estimar el efecto del contexto (profesor en este caso) analizando los datos con variables agregadas en lugar de utilizar las medidas individuales de todos los estudiantes. En el ejemplo seguido equivaldría a estudiar la relación entre los errores medios de los estudiantes atendidos por los distintos profesores en función de las notas medias en Lengua obtenidas en el curso anterior de todas las clases. En esta nueva situación no se tienen notas y errores de los estudiantes sino que se tiene la media de estas variables en los distintos grupos de alumnos (clases) que atienden los diversos profesores. Si así es, las tres variables pertenecen al nivel contextual y tampoco en este caso existe jerarquía entre ellas. En este nuevo modelo las variables del contexto también son consideradas parámetros fijos lo que hace que las inferencias sean exclusivas para los profesores concretos muestreados, y no para toda la población de la que proceden. Si se analizan así los datos, con variables agregadas, y se pretenden deducir relaciones para los estudiantes, en el caso de encontrarlas, ya que el hecho de perder una gran cantidad de información afecta negativamente a la potencia de la prueba (Hill & Rowe, 1996; Hox, 1998; Goldstein, 1995), se incurriría en la denominada *Falacia Ecológica* (Morgenstern, 1995; Robinson, 1950).

Las diferentes formas de análisis que se acaban de mencionar tienen en común una particularidad, sólo contemplan una unidad de análisis (estudiante o contexto) cuando en realidad existen dos unidades de análisis, que, además de ser necesariamente variables aleatorias, están organizadas jerárquicamente de modo natural. Esto implica, de una parte, reconocer que la unidad de nivel superior no refleja el comportamiento de los estudiantes individuales, sino su variabilidad y, por lo tanto, los resultados con datos agregados pueden ser muy distintos a los obtenidos con las puntuaciones individuales de todos los estudiantes, generándose así *Sesgos de Agregación* (Roberts & Burstein, 1980). Además, a pesar de esa variabilidad, las unidades que están anidadas en el nivel superior necesariamente mantienen cierto parecido entre sí por el hecho de haber estado expuestas a las mismas condiciones (De la Cruz, 2008), en otras palabras, los estudiantes que están bajo la tutela de un profesor adquieren comportamientos de grupo distintos a los que están bajo la tutela de otro profesor. Por lo tanto, necesariamente hay que suponer que las respuestas de los distintos estudiantes del nivel inferior no sean independientes.

Los modelos de regresión anteriores ([1]-[4]) son posibles y, de hecho, se pueden encontrar en todos los ámbitos de las investigaciones, desde el campo de las Ciencias Sociales (Psicología, Educación...), las Ciencias Médicas (Enfermería, Epidemiología, Medicina...), las Ciencias Económicas, etc., siempre y cuando aquello que se desee estudiar esté exento de los efectos contextuales o individuales o, intencionalmente, se hayan aislado. El investigador anterior ha planteado todos estos modelos dentro de los límites del MLG, es decir, ha asumido que:

$$y = X\beta + \varepsilon \qquad [5]$$
$$E(y) = X\beta$$
$$Var(y) = Var(\varepsilon) = V = \sigma^2 I$$

Donde en [5] $\boldsymbol{y}$ es un vector de la dimensión $n \times 1$, conteniendo los valores de la variable de respuesta y_i para la observación i en el grupo j (o bien para el sujeto i dentro del tiempo j); $\mathbf{X}$ es una matriz de diseño de dimensión $n \times p$, la cual especifica los valores de los efectos fijos que corresponden a cada parámetro para cada una de las observaciones (vectores de ceros y unos que denotan la ausencia y presencia de efectos categoriales para variables carentes de estructura numérica, mientras que los vectores numéricos denotan los efectos de las variables medidas en una escala cuantitativa). $\mathbf{X}$ incluye las covariables explicativas, $\boldsymbol{\beta}$ es un vector de parámetros no aleatorios estimados asociados a las covariables del modelo (que pueden incluir variables de diversa naturaleza), y por último $\boldsymbol{\varepsilon}$ es un vector de errores aleatorios desconocidos de dimensión $n \times 1$ distribuido normal e independientemente, con media cero y varianza constante.

Sin embargo, actualmente las investigaciones en educación buscan todos aquellos factores asociados al aprendizaje escolar (Cornejo & Redondo, 2007) y este modelo clásico no sirve, básicamente, por dos razones. La primera de ellas, es que en el ámbito de la educación hay más de una unidad de análisis, existe más de una variable aleatoria (estudiantes y clases) y este modelo sólo contempla una. La segunda es que el MLG asume que las respuestas de los estudiantes que pertenecen a una misma unidad de análisis del nivel superior son independientes, cuando en realidad están relacionadas y, como consecuencia, los errores estándar se subestimarían y se concluirá en muchas ocasiones que los resultados son estadísticamente significativos cuando en realidad no lo son (Hox, 1995).

2.2. Del Modelo Lineal General al Modelo Lineal General Mixto: Los Modelos Multinivel

Hace casi ya tres décadas, que dos matemáticos ingleses, Aitkin y Longford (1986), redactaron un artículo que conmocionó el mundo de la investigación educativa. En dicho manuscrito, simplemente, se puso de manifiesto lo que es hoy evidente, que la realidad de los sistemas educativos (donde los estudiantes están agrupados en clases, éstas a su vez en escuelas, y estas últimas en provincias o regiones), conlleva que los estudiantes de un mismo grupo compartan experiencias diferentes a los de otras clases, al igual que las aulas de una escuela tienen la misma dirección o el mismo clima escolar y, así las cosas, no se puede negar la dependencia entre las unidades de una misma unidad de análisis. Precisamente estos autores pusieron de manifiesto lo que anteriormente se acaba de exponer, que el MLG no asume esto que de modo natural sucede en las estructuras organizadas jerárquicamente, y que desde entonces no deja de repetirse (Gelman & Hill, 2006; Heck & Thomas, 2000; Hox, 1998).

A partir de ese examen crítico, Aitkin y Longford (1986) dieron a conocer al campo de la educación una nueva técnica de análisis, los Modelos Lineales Jerárquicos o Modelos Multinivel. No obstante pese a que estos modelos eran desconocidos en dicho campo, otros ámbitos como la Econometría, la Estadística, etc. ya los tenían incorporados. De hecho, la apropiación de esta técnica de análisis en las diferentes disciplinas ha dado lugar a una ingente cantidad de denominaciones en la literatura científica especializada. Se habla entonces de Modelos Lineales Multinivel -Multilevel Linear Models- (Goldstein, 1987); Modelos de Componentes de Varianza -Covariance Components Models- (Dempster, Rubin, & Tsutakawa, 1981); Modelos de Efectos Fijos y de Efectos Aleatorios -Mixed-Effects Models and Random-Effects Models- (Elston & Grizzele, 1962) y Modelos de Regresión de Efectos Aleatorios -Random Coefficient Regression Model- (Rosenberg, 1973), entre los más relevantes.

Ahora bien, estos modelos bajo cualquiera de sus sobrenombres, tanto en su dimensión teórica como aplicada, tratan conjuntos de datos anidados dentro de una población con estructura jerárquica, entendiendo que las distintas jerarquías se corresponden con diferentes niveles del modelo. El siguiente ejemplo permite aclarar mejor la configuración. Supónganse una estructura de datos en dos niveles, tal como muestra la Figura 1. El Nivel Macro (High-Level) se relaciona con los contextos (por ejemplo, una clase determinada de un colegio) y el Nivel Micro (Lower-Level) hace

referencia a los estudiantes que están anidados en el nivel superior (por ejemplo, los estudiantes que se encuentran en una clase).

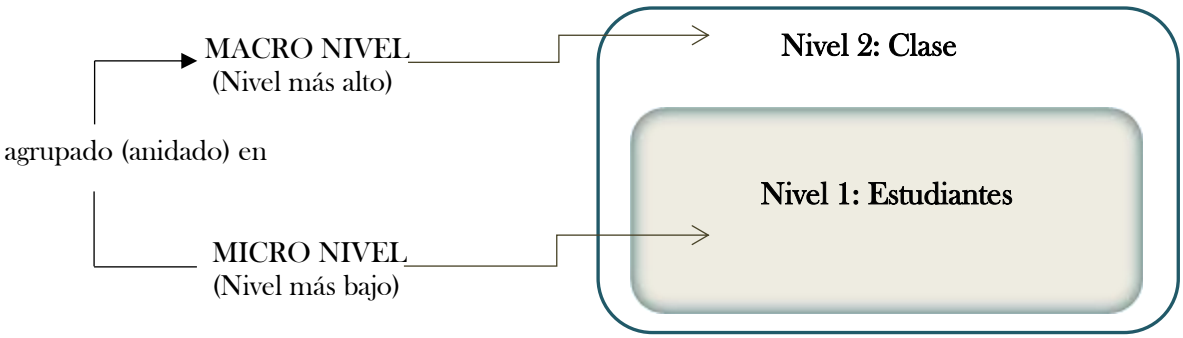


Figura 1. Disposición de los Datos en un Modelo Multinivel de 2 Niveles.
Adaptado de Amador & López-González, 2007.

Respetar la organización agrupada que presentan los datos educativos en su forma natural es muy importante. Los modelos multinivel se tornan, por tanto, en la solución trabajando así con dos, tres, o más unidades de análisis aleatorias de forma simultánea y estimando los componentes de varianza correspondientes a todas ellas. De esta forma se conoce en qué medida cada unidad de análisis es capaz de explicar la conducta que se desea observar.

Además, los modelos jerárquicos permiten estudiar de manera independiente la variabilidad de cada una de las unidades de análisis (en las que se hallan agrupados los datos) que han resultado estadísticamente significativas (asumiendo la dependencia entre los elementos que la constituyen), y especificar un modelo adecuado para cada una de ellas. Después concatenan todos los modelos y combinan la información obtenida a través de los diferentes niveles, examinando simultáneamente los efectos de cada una de las variables presentes en la investigación, así como las interacciones entre las variables de un mismo nivel y de distintos niveles. Bajo estas circunstancias, cometer la falacia atomista o la ecológica va a requerir más que arte.

Continuando con el estudio del investigador que se viene desarrollando, por ejemplo, partiendo de la ecuación [1], para considerar que los elementos de la unidad de análisis del nivel 1 (estudiantes) están anidados en los elementos de la unidad de análisis del nivel 2 (profesores), basta hacer un simple gesto sobre la ecuación del MLG,

añadiendo un subíndice extra a los coeficientes de las variables exclusivas del nivel 1 del siguiente modo:

$$Y_{ij=}\beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \qquad [6]$$

En este nuevo modelo ya no se puede decir que no importa qué profesor tengan los estudiantes porque cada profesor tiene un valor específico de β_0 y β_1, esto es, el modelo ahora permite a cada profesor tener su propia intersección y su propia pendiente. Y, ciertamente, esta variabilidad en el nivel 2 (profesores) es lo que caracteriza a un modelo multinivel: el modelo refleja cómo se relacionan X e Y en las unidades del nivel 1 (estudiantes) en cada uno de los subgrupos definidos por el nivel 2 (profesores).

El caso más simple de un modelo multinivel es el que, como éste, tiene sólo dos niveles: nivel 1 (estudiantes) y nivel 2 (profesores), que en este caso se confunde con la variable clase. Sin embargo, el modelo aún no está adecuadamente formulado. En realidad un modelo multinivel resulta, como se ha señalado anteriormente, de integrar un sistema de ecuaciones en dos pasos (o más, dependiendo de las unidades de análisis o niveles que contemple). En el primer paso (nivel 1), se define una ecuación de regresión para cada unidad del primer nivel con las variables de este nivel (estudiante) que determinen la variable dependiente. En este caso sería idéntica a la ecuación [6]. El parámetro β_{0j} refleja la media de errores en lectura que tienen los estudiantes anidados en el *j*-ésimo profesor (clase). β_{1j} refleja el cambio promedio en el número de errores asociado con una unidad de cambio en las notas de Lengua, y e_{ij} denota la diferencia entre los errores obtenidos por el estudiante i y la media de los alumnos del *j*-ésimo profesor (clase), es decir, la varianza que permanece sin explicar después de controlar la nota en Lengua del año anterior. Por simplicidad, se asume que los errores del nivel 1 son variables aleatorias que tienen una distribución normal con media cero y varianza constante a través de los distintos profesores (clases). Esto es, $\varepsilon_{ij} \sim N(0, \sigma^2)$.

A continuación se procede a incorporar la naturaleza jerárquica de los datos en el modelo del nivel 2. Por ejemplo, se puede definir β_{0j} y β_{1j} como sigue:

$$\begin{aligned} \beta_{0j} &= \gamma_0 + u_{0j} \\ \beta_{1j} &= \gamma_1 + u_{1j} \end{aligned} \qquad [7]$$

El parámetro β_{0j} está formado por una parte fija o sistemática γ_0 que representa los errores medios de los estudiantes de todos los profesores, y una parte aleatoria u_{0j} que refleja la variabilidad de los alumnos de cada profesor respecto de esa media

poblacional. Se asume que estos dos términos son independientes. Del mismo modo, el término β_{1j} está formado por una parte fija o sistemática γ_1 que es la pendiente media que relaciona la mejora en el número de errores con la nota en Lengua en la población de profesores, y una parte aleatoria u_{1j} que refleja la variabilidad de las pendientes de los distintos profesores respecto de esa pendiente poblacional media. También se asume que estos dos términos son independientes. Además, se toman los términos u_{0j} y u_{1j} como variables aleatorias con valor esperado cero y varianzas σ_{u0}^2 y σ_{u1}^2 respectivamente, y covarianza σ_{01}.

Este paso es muy importante dado que, ahora, los parámetros β_{0j} y β_{1j} ya no se interpretan como constantes fijas, como en el modelo de regresión clásico [1], sino como variables cuyos valores pueden cambiar de un profesor a otro como se ha indicado; y varían tanto en función de su media como en función del efecto aleatorio asociado con cada uno de los profesores (unidades del nivel 2). Si se sustituyen en [6] los parámetros β_{0j} y β_{1j} de [7] por su valor, ya se tiene el modelo completo:

$$Y_{ij=}\gamma_0 + \gamma_1 x_{ij} + (u_{0j} + u_{1j}x_{ij} + e_{ij}) \qquad [8]$$

Hay que resaltar aquí que entre los términos β_{0j} y β_{1j} no se asume independencia. La relación, entre ambos, viene dada por: $\rho(\beta_{oj}, \beta_{1j}) = cov(\beta_{oj}, \beta_{1j})/(\sigma_{uo}\sigma_{u1})$.

Si el tamaño de las medias es independiente del tamaño de las pendientes, entonces $\rho(\beta_{oj}, \beta_{1j}) = 0$. Esto no es trivial. Si todas las clases comparten la misma ecuación de regresión ($\sigma_{uo} = \sigma_{u1} = 0$) pueden suceder dos cosas. O bien que la nota en Lengua no influya de ninguna forma sobre los errores en lectura en ninguno de los profesores (entonces, $\gamma_1 = 0$ en todos los profesores) o, puede que las notas en Lengua sí intervengan en el número de errores pero en la misma medida en todos los profesores (entonces, $\gamma_1 \neq 0$ pero idéntico en todos los profesores).

No obstante, no se agotan aquí todas las posibilidades. Puede suceder que $\rho(\beta_{oj}, \beta_{1j}) = 0$ y sin embargo ($\sigma_{uo} \neq \sigma_{u1} \neq 0$) . Por ejemplo, puede ocurrir que todas las clases (profesores) tengan medias distintas ($\sigma_{uo} > 0$) pero la misma pendiente ($\sigma_{u1} = 0$), o puede suceder que las clases difieran tanto en las medias ($\sigma_{uo} > 0$) como en las pendientes ($\sigma_{u1} > 0$).

Por otra parte, si las pendientes de las clases (profesores) son tanto mayores cuanto mayores son las medias, entonces $\rho(\beta_{oj}, \beta_{1j}) > 0$, y si las pendientes de las clases (profesores) son tanto menores cuando mayores son las medias, entonces, $\rho(\beta_{oj}, \beta_{1j}) < 0$.

Dado que tanto las medias como la relación entre X e Y pueden variar de profesor a profesor, puede ser interesante incluir en el modelo una o más variables del nivel 2 que puedan dar cuenta de dicha variación. Por ejemplo, los profesores quizá se diferencien en función de su método de enseñanza. Así, puede haber profesores basados en un enfoque de enseñanza orientado a la transmisión de la información, más centrado en el propio docente - Information Transmission/Teacher-Focused (ITTF) approach - y profesores con un enfoque orientado al cambio conductual, más centrado en el estudiante - Conceptual Change/Student-Focused (CCSF) approach - (ITTF=0, CCSF=1). Podría entonces darse el caso de que esta distinción del nivel 2 fuera la responsable (o al menos en parte) de la variabilidad existente, no ya sólo entre las medias de errores en lectura de las clases (profesores), sino entre las pendientes que relacionan los errores en lectura con la nota de Lengua del curso anterior. Para incluir en el modelo esta variable del nivel 2 se deben definir los parámetros β_{0j} y β_{1j} de otra forma, como la que a continuación se expresa:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \end{aligned} \qquad [9]$$

Si en la ecuación [6] se sustituyen los nuevos valores de β_{0j} y β_{1j} se obtiene:

$Y_{ij=}\gamma_{00} + \gamma_{01} Z_j + u_{0j} + \gamma_{10} x_{ij} + \gamma_{11} x_{ij} Z_j + u_{1j} x_{ij} + e_{ij}$. Reordenando los términos del modelo, colocando los efectos fijos al principio y los aleatorios al final (entre paréntesis) la ecuación queda del siguiente modo:

$$Y_{ij=}\gamma_{00} + \gamma_{01} Z_j + \gamma_{10} x_{ij} + \gamma_{11} x_{ij} Z_j + (u_{0j} + u_{1j} x_{ij} + e_{ij}) \qquad [10]$$

En el caso de que hubiese más niveles y más variables, el planteamiento y el desarrollo sería análogo al realizado con dos niveles.

El modelo representado en la ecuación [10] incluye los efectos de las variables grupales (γ_{01}), las variables individuales (γ_{10}) y su interacción (γ_{11}) en el resultado individual Y_{ij}. Estos coeficientes (γ_{01}, γ_{10} y γ_{11}), además de γ_{00}, son comunes a todos los estudiantes, independientemente del grupo al que pertenecen, y suelen llamarse

efectos o coeficientes fijos. El modelo también tiene un componente de intersección aleatoria u_{0j} y un componente de pendiente aleatoria u_{1j}. Los valores de estos componentes varían aleatoriamente entre los profesores (clases), por lo que se denominan efectos o coeficientes aleatorios. Los parámetros de las ecuaciones anteriores (coeficientes fijos, coeficientes aleatorios, varianza de los efectos aleatorios y varianza residual) se estiman simultáneamente mediante métodos iterativos. Las varianzas de nivel 1 y de nivel 2 (σ^2, σ_{u0}^2 y σ_{u1}^2) respectivamente, y la covarianza σ_{01}, se llaman componentes de varianza.

Existen diferencias notorias entre el modelo [6] y el modelo [10]. En el primero, se asume independencia entre los errores de los estudiantes de cada clase (profesor) e igualdad de varianzas entre las clases (profesores). En el segundo no. La parte aleatoria del modelo [10] es más compleja que en el modelo de regresión lineal convencional (únicamente incluye e_{ij}). Los residuos de un modelo multinivel no son independientes dentro de cada clase (profesor) porque los componentes u_{0j} y u_{1j} son comunes a todos los estudiantes de la misma clase (con el mismo profesor). Por otro lado, la varianza de los residuos no es la misma en todas las clases (profesores) dado que tanto u_{0j} como y u_{1j} pueden variar de clase a clase.

Los modelos mixtos, por tanto, descansan en la partición del error no explicado en un componente común a las observaciones procedentes de una misma unidad de muestreo y un término residual del error, propio de cada observación y, en principio, independiente de los términos residuales del resto de las observaciones. Además incluyen en su formación parámetros fijos, comunes a toda la población, y parámetros aleatorios, específicos de cada unidad de muestreo. Los parámetros aleatorios se consideran realizaciones aleatorias de un proceso de media cero y cuya varianza define la componente del error asociada a la unidad de muestreo.

En suma, ya no se está bajo la tutela del MLG, sino del Modelo Lineal General Mixto (en adelante MLM). Para finalizar, se puede simplificar la ecuación [10] escribiéndola para el conjunto de los *N* estudiantes y de las *J* clases (profesores) en la notación matricial que sigue:

$$\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}\mathbf{u} + \boldsymbol{\varepsilon} \qquad [11]$$

Donde en [11], $\mathbf{Y}$ es un vector de respuestas, $\mathbf{X}$ y $\mathbf{Z}$ son matrices de diseño, $\boldsymbol{\beta}$ es un vector que contiene todos los parámetros de efectos fijos de la ecuación, y $\mathbf{u}$ y $\boldsymbol{\varepsilon}$ son

vectores que contienen los efectos aleatorios de los profesores (nivel 2) y de los estudiantes (nivel 1), respectivamente. Se asume que el vector de coeficientes aleatorios $\mathbf{u}$ se distribuye normal e independiente de $\boldsymbol{\varepsilon}$ con media cero y matriz de covarianza $\mathbf{G}$, donde $\mathbf{G}$ es una matriz $\mathbf{G}[\mathbf{u}\sim \mathrm{N}(0, \mathbf{G})]$ de covarianza de parámetros aleatorios (que en el caso más general de modelos anidados con datos registrados de modo transversal tiene una formación no estructurada), diagonal, de bloques de dimensión *k* x *k*. También se asume que el vector de coeficientes aleatorios $\boldsymbol{\varepsilon}$ se distribuye normal e independiente con media cero y matriz de covarianza $\mathbf{R}$, $\mathbf{R}[\boldsymbol{\varepsilon}\sim \mathrm{N}(0, \mathbf{R})]$.

En el MLM, los elementos de $\boldsymbol{\varepsilon}$ (a diferencia del tradicional) no requieren ser independientes, ahora bien, en el modelo anidado [10] cuyos datos responden a un diseño transversal en el que sólo se tiene un registro para cada estudiante, se asume que $\mathbf{R}$ es una matriz diagonal de dimensión $\mathbf{I}$. Por consiguiente, bajo los supuestos de la distribución especificados para $\mathbf{u}$ y para $\boldsymbol{\varepsilon}$, la distribución marginal de $\mathbf{y}$ es $\mathbf{y}\sim \mathrm{N}[\mathbf{X}, \boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})]$ donde $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ es la matriz de covarianza global para el modelo de la ecuación.

El MLG que se expuso anteriormente es un caso particular del MLM porque cuando $\mathbf{R} = \sigma^2\mathbf{I}$ y $\mathbf{Z} = 0$, ambos enfoques son plenamente coincidentes.

EL MLM EN EL CONTEXTO DE LOS DISEÑOS LONGITUDINALES

Hasta ahora se ha contemplado el análisis de unos datos que han sido recogidos siguiendo el patrón de algún diseño transversal, es decir, tomados para todos los estudiantes en un único momento, bajo unas circunstancias concretas. Sin embargo, otro modo de hacerlo es siguiendo un patrón de algún diseño longitudinal, o lo que es lo mismo, registrando la variable de interés en sucesivas ocasiones pautadas, bien por la administración de tratamientos distintos, bien por el tiempo. En fin, el propósito de la investigación y las hipótesis que se pueden poner a prueba, con datos registrados en sucesivas ocasiones desde la misma unidad de análisis puede adquirir múltiples matices. De hecho, son los diseños que obedecen a esta estructura los más utilizados tanto en investigación experimental, cuasi-experimental como no experimental (Fernández, Livacic-Rojas, & Vallejo, 2007).

Estos diseños (en cuanto al modo de recoger los datos) tienen unas ventajas notables respecto a las investigaciones transversales, entre las que merece destacarse una principalmente. El estudiante se convierte en control de sí mismo y por lo tanto las diferencias individuales son menores, lo que hace que el término de error sea menor que en cualquier otro diseño. Por este motivo la potencia de la prueba que ostentan es valorada de forma extraordinaria.

Pero la perfección no existe, y estos diseños requieren labores de experto en todas las etapas de la investigación para que el error no resulte enrarecido y para conseguir no perder registros ni estudiantes. Esto es así fundamentalmente por dos razones: porque la independencia de las observaciones es difícil de mantener (de modo natural las conductas emitidas por una mismo estudiante que están más cercanas en el tiempo, están más correlacionas que las más alejadas, admitiendo también que otros muchos patrones de relación son posibles, pero patrones al fin y al cabo alejados de la higiene que requiere el patrón que asume el MLG), y porque es común que el interés de los estudiantes por participar en la investigación decrezca con el paso del tiempo.

Fernández et al. (2007) ofrecen una panorámica en la búsqueda de procedimientos alternativos, unos univariados y otros multivariados, para solventar estos problemas. Y todos ellos funcionan muy bien en determinadas ocasiones, pero casi ninguno es capaz de atender a todas ellas. No obstante uno, el MLM, ha marcado un antes y un después en el análisis de los datos longitudinales.

El MLM presenta muchas ventajas sobre el resto de procedimientos de análisis que lo convierten en una estrategia sólida para realizar inferencias más exactas tanto de los parámetros del modelo -efectos de los tratamientos e interacción- (Kowalchuck, Keselman, Algina, & Wolfinger, 2004; Littell, 2002; Wolfinger, 1996) como de sus errores estándar (Núñez-Antón & Zimmerman, 2001), a saber:

- Permite analizar los datos cuando se tienen observaciones perdidas (datos incompletos), es decir, cuando para algunos estudiantes no se dispone de todos los registros, con independencia de que el número de niveles de la variable intra-estudiante exceda o no el número de alumnos, esto es, sin importar el tamaño de la muestra. También permite que para cada estudiante se realicen diferente número de registros (tal vez porque se ha logrado determinada meta, por ejemplo, que hayan alcanzado una determinada destreza). Hay que señalar aquí que, sin excepción, todos los demás procedimientos de análisis desarrollados para estos diseños prescinden de los estudiantes que no tienen todas las observaciones. El MLM también permite que los intervalos de tiempo entre cada aplicación de tratamiento sean específicos para cada alumno.

- Admite que el modelo contenga tanto variables fijas como efectos aleatorios (Raudenbush & Bryk, 2002; Kuehl, 2001).

- Posibilita el manejo de covariadas cuantitativas y categóricas no sólo estables, sino con la posibilidad de ir cambiando con el transcurso del tiempo y, de este modo, proporciona pruebas más potentes de los efectos de los tratamientos.

- Permite obtener de una forma más eficiente los estimadores de los parámetros usados en la modelización de la estructura de medias y, también, lograr errores estándar más adecuados de estos estimadores.

Finalmente, hay que tener en cuenta que dado que el investigador generalmente desconoce cuál es la estructura de covarianza que subyace a sus datos, la aplicación de criterios de selección de modelos para detectar la estructura de covarianza más

parsimoniosa favorece la precisión y eficiencia del análisis a la hora de detectar el patrón de cambio operado en los participantes del estudio (Vallejo et al., 2010). Permite modelar las variaciones entre e intra-individuos tanto para datos completos como para incompletos hasta encontrar cual es la estructura de la matriz **Σ** que mejor se ajusta antes de proceder al análisis (Vallejo, Fernández, & Secades, 2004).

3.1. Formulación General del MLM para Datos Longitudinales

El modelo que subyace a los datos recogidos de modo secuencial bajo el paraguas del MLM es, en su modo más general, idéntico al expuesto previamente para un diseño transversal jerárquico en la ecuación [11]:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \qquad \text{[11 repeated]}$$

Ahora bien, llegados a este punto hay que hacer aquí una distinción de dos situaciones distintas en función del objetivo de la investigación. Una de ellas se plantea en la Investigación Experimental y en la Investigación Cuasi-experimental, consistente en examinar el efecto que ejercen determinados tratamientos sobre una variable dependiente de interés, la otra es la denominada Investigación de Curvas de Crecimiento en la que interesa no sólo la evolución del grupo sino también la evolución individual a lo largo del tiempo. La primera la vamos a denominar Diseño de Medidas Repetidas, la segunda Curvas de Crecimiento. Entre una situación y otra no cabe duda de que hay una enorme cantidad de variaciones, siendo una de ellas en la que se han fundamentado los actuales Modelos de Valor Añadido llevados a cabo de manera intensiva en la investigación educativa.

3.2. Diseño de Medidas Repetidas

Del modelo anterior expuesto se va a prescindir del término **Zu**, esto es $\mathbf{Z} = 0$. En este caso se asume que no existen más efectos aleatorios que los del error asociado a cada respuesta de los estudiantes. En teoría este enfoque puede resultar útil para aplicaciones que impliquen un número reducido de observaciones de cada estudiante suficientemente espaciadas y como resultado de la aplicación de determinados tratamientos, o de seguimiento en pocas ocasiones con ánimo de observar el alcance de los efectos de los tratamientos.

Así las cosas, el modelo lineal estándar para analizar medidas repetidas obtenidas desde p grupos de n_j estudiantes (i=1,..., n_j; $\sum n_j$ =N) en un conjunto de q ocasiones (bajo distintos niveles de tratamiento), puede ser descrito como:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \qquad [12]$$

Donde en [12], $\mathbf{Y}_i = (y_{i1}, \ldots, y_{it})'$ es un vector de medidas repetidas de dimensión qx1 efectuadas en el *i*-ésimo participante, $\mathbf{X}_i = (X'_{i1}, \ldots, X'_{it})'$ es una matriz de diseño conocida de orden $q \times h$, $\boldsymbol{\beta}$ es un vector de parámetros desconocidos de la población de orden $h \times 1$ asumiendo valores fijos (efectos fijos) y $\boldsymbol{\varepsilon_i}$ es un vector de errores aleatorios de orden $q \times 1$.

Se asume sólo efectos fijos y, por lo tanto, la matriz de varianza-covarianza ($\boldsymbol{\Sigma}$) de las observaciones en todos los estudiantes, la varianza de $\mathbf{Y}$, denotada por $\mathbf{V}$, es:

$$\mathbf{V} = \mathbf{I}_n\boldsymbol{\Sigma}$$

De esta manera se acepta que $\boldsymbol{\Sigma} = \mathbf{R}$. En esta situación puede ocurrir que se cumplan las asunciones del modelo mixto de Scheffé y entonces la matriz $\mathbf{R}$ representaría la correlación constante entre todos los pares de observaciones de un mismo estudiante y varianzas homogéneas. Sin embargo, la experiencia de muchos investigadores converge en que lo habitual es que el registro de medidas repetidas no satisfaga el simple supuesto de simetría combinada o compuesta, ni tampoco el de esfericidad. La ventaja principal entonces del MLM radica en el hecho de que permite modelar una gran cantidad de estructuras de covarianza y elegir la más apropiada mediante criterios de ajuste. Por este motivo, algunos autores lo han denominado como Modelos de Patrones de Covarianza (Jennrich & Schluchter, 1986; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011a). Una ilustración de cómo aplicar el modelo de medidas repetidas en el ámbito de la educación se puede encontrar en el reciente trabajo de Núñez, Rosário, Vallejo y González-Pienda (2013).

3.3. Curvas de Crecimiento

Cuando un investigador decide estudiar curvas de crecimiento registra una serie de medidas en sucesivos intervalos temporales de amplitud constante, en una o más muestras de alumnos, con el fin de examinar el proceso de desarrollo de cada estudiante y las posibles diferencias existentes en dicho proceso, entre distintas muestras de

alumnos. El investigador considera que esos datos configuran una estructura jerárquica a dos niveles: las observaciones son las unidades del primer nivel y los estudiantes son las unidades del segundo nivel (Cnaan, Laird, & Slasor, 1997; Goldstein, 1989, 2003; Raudenbush & Bryk, 2002; Hertzog, Lindenberger, Ghisletta, & Von Oertzen, 2008; Snijders & Bosker, 2012).

Dado que los datos obedecen a un modelo multinivel lineal jerárquico, requieren un análisis totalmente distinto al realizado con los métodos tradicionales (Wu, Clopper, & Wooldridge, 1999), deben ser analizados en dos estadios. En el primero se define una regresión lineal para las observaciones registradas en cada estudiante. En el segundo, los coeficientes de regresión o los parámetros de las curvas de crecimiento individuales, modelados en el primer estadio, son considerados como variables dependientes aleatorias. El propósito del segundo estadio reside en analizar la distribución de estos parámetros o efectos aleatorios en la población y analizar las circunstancias que explican la variabilidad de las trayectorias individuales. Por otra parte, dado que las curvas de crecimiento representan un proceso de desarrollo que se produce en función del tiempo, una forma adecuada de modelarlas radica en describir los valores esperados de las observaciones como funciones polinómicas del tiempo (Van der Leeden, 1998b).

3.4. Una de las Imágenes del Estudio de Curvas de Crecimiento: Los Modelos de Valor Añadido

Bajo la denominación de Modelos de Valor Añadido se incluyen hoy en día diferentes modelos estadísticos aplicados a la evaluación de sistemas educativos que varían muy sustancialmente entre ellos, tanto en complejidad como en los supuestos que subyacen a los mismos (Ponisciak & Bryk, 2005; Sanders, 2006; Wiley, 2006). No obstante, todas las aproximaciones y propuestas englobadas bajo este término comparten un misma finalidad: vincular los cambios registrados en el rendimiento individual de los estudiantes con las escuelas a las que asisten o con los profesores responsables de la clase a la que pertenecen (Martínez Arias, Gaviria, & Castro, 2009). Dado que son modelos orientados al análisis del cambio, un segundo denominador común a todos ellos, y derivado del anterior, es que han de operar sobre datos que permitan el seguimiento individual del crecimiento en el rendimiento a lo largo del tiempo, con el fin de estimar la contribución de la escuela y/o el profesor a dicho crecimiento (Braun, 2005). En consecuencia, la estructura longitudinal de las medidas de rendimiento del alumnado y

las implicaciones que de ésta se derivan conforman un objeto de análisis de importancia central en las aproximaciones metodológicas a la estimación de medidas del valor añadido. En este sentido, se define el valor añadido como la contribución que hace la escuela al progreso del estudiante, en relación con una serie de objetivos, una vez eliminada la influencia de factores ajenos a la escuela en dicho progreso. El valor añadido así entendido se considera como un factor determinante de la calidad y la eficacia de un centro educativo. Por consiguiente, los modelos de valor añadido aportan la medición longitudinal de los resultados durante dos o más años, incorporando la dimensión de "progreso" a la foto fija habitual de la evaluación del rendimiento escolar (Mata & Ballesteros, 2012).

De acuerdo con lo que constituyen las propuestas recientes más aceptadas, el estudio adopta la aproximación proporcionada por los modelos jerárquico lineales para la modelización longitudinal (Singer & Willett, 2003), alineándose igualmente con los desarrollos metodológicos más actuales en la medida del valor añadido.

El problema se plantearía de la siguiente manera, en el contexto de las evaluaciones longitudinales del rendimiento, se asume que la magnitud de la correlación observada entre dos medidas dadas (y por tanto la capacidad predictiva de una medida previa sobre una posterior) es tanto mayor cuanto mayor es también la proximidad temporal entre las mismas. De este modo se espera que, en un esquema de evaluación con cuatro momentos de medida, la correlación entre la primera y la segunda medida sea superior a la que se registre entre la primera y la tercera, correlación que se espera a su vez superior a la correspondiente a la primera y la cuarta medida. Igualmente se asume que la magnitud de las correlaciones con idéntica proximidad temporal será básicamente equiparable bajo condiciones de fiabilidad constante, de modo que la correlación entre la primera y la segunda medida será similar a la registrada entre la tercera y la cuarta (Blanco, González, & Ordóñez, 2009).

Considerado lo anterior, el planteamiento del problema parte de un modelo jerárquico lineal básico para la determinación de las puntuaciones en el constructo rendimiento, con tres niveles y en el que se incorpora como predictora la variable tiempo:

Primer nivel: Ocasión/Tiempo $Y_{tjk} = \beta_{ojk} + \beta_{1jk}(t - t_0) + \varepsilon_{tjk}$

Segundo nivel: Estudiante $\beta_{0jk} = \beta_{ok} + \mu_{0jk}$; $\beta_{1jk} = \beta_{1k} + \mu_{1jk}$

Tercer nivel: Escuela $\beta_{0k} = \beta_{00} + \mu_{0k}$; $\beta_{1k} = \beta_{10} + \mu_{1k}$

No obstante, la modelización de estos niveles puede alcanzar una enorme complejidad, y por ese motivo la modelización de datos longitudinales para estudiar el valor añadido plantea un número importante de cuestiones estadísticas y psicométricas a tener presentes para desarrollar una óptima investigación (una amplia revisión puede ser consultada en Martínez Arias, 2009 y también en McCaffrey, Lakewood, Koretz, & Hamilton, 2003).

LA MODELIZACIÓN MULTINIVEL

4.1. El Proceso de Modelado Estadístico Multinivel

En particular, el análisis multinivel persigue obtener el modelo que, partiendo del marco teórico previo, mejor se ajuste a los datos (Hill & Rowe, 1996). De esta manera, el objetivo fundamental del modelado estadístico es la búsqueda del modelo más parsimonioso que sea capaz de explicar la variable de interés con el mínimo error posible. No obstante, este proceso de modelado no es tarea fácil, y diversos autores (Heck & Thomas, 2000; Kim, 2009; Peña, 2011) son coincidentes en la importancia que tiene el seguir una secuencia de aplicación de procedimientos de modelado para que el investigador pueda determinar si un modelo es, o no, aceptable como explicación de la variable. Por lo tanto, el proceso de modelado estadístico requiere el empleo de 4 etapas claramente diferenciadas, como se puede observar en la Figura 2. Para ilustrar el desarrollo del proceso de modelado estadístico, se continuará con el ejemplo de dos niveles basándose en el trabajo de Ato y Vallejo (2007).

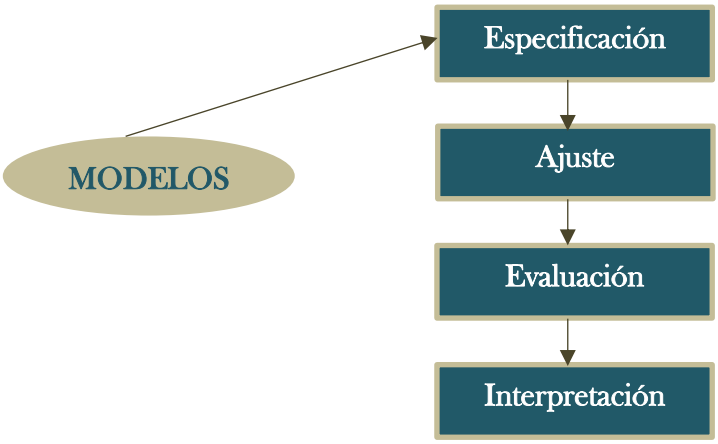


Figura 2. Etapas Principales del Proceso de Modelado Estadístico.
Tomado de Ato & Vallejo, 2007.

Etapa 1: Especificación

En esta primera etapa del proceso, el investigador, basándose en su propia intuición o en el conocimiento sustantivo de la temática de estudio, o en resultados derivados de modelos anteriores sometidos a prueba, *especifica* un modelo teórico objeto de su interés.

En este sentido, para la modelización multinivel que nos ocupa, el investigador debe formular inicialmente un modelo nulo, también denominado modelo vacío o modelo incondicional, que permitirá conocer la cantidad de varianza que pudiera explicarse a nivel individual (nivel 1) y a nivel de profesor (clase, nivel 2). Esta fase es sumamente importante. Si el componente de varianza del nivel 2 no es estadísticamente significativo no tendría sentido considerar la naturaleza jerárquica de estos datos y se analizarían los datos mediante el MLG. Si es estadísticamente significativo, se considera la naturaleza jerárquica de los datos y se utiliza el MLM para realizar el *análisis multinivel*. Además, en este primer paso, examinando la correlación intra-clase se advierte en qué medida los estudiantes agregados en las unidades de análisis del nivel 2 (profesores, clases) están relacionados entre sí y no son independientes. Este modelo servirá como referente para *evaluar la bondad de ajuste de modelos condicionales más complejos*.

Ahora bien, cuando el investigador está interesado en la comparación de modelos, durante esta primera etapa, es habitual comparar entre dos modelos, para lo cual se deben de especificar dos modelos alternativos: un modelo restringido, que contiene un número básico de parámetros y un modelo aumentado (ampliado), que contiene uno o más parámetros adicionales respecto del modelo restringido.

Etapa 2: Ajuste

Este segundo proceso de denomina *ajuste global*, dado que se realiza sobre un modelo estadístico particular. El objetivo aquí es comparar los datos empíricos con el valor esperado del modelo. Tal comparación se realiza de modo que se minimice algún criterio estadístico, y se evalúa mediante alguna medida de discrepancia, siendo la más común la desvianza (deviance).

Continuando con el ejemplo, el investigador tiene que ajustar el modelo correspondiente al nivel 2 con el fin de conocer en qué medida las variables del contexto (clase, profesor) explican la variable dependiente. Seguidamente, se ajustará el modelo correspondiente a las variables del estudiante (nivel 1). Esto es, al modelo ajustado en el

paso anterior se le añaden las variables que se consideran importantes para explicar la variable dependiente del nivel 1. De esta manera, se conoce el grado en que las variables del estudiante predicen la variable dependiente. Las variables que se consideran importantes del nivel 2 se pueden introducir de una en una o todas a la vez a modo de bloque. En cualquier caso, se debe decidir bien qué variables deben quedar en el modelo utilizando criterios de ajuste. En este punto el modelo es todavía provisional y aún sería posible optimizarlo añadiendo efectos de interacción entre las variables de ambos niveles, por tanto, se debe de proceder con el estudio del grado de interacción existente entre las variables del nivel 1 y del nivel 2. Este paso es una elaboración de un modelo capaz de incorporar todas las variables seleccionadas en el modelo teórico y que han resultado significativas. Repárese que, de lo expuesto, se aprecia rápidamente que estimar un modelo multinivel equivale a estimar un modelo combinado o mixto. Pues, aunque se pueden formular modelos separados para cada nivel, dichos modelos están conectados estadísticamente.

Por otro lado, cuando la motivación del investigador es la comparación de dos modelos, siendo uno de ellos parte del otro, del que difiere en uno (o más) parámetros, el proceso de ajuste es el denominado *ajuste condicional*. Dicho ajuste permite concluir si el modelo más simple (el modelo restringido, que cuenta con menos parámetros) representa una discrepancia significativa respecto del modelo más complejo (el modelo ampliado, que cuenta con mayor número de parámetros). El proceso es similar a una prueba de hipótesis nula, que asume que la diferencia entre el modelo ampliado y el restringido es cero, y la prueba de razón de verosimilitud (LRT) permite decidir entre uno de los dos modelos. En estos casos pueden emplearse también algún *criterio de ajuste estadístico*, como el Criterio de Información Bayesiano o el Criterio de Información de Akaike, entre otros. En general, dados dos modelos bien ajustados, uno de los cuales es parte del otro, es mejor aquel que tiene el más bajo valor del criterio de ajuste estadístico utilizado. Determinar la bondad de ajuste del modelo elegido durante el proceso de modelado es clave (Vallejo et al., 2010), por ello todas las precauciones que se tomen durante esta etapa son pocas.

Etapa 3: Evaluación

En esta tercera etapa del proceso, y una vez aceptado un modelo teórico como representación de los datos observados, se precisa determinar si tal modelo es válido practicando una evaluación concienzuda de sus propiedades. Por lo tanto, se ha de verificar el cumplimiento de los supuestos.

En este sentido, el investigador del ejemplo desarrollado debe evaluar los principales supuestos que recaen sobre el error del modelo y su certificación se realiza mediante el análisis de residuos. Se comprueba, entonces, que el error tiene media nula y varianza constante. A su vez, es igualmente necesario confirmar que los componentes aleatorios son ortogonales.

Al final de esta etapa de evaluación de supuestos, sólo quedaría por comprobar que el error tiene una distribución normal para que los resultados puedan ser inferidos de la muestra a la población.

Etapa 4: Interpretación

Para concluir, si el modelo no presenta problemas que recomienden su rechazo, en la cuarta etapa del proceso de modelado estadístico hay que *interpretar* el modelo, comprobando previamente que se cumple con tres criterios básicos que articulan tal proceso, a saber: *la parsimonia* (el modelo aceptado debe ser suficientemente simple para describir con relativa aproximación el complejo mecanismo subyacente que produjo la respuesta), *la bondad de ajuste* (para que resulte óptimo, el modelo aceptado no debe diferir significativamente del modelo saturado o global y no debe de haber ningún otro modelo más apropiado con el mismo o menor número de parámetros -ajuste condicional-) y *la integración teórica* (el modelo aceptado debe tener significado desde el punto de vista de la teoría que lo generó). Así las cosas, y en consonancia con el ejemplo ilustrado el investigador puede verificar cuánta varianza del primer nivel y del segundo es explicada por el modelo.

No obstante, hay que tener presente que durante el proceso de modelado se han estimado parámetros, contrastado hipótesis, y tomado decisiones sobre qué modelo se ajusta mejor. Veamos a continuación de qué modo.

4.2. La Estimación de los Parámetros (β, u y V(θ))

El modelo definido en la Ecuación [11] necesita estimar tres tipos de parámetros diferentes: los coeficientes de efectos fijos $\boldsymbol{\beta}$, la matriz de covarianza de los efectos aleatorios (**G**) y la matriz de covarianza de los términos de error (**R**). Cuando las matrices **G** y **R** son conocidas, los estimadores estándar $\boldsymbol{\beta}$ y **u** pueden ser obtenidos resolviendo las ecuaciones del conocido modelo mixto de Henderson (1975),

$$\begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z}+\hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} \qquad [13]$$

En concreto, la solución a dicho modelo se puede escribir como sigue:

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= [\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}]^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{y} \\ \hat{\mathbf{u}}(\boldsymbol{\theta}) &= \mathbf{G}\,\mathbf{Z}'\mathbf{V}(\boldsymbol{\theta})^{-1}[\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})], \end{aligned} \qquad [14]$$

donde en [14], $\hat{\boldsymbol{\beta}}$ representa el mejor estimador lineal insesgado de los parámetros fijos del modelo [BLUE *(Best Linear Unbiased Estimator)*] y $\hat{\mathbf{u}}$, el mejor predictor lineal insesgado [BLUP *(Best Linear Unbiased Predictor)*] de los efectos aleatorios (MacCulloch & Searle, 2001), también referido como estimador empírico de Bayes o estimador encogido.

A su vez, las matrices de varianza-covarianza aproximadas de los estimadores correspondientes son:

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= [\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}]^{-} \\ V(\hat{\mathbf{u}}) &= \mathbf{G}-\mathbf{G}\,\mathbf{Z}'\mathbf{P}\,\mathbf{Z}\,\mathbf{G}, \end{aligned} \qquad [15]$$

donde en [15,] el signo menos indica que una inversa generalizada es requerida si **X** no tiene rango pleno y $\mathbf{P}=\mathbf{V}(\boldsymbol{\theta})^{-1}-\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}]^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}$. El vector $\boldsymbol{\theta}$ contiene los elementos únicos de **G** y los parámetros en **R**.

La ecuación [14] proporciona la maquinaría para probar las hipótesis acerca de los efectos fijos y aleatorios. Sin embargo, para obtener los estimadores $\hat{\boldsymbol{\beta}}$ y $\hat{\mathbf{u}}$ se requiere conocer $\mathbf{V}(\boldsymbol{\theta})$. Debido a que en la práctica las matrices **G** y **R** casi nunca resultan conocidas, el investigador se ve obligado a seguir otro enfoque.

Si el diseño de investigación está equilibrado, se pueden utilizar procedimientos algebraicos basados en el método de los momentos (Searle, Casella, & McCulloch, 1992).

Sin embargo, el tradicional método de los momentos consistente en resolver sistemas de ecuaciones simultáneas relacionando los valores esperados con los observados, tiene difícil acomodo cuando se usan modelos multinivel, debido a que en los ámbitos aplicados las unidades de muestreo suelen estar anidadas en grupos no equilibrados con matrices de dispersión parametrizadas arbitrariamente.

Cuando el diseño de investigación está desequilibrado, los componentes de la matriz $\mathbf{V}(\boldsymbol{\theta})$ se estiman iterativamente mediante procedimientos numéricos. Por regla general, estos procedimientos están basados en técnicas de estimación de máxima verosimilitud (ML), o de máxima verosimilitud restringida (REML) para evitar obtener estimaciones sesgadas. Los estimadores ML de $\boldsymbol{\theta}$ son obtenidos maximizando el logaritmo natural de la función de verosimilitud correspondiente a la densidad del vector $\mathbf{y}$ para $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$, donde:

$$l_{\mathrm{c}}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\tfrac{1}{2}[(\mathrm{n}\log(2\pi) + \log|\mathbf{V}(\boldsymbol{\theta})| + \mathbf{r}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{r}] \qquad [16]$$

Donde $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Si n es pequeño, más que estimar los componentes de la varianza desde la verosimilitud global, puede interesar maximizar la parte de la verosimilitud que es invariante de los efectos fijos del modelo mediante el método REML. En concreto, de acuerdo con derivaciones efectuadas por Harville (1977), maximizando el logaritmo de la función de verosimilitud:

$$l_{r}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\tfrac{1}{2}[(\mathrm{n}-\mathrm{h})\log(2\pi) + \log|\mathbf{V}(\boldsymbol{\theta})| + \log|\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}| + \mathbf{r}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{r}] \qquad [17]$$

En la actualidad existen varios procedimientos que permiten calcular los estimadores ML o REML de los parámetros, $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$, del modelo multinivel (ver Dedrick et al., 2009; De Leeuw & Meijer, 2008; Golstein 2003). Sin embargo, bajo el supuesto de normalidad multivariada los parámetros son usualmente obtenidos mediante el algoritmo de Newton-Raphson (NR) o el algoritmo Expectación-Maximización (EM) descrito por Dempster, Laird y Rubin (1977) o incluso, un híbrido de los dos anteriores (Pinheiro & Bates, 2000). Aunque existen algunas razones para preferir el algoritmo NR al EM. De acuerdo con Lindstrom y Bates (1988), el algoritmo NR requiere un menor número de iteraciones para converger que el algoritmo EM. Otra ventaja del algoritmo NR sobre el EM, reside en que el primero permite computar los errores estándar de los elementos de $\boldsymbol{\theta}$ desde la inversa de la matriz de información empírica (inversa de la matriz hessiana

cambiada de signo). Cuando se utiliza el algoritmo EM no se calcula dicha matriz y, por ende, los errores estándar para los elementos de $\boldsymbol{\theta}$ (Jennrich y Schluchter, 1986).

Otro procedimiento disponible para estimar los elementos de la matriz $\mathbf{V}(\boldsymbol{\theta})$ se basa en el enfoque Bayesiano. A diferencia del enfoque Clásico o Frecuentista, este enfoque no requiere grandes tamaños de muestra y se puede implementar usando métodos de Monte Carlo basados en Cadenas de Markov (para más información véase Hox, 2010). No obstante, como pone de relieve Van der Leeden, (1998b) el esfuerzo computacional exigido por este procedimiento puede ser considerable cuando los modelos son complejos y los tamaños de muestra de los diferentes niveles muy elevados. Detalles y extensiones de la técnica pueden descubrirse en numerosos textos, incluyendo De Leeuw y Meijer (2008), Hox y Roberts (2011) así como Snijders y Bosker (2012).

4.3. El Contraste de Hipótesis en el MLM

Una cosa es la estimación de los parámetros fijos, aleatorios y de covarianza, y otra distinta es el contraste de hipótesis de esos parámetros. Para contrastar los efectos fijos y aleatorios se utilizan los mismos procedimientos. Las hipótesis nulas correspondientes a contrastes simples se pueden probar mediante F o t (calculando la razón entre los estimadores ML o REML y sus respectivos errores estándar). Las hipótesis nulas correspondientes a contrastes múltiples se pueden probar calculando el estadístico W de Wald. Cuando los datos no están equilibrados, como suele ser usualmente el caso, y el número de unidades del nivel 2 es pequeño, estas pruebas pueden ofrecer resultados liberales. Para solucionar este problema Fai y Cornelius (1996) y, también, Kenward y Roger (1997, 2009) han desarrollado dos procedimientos diferentes. Además, Keselman, Algina, Kowalchuck y Wolfinger (1999) han demostrado que la prueba de Wald es liberal cuando los datos violan el supuesto de la esfericidad multimuestral, y estos autores han evaluado los efectos de aplicar el procedimiento de Satterthwaite bajo el MLM.

Para poner a prueba las hipótesis nulas correspondientes a los componentes de varianza (recuérdese, **G** y **R**) se puede utilizar la prueba de razón de verosimilitud (LRT) o el estadístico Z de Wald (se obtiene dividiendo cada uno de los parámetros estimados mediante ML o REML por su correspondiente error estándar). Este estadístico únicamente es exacto asintóticamente (Wolfinger, 1996), esto es, para muestras

relativamente grandes. Así pues, cuando el número de unidades del segundo nivel es reducido, hay que ser muy prudente en la interpretación de los resultados.

APORTACIONES DE LOS MODELOS MULTINIVEL

Llegados a este punto, es conveniente hacer una recapitulación de las aportaciones más relevantes de los Modelos Multinivel al Campo de la Educación y, cómo no, al estudio de la Eficacia Escolar. Esta metodología ha supuesto una verdadera revolución, permitiendo grandes contribuciones tanto de carácter más sustantivo como de carácter claramente metodológico (Hox & Kreft, 1994; Núñez, Vallejo, Rosário, Tuero-Herrero, & Valle, en prensa).

5.1. Contribuciones Sustantivas

La aportación de carácter sustantivo más destacada que presentan los Modelos Jerárquicos es que tienen en cuenta las diferencias en el contexto. En este sentido, se entiende que los estudiantes producen diferencias al igual que los contextos, luego se precisan modelos que no reduzcan a los alumnos a las agregaciones estadísticas y que no limiten los contextos a vagas generalizaciones. Así, en los estudios sobre Eficacia Escolar se necesita considerar simultáneamente las variables de los estudiantes (nivel 1), tales como la situación socioeconómica de las familias, etc. y las variables de escuela (nivel 2) como el clima del centro, la titularidad... Esta consideración de las diferencias contextuales se concreta en:

- *Consideración de la Heterogeneidad.* Los efectos de los contextos pueden ser potencialmente complejos, con relaciones que fluctúan en diferentes sentidos. Por tanto, se hace necesario estudiar quién es un alumno en relación con el sitio en el que está.

- *Interacción entre los Estudiantes y sus Contextos.* Hay que tener muy presente la posibilidad de que un alumno interactúe con su contexto próximo de forma distinta a la que lo hace su grupo social de referencia. Dicho en otras palabras, las diferencias entre los contextos deben ser examinadas en relación con las características de los estudiantes en combinación con las características sociales de los lugares.

- *Pluralidad de Contextos.* Es altamente probable que no exista un solo contexto. De hecho, en el caso del rendimiento académico, los resultados de los estudiantes pueden estar influenciados por la escuela a la que asisten, pero también por su contexto familiar.

Otra aportación sustantiva a considerar es que los Modelos Multinivel permiten analizar simultáneamente contextos y heterogeneidad individual, dado que no sólo se deben considerar las diferencias entre contextos. Así, siguiendo a Coleman et al., (1966), los estudiantes de nivel sociocultural más bajo no sólo pueden diferir en la media de rendimiento académico, sino que también pueden ser más o menos versátiles en sus puntuaciones.

Un último aspecto a tener presente en las aportaciones de carácter sustantivo es que permiten combinar la investigación intensiva con la extensiva, esto es, la cualidad y la cantidad. Los comportamientos y las acciones de los estudiantes tienen estos dos componentes, uno cualitativo (ocurre) y otro cuantitativo (cuán frecuente sea). Ambos elementos deben considerarse simultáneamente (por ejemplo, el fracaso académico es muy bajo en algunas escuelas, pero aquellos alumnos que fracasan lo hacen de manera estrepitosa). La investigación extensiva permite de esta manera identificar patrones, y a su vez, permite reconocer grupos específicos que necesitan estudios más intensivos.

5.2. Contribuciones Técnicas

La característica más relevante de los Modelos Jerárquicos es que aportan un entorno natural dentro del cual se pueden comparar teorías sobre relaciones estructurales entre variables en cada uno de los niveles en los que se organizan los datos. Estos modelos ofrecen una estructura de análisis dentro de la cual se pueden reconocer los distintos niveles en los que se articulan los datos, al estar representados cada uno con su propio submodelo (Draper, 1995). Cada submodelo, a su vez, expresa la relación entre las variables dentro de un determinado nivel y especifica cómo las variables de un nivel influyen en las relaciones que se establecen en los otros niveles.

En esta Tesis ya se han analizado las técnicas metodológicas más tradicionales para la investigación en la Eficacia Escolar y, por tanto, ha quedado claro que dichas técnicas conllevaban una serie de inconvenientes cuando abordaban datos con una estructura anidada (Bryk & Raudenbush, 1992; Hox & Kreft, 1994). Sin embargo, la

metodología multinivel proporciona una serie de ventajas, como las que a continuación siguen.

- *Mejoran la Estimación de los Efectos Entre las Unidades Individuales.* En este sentido, desarrollan una estimación mejorada del modelo de regresión para una escuela apoyándose en las estimaciones semejantes que existen para otras escuelas.

- *Permiten Formular y Probar Hipótesis sobre los Efectos Cruzados Entre Niveles.* Así, por ejemplo, se puede estudiar la relación entre la titularidad de la escuela y el rendimiento académico de los estudiantes en función de su nivel sociocultural. Sin duda, la oportunidad que nos brindan los Modelos Multinivel de estudiar las interacciones entre las variables definidas en distintos niveles de una jerarquía es una cuestión muy relevante, dado que de no considerarse, se pueden cometer inferencias erróneas e inadecuadas (por ejemplo, usar datos del nivel de contexto para inferencias individuales y que las variables puedan tener distintos significados en niveles diferentes).

- *Realizan la Partición de Componentes de Varianza y Covarianza Entre Niveles.* Así, permiten descomponer las correlaciones entre las variables relacionadas con los estudiantes en componentes Intra e Inter centros).

- *Estiman Adecuadamente los Parámetros.* Los Modelos Multinivel posibilitan una estimación apropiada y adecuada de los parámetros en presencia de correlaciones intragrupos (autocorrelación). Las observaciones dentro de un grupo están muy próximas en el tiempo o incluso en el espacio, por tanto, es esperable que sean más similares que las observaciones de distintos grupos (dada la no asignación aleatoria de los estudiantes a los grupos, y dado el conjunto de estímulos y experiencias compartidas por los alumnos de un mismo grupo). La cantidad de covariación entre las observaciones que comparten el mismo contexto, suele expresarse por medio de la correlación intraclase. Por ello, cuando se emplean los estadísticos de contraste ordinario, al considerar al estudiante como unidad de análisis, violan el supuesto de independencia de los errores. Incluso valores de correlación intraclase muy pequeños conllevan Errores de Tipo I mayores que el nivel del alpha nominal. De hecho, la no dependencia de las observaciones y la heterogeneidad no son fallos de nuestros datos, sino de

sus características, por tanto, con la metodología multinivel son esperados y modelados. De igual forma, estos modelos ofrecen una estructura explícita dentro de la cual se puede expresar la similitud de los juicios destinados a combinar la información entre unidades (niveles diferentes) para realizar mejores estimaciones y predicciones a partir de las observaciones.

▪ *Incorporan Efectos Aleatorios.* Los Modelos Jerárquicos asumen un muestreo aleatorio de estudiantes en contextos también aleatorios. De esta manera, los análisis de los modelos multinivel pueden incorporar efectos aleatorios (recuérdese que el modelo de regresión clásico solo asume coeficientes fijos y por tanto, sus inferencias afectan únicamente a los tratamientos incluidos en el estudio).

5.3. Aplicaciones de los Modelos Multinivel a la Investigación Educativa

Los Modelos Multinivel abren una ingente cantidad de posibilidades para la investigación, sobre todo en el ámbito que nos compete, el educativo. Así, además de las múltiples ventajas que se acaban de mencionar, permiten una buena cantidad de aportaciones al estudio de la Eficacia Escolar entre las destacan las siguientes:

▪ *Estimar el Efecto Escolar de cada Centro mediante la Metodología de Valor Añadido en Educación del centro.* Es decir, permiten saber qué aporta el centro exactamente al desarrollo del estudiante descontando variables tales como el nivel sociocultural de la familia, el rendimiento previo del alumno, etc.

▪ *Tener una Estimación de los Efectos Escolares Diferenciales.* De tal forma que, se conozca el grado de equidad de los centros en el fomento del desarrollo de todos los estudiantes.

▪ *Conocer y Estimar con Precisión la Magnitud de la Aportación de las Variables de Clase, Escuela o Contexto sobre la Variable Producto del Estudiante.* Así, se trabaja con múltiples niveles de análisis de forma simultánea y se analizan simultáneamente contextos y heterogeneidad individual.

▪ *Analizar la Interacción Entre Variables de Niveles Diferentes.* Esto permite, por ejemplo, conocer si determinadas características del docente, el aula o la escuela, inciden de manera distinta en los diferentes grupos de alumnos.

En definitiva, para realizar investigación de calidad, aportando resultados que contribuyan a un mejor conocimiento de las variables que influyen en la mejora de la Eficacia Escolar, es imprescindible que se utilicen los recursos metodológicos más adecuados. El camino conveniente en estos estudios se llama, sin duda, Modelos Multinivel. Ahora bien, nadie dijo que el camino a seguir fuera fácil, por lo que habrá que tener suma precaución en la selección del mejor modelo multinivel a considerar para explicar nuestros datos.

DESAFÍOS EN LOS MODELOS MULTINIVEL

6.1. La Problemática de la Selección de Modelos

Los Modelos Lineales Mixtos en la investigación educativa permiten analizar datos de carácter transversal y longitudinal y, de hecho, es en éstos últimos donde tienen una mayor importancia. Con datos temporales, estos modelos, posibilitan ajustar y realizar inferencias acerca de la estructura de medias (efectos fijos, para describir el promedio de las respuestas dadas en función del tiempo) y modelar la estructura de covarianza (efectos aleatorios, para describir la variación entre las medidas repetidas de los estudiantes). Así, para mejorar la calidad de las inferencias obtenidas con estos modelos, es decir, con el fin de obtener pruebas más válidas y potentes de los parámetros de efectos fijos, es muy conveniente elegir un modelo con una estructura de medias y de covarianzas apropiada (Littell, Pendergast, & Natarajan, 2000; Fitzmaurice, Laird, & Ware, 2011). Justamente, cuando se selecciona un modelo apropiado, la precisión y la exactitud de la estimación de los parámetros mejora. En este sentido, se han desarrollado programas de software estadístico para facilitar el modelado de la matriz de covarianza. Por ejemplo, el Proc Mixed de SAS permite ajustar y comparar modelos variados (de simetría compuesta, de esfericidad, autorregresivos, de media móvil, etc.), además permite especificar estructuras de covarianza heterogéneas dentro y a través de los grupos (evitando la equicorrelación de las observaciones y la homogeneidad de las matrices de dispersión).

Sin embargo, una de las principales dificultades con las que se encuentra el investigador radica en la **selección del modelo**, pues una decisión no correcta puede llevar a conclusiones erróneas. Diversos autores (Ato, Losilla, Navarro, Palmer, & Rodrigo, 2000; Claeskens & Hjort, 2008; García, 1996, Vallejo et al., 2010) coinciden en que para una misma realidad concreta no hay un único modelo válido sino todo un conjunto de modelos candidatos. Por ello, el objetivo que se persigue con el modelado estadístico es el de encontrar el modelo más *"óptimo"*, entendiéndose éste como aquel modelo que explica la máxima variabilidad con el mínimo de parámetros posibles. Modelar la complejidad de un fenómeno pasa por llevar a cabo un proceso de

simplificación, evitando de esta manera que el excesivo número de parámetros dificulte el análisis. Este proceso de simplificación se guía por el principio de parsimonia, que se basa en el postulado conocido como *"la Navaja de Occam",* según el cual a igualdad de condiciones es preferible un modelo simple a un modelo complejo. Así, el principio de parsimonia proporciona una pauta importante para la simplificación de los modelos, anclada en la ponderación del criterio de máximo ajuste por el criterio de mínimo número de parámetros.

Como se puede apreciar en la literatura estadística científica, y como previamente ya se ha mencionado, para ajustar distintos modelos a un mismo conjunto de datos, es necesario utilizar criterios para la comparación de los ajustes y, por lo tanto, para la selección de un modelo. Para comparar modelos el criterio más utilizado ha sido el de la prueba de razón de verosimilitudes (LRT), bien con la desvianza obtenida a partir de la función de máxima verosimilitud completa (ML), o de máxima verosimilitud restringida (REML), según se trate de elegir entre modelos con idéntica estructura de covarianza o de medias (Kreft & de Leeuw, 1998). Sin embargo, la LRT para la selección de modelos presenta alguna que otra limitación (véase detalladamente Hamaker, Van Hattum, Kuiper, & Hoijtink, 2011). En este sentido, para paliar en la medida de lo posible todas estas limitaciones hoy se recomienda el uso de criterios de selección de modelos incluyendo criterios de información, criterios predictivos, y técnicas gráficas; si bien de todos ellos son los criterios de información los que más aceptación tienen (Gurka & Edwards, 2008; Vallejo et al., en prensa).

Los Criterios de Información suponen un enfoque distinto y más reciente llevado a cabo mediante el método de máxima verosimilitud. Estos criterios en mayor o menor medida, penalizan el logaritmo de la función de verosimilitud por el número de parámetros (Lee & Ghosh, 2009). La mayoría de las veces desde la formulación marginal del modelo (se ignoran de manera explícita, los efectos aleatorios a la hora de modelar la variación de los datos multinivel) y eligen aquel modelo que minimiza el valor de los mismos. Algunos de los Criterios de Información comúnmente más usados (Vallejo & Lozano, 2006) son: el Criterio de Información de Akaike, en adelante **AIC**, (Akaike, 1974); el Criterio de Información Bayesiano de Schwarz, en adelante **BIC** (Schwarz, 1978); el Criterio de Información de Hannan y Quinn, en adelante **HQIC** (Hannan & Quinn, 1979) el Criterio de Información de Akaike Consistente, en adelante, **CAIC** (Bozdogan, 1987) el Criterio de Información Akaike Corregido, en adelante **AICc**

(Hurvich & Tsai, 1989), así como diversas versiones que surgen a partir de estos criterios. El uso ascendente de los principales Criterios de Información viene marcado esencialmente por dos aspectos de suma importancia, por un lado estos criterios gozan de cierta flexibilidad y pueden comparar modelos anidados como modelos no anidados (Gurka, 2006; Vallejo et al., 2003, 2010), y por otro lado algunos de los criterios más comúnmente utilizados, como el AIC y el BIC, ya vienen implementados en la mayor parte de los programas estadísticos que ajustan modelos mixtos (Gómez, Torres, García, & Navarro, 2012). De hecho, en estos momentos se pueden utilizar de forma efectiva los distintos Criterios de Información para seleccionar el mejor modelo de entre todos los posibles. Estos análisis pueden llevarse a cabo de una manera relativamente fácil con varios paquetes estadísticos de carácter generalista incluyendo los módulos PROC MIXED y PROC GLIMMIX de SAS y la función *lme* de R y Splus, así como mediante los comandos *mixed* y *xtmixed* de SPSS y STATA, respectivamente. No obstante, los dos procedimientos de SAS tienen la ventaja de ofrecer un menú más amplio y de incorporar la solución introducida por Kenward y Roger (1997, 2009) para realizar inferencias con muestras pequeñas.

Sin embargo, a pesar de la gran variedad de estrategias para la bondad de ajuste, actualmente no existe un consenso sobre lo que constituye la herramienta más adecuada para elegir el mejor modelo multinivel. No obstante, en términos generales algunas investigaciones apuntan en la dirección, que independientemente del criterio de selección utilizado, la elección del modelo óptimo prospera cuando el tamaño de muestra aumenta y la complejidad del modelo disminuye. Por ello es necesario el estudio pormenorizado de la fase de ajuste del proceso de modelado estadístico multinivel y de los criterios existentes para dicho ajuste. En este sentido, se presentan a continuación 4 estudios realizados con el propósito de brindar a los investigadores, interesados en el ámbito educativo, una mayor comprensión de las herramientas más útiles para el desarrollo de un buen ajuste y, por lo tanto, de una mejor selección del modelo multinivel en referencia a los datos observados.

Ejecución de la Investigación

Todo investigador debe

pensar durante la realización de su estudio

OBJETIVO GENERAL

Los Modelos Lineales Jerárquicos forman una clase de modelos que permiten la modelación en una gran variedad de situaciones en las cuales se tienen datos que presentan una estructura jerárquica. Como se ha podido constatar en páginas anteriores, aunque tienen una gran historia su proliferación se remonta a finales de la década de los ochenta, cuando los avances en las técnicas estadísticas y el desarrollo de los programas de software hicieron posible su aplicación. De hecho, el desarrollo de estos modelos no habría sido posible sin la realización de investigaciones educativas a gran escala mediante pruebas estandarizadas, nacionales e internacionales, sin la disponibilidad de ordenadores para la creación y almacenamiento de bases de datos y, como no, sin los desarrollos de software estadístico que permiten implementar los complejos procesos de estimación.

Así las cosas, la modelización lineal multinivel permite la construcción empírica de modelos (los modelos ajustados a los datos), con lo que es posible desarrollar descripciones y explicaciones, y probar hipótesis respecto al comportamiento de los fenómenos que nos interesan en el área de la Educación. La aplicación correcta de la modelación implica el postulado realista de ecuaciones que establecen relaciones causales para describir el fenómeno bajo estudio. Pero esta circunstancia enfrenta al investigador al reto de considerar variables que miden distintos niveles de agregación de las unidades de estudio, además de que debe considerar la estructura de anidamientos y entrelazamientos de la muestra de la que se obtienen los datos (Vallejo, et al., 2008). Esta problemática ha trazado una línea de desarrollo para la modelación que se expresa principalmente en dos vertientes, a saber: modelos cada vez más generales y más complejos, y métodos de estimación y herramientas de evaluación para la selección de los modelos más parsimoniosos adecuados a cada situación (Ojeda & Velasco, 2012). No obstante es esta última la que tiene una mayor trascendencia para la investigación educativa dado que realizar la selección del modelo óptimo resulta sustancialmente crucial para interpretar adecuadamente los datos. Sin embargo cuando para una misma evidencia muestral existen modelos alternativos surge el problema de la elección (Vallejo et al., 2010, 2011b). *¿Cuál es el mejor modelo de todas las alternativas formuladas?*

¿*Tiene sentido seleccionar un modelo en función del uso posterior que se vaya a dar al mismo?* Estas y otras preguntas son formuladas.

Los distintos métodos de selección de modelos han sido objeto de comparación en la literatura sin que exista una postura unánime sobre cuál es la mejor forma de seleccionar el modelo óptimo. Gran parte de esta controversia yace en el hecho de que no todos los criterios han sido definidos con el mismo fin, esto es, para todos los autores la idea de *"mejor modelo"* no es igual, no es la misma. Pese a ello, sí parece claro que hay algunas propiedades deseables que todo criterio de selección debería de cumplir, es decir, existen algunas formas objetivas de comparar los distintos criterios de selección y concluir cuál de ellos es el mejor, al menos en esa parcela. En este sentido, en la literatura científica estadística se ha comparado el comportamiento de los criterios de selección cuando cambia el tamaño muestral (Geweke & Meese, 1981), la estructura del proceso que generó los datos (Koehler & Murphree, 1988), el grado de colinealidad entre las variables o la distribución del término de error (Mills & Prasad, 1992). Otras investigaciones también se han ocupado de estudiar si el número de modelos influye en la selección (García, 1996). Más recientemente, se han desarrollado estudios más exhaustivos y pormenorizados en los que el interés radicó en la selección del mejor modelo mixto bajo distintos escenarios. Por ejemplo, los trabajos de Gurka (2006) y Wang y Schaalje (2009) dieron cuenta de la selección del mejor modelo correcto de medias dada una estructura de covarianza particular; los estudios de Ferron, Dailey y Yi (2002), y Vallejo, Ato y Valdés (2008) mostraron la capacidad de seleccionar el modelo correcto de covarianza cuando la estructura de medias del modelo era conocida; y la capacidad de selección de la correcta estructura de medias y de covarianza en los modelos también fue examinada (Gurka, 2006). Y es precisamente a colación de las investigaciones realizadas en los últimos años lo que ha propiciado la motivación de los estudios que conforman esta Tesis, en especial los estudios de Gurka (2006) por ser unos de los pioneros en la introducción de distintos criterios de información para la selección de los modelos (no quedándose únicamente con el desempeño de los criterios usualmente adoptados: AIC y BIC). Por consiguiente, el objetivo principal de las investigaciones que aquí se presentan es aportar una mayor comprensión del desempeño de los Criterios de Información, implementados en el módulo PROC MIXED del programa SAS, en diseños de corte longitudinal y en diseños de corte transversal. Dicho lo cual, a continuación se muestra un desglose de los estudios realizados para dar cuenta del objetivo general perseguido.

7.1. Objetivos Específicos

7.1.1. Objetivo 1

En las investigaciones educativas suele ser habitual el estudio del cambio en función del tiempo, por cuya razón se obtienen datos longitudinales de una muestra dada de estudiantes que es medida repetidas veces en la misma variable de respuesta. En lo concerniente a estos datos, todo modelo de análisis que pretenda dar verdadera cuenta de lo que realmente interesa, debe afrontar como uno de los retos la posible correlación entre las medidas repetidas. Las correlaciones entre los datos u observaciones repetidas del mismo estudiante quedan plasmadas en la estructura de covarianza. En este sentido, el uso del enfoque basado en los modelos mixtos va a permitir modelar de forma ajustada la estructura de covarianza, proporcionando así errores estándar más válidos.

La mayoría de estudios realizados hasta estos momentos se han centrado en ajustar la matriz de dispersión usando criterios de selección de modelos para elegir entre estructuras de covarianza no anidadas. Por lo tanto, el **Primer Objetivo** de esta Tesis consiste en llevar a cabo un estudio de simulación Monte Carlo para determinar cuán efectivos son los Criterios de Información: AIC, AICC, BIC, CAIC y HQIC para revelar el verdadero proceso generador de los datos en una familia de modelos anidados enmarcados en un contexto longitudinal. Estos criterios serán evaluados bajo estimación ML y REML cuando se manipulan diversas estructuras de medias y/o de covarianzas. Además, para proporcionar un punto de referencia para la comparación también será utilizado el criterio de ajuste condicional LRT. Con el objeto de arrojar un poco más de luz al objetivo planteado, serán usados sendos diseños crossover en los que se violan separada y conjuntamente los supuestos de normalidad de los datos y de homogeneidad de las matrices de dispersión.

Paper I

Psicothema 2010. Vol. 22, nº 2, pp. 323-333
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional

Guillermo Vallejo Seco, Jaime Arnau Gras*, Roser Bono Cabré*,
Paula Fernández García y Ellián Tuero Herrero
Universidad de Oviedo y ** Universidad de Barcelona

Un marco teórico potente resulta clave para especificar el modelo mixto que explica mejor la variabilidad de datos longitudinales. A falta de teoría, la mayoría de las investigaciones realizadas hasta la fecha, se ha centrado en ajustar la matriz de dispersión usando criterios de selección de modelos para elegir entre estructuras de covarianza no anidadas. En este trabajo, comparamos el desempeño del estadístico razón de verosimilitud (LRT) condicional y de varias versiones de los criterios de información para seleccionar estructuras de medias y/o de covarianzas anidadas, asumiendo conocido el verdadero proceso generador de datos. Los resultados numéricos indican que los criterios de información eficientes funcionaban mejor que sus homólogos consistentes cuando las matrices de dispersión usadas en la generación eran complejas y peor cuando eran simples. Globalmente, el desempeño del LRT condicional basado en el estimador de máxima verosimilitud completa (FML) era superior al resto de los criterios examinados. Sin embargo, el desempeño era inferior cuando se basaba en el estimador máxima verosimilitud restringida (REML). También encontramos que la estrategia sugerida en la literatura estadística de usar el estimador REML para seleccionar la estructura de covarianza y el estimador FML para seleccionar la estructura de medias debería ser evitada.

Nested model selection for longitudinal data using information criteria and the conditional adjustment strategy. Knowledge of the subject matter plays a vital role when attempting to choose the best possible linear mixed model to analyze longitudinal data. To date, in the absence of strong theory, much of the work has focused on modeling the covariance matrix by comparing non-nested models using selection criteria. In this paper, we compare the performance of conditional likelihood ratio test (LRT) and several versions of information criteria for selecting nested mean structures and/or nested covariance structures, assuming that the true data-generating processes are known. Simulation results indicate that the efficient criteria performed better than their consistent counterparts when covariance structures used in the data generation were complex, and worse when structures were simple. The conditional LRT under full maximum likelihood (FML) estimation was better overall than the other criteria in terms of selection performance. However, under restricted maximum likelihood (REML), estimation was inferior. We also find that the strategy suggested in the statistical literature of using REML for covariance structure selection, and FML for mean structure selection may be misleading.

Actualmente, cada vez son más las disciplinas que utilizan enfoques basados en la teoría del modelo lineal mixto para analizar datos que presentan una estructura jerarquizada (Vallejo, Arnau y Bono, 2008a). Además del gran desarrollo teórico que estos modelos han experimentado en las tres últimas décadas, el establecimiento definitivo de los mismos se ha visto favorecido por la incorporación de procedimientos analíticos específicos dentro de los principales paquetes estadísticos profesionales, incluyendo el módulo *Proc Mixed* en SAS, la función *lme* en S-PLUS/R o los comandos *Mixed y xtmixed* en SPSS y STATA, respectivamente. Los modelos lineales mixtos permiten analizar datos de corte longitudinal y transversal, aunque son especialmente útiles cuando se trabaja con datos temporales, ya que permiten ajustar y realizar inferencias acerca de la estructura de medias (como sucede con los clásicos enfoques univariante y multivariante de medidas repetidas) y modelar la estructura de covarianza (en términos de efectos aleatorios y error puro). Mediante este enfoque, más que asumir una matriz de dispersión demasiado parca (p. e., la matriz de simetría compuesta —CS— típica del enfoque univariante) o una complemente general (p. e., la matriz no estructurada —UN— típica del enfoque multivariante), se trata de buscar un equilibrio entre los criterios de flexibilidad y parsimonia o simplicidad científica (Ato y Vallejo, 2007). Al respecto, hay que advertir que si un investigador especifica un modelo excesivamente simple corre el riesgo de efectuar inferencias erróneas, debido a la subestimación de los errores estándar. Si, por el contrario, formula un modelo ex-

Fecha recepción: 12/7/2009 • Fecha aceptación: 26/10/2009
Correspondencia: Guillermo vallejo Seco
Facultad de Psicología
Universidad de Oviedo
33450 Oviedo (Spain)
e-mail: gvallejo@uniovi.es

cesivamente complejo corre el riesgo de efectuar inferencias ineficientes.

Con el fin de mejorar la calidad de las inferencias obtenidas con el enfoque del modelo mixto, resulta crucial modelar dos aspectos diferentes de los datos (Littell, Pendergast y Natarajan, 2000). Por un lado, los efectos fijos usados para describir el promedio de las respuestas en función del tiempo (en adelante, estructura de medias). Y, por otro lado, los efectos aleatorios usados para describir la variación entre las medidas repetidas dentro de los sujetos (en adelante, estructura de covarianza). Cuando se modela de forma efectiva la estructura de covarianza y también la de medias, dado que la forma de primera depende de la elección que se haga de la segunda (Fitzmaurice, Laird y Ware, 2004), se obtienen estimaciones más exactas (con menor sesgo) y precisas (con menor varianza) de los parámetros. Vallejo, Ato y Valdés (2008b) confirman la importancia de identificar el verdadero proceso generador de datos (PGD). En este estudio las tasas de error basadas en el verdadero PGD nunca excedían su valor nominal. Sin embargo, los errores estándar resultaban sesgados cuando se especificaba erróneamente el verdadero PGD.

La selección del modelo óptimo resulta central para interpretar adecuadamente los datos, no obstante, dicho objetivo es difícilmente alcanzable porque para una misma evidencia muestral existen múltiples modelos candidatos (Claeskens y Hjort, 2008). Para facilitar el modelado de la matriz de covarianza, SAS, probablemente el programa más popular y versátil de cuantos existen actualmente (Feng, Zhou, Zhang y Zhang, 2009), y otros programas estadísticos incorporan un completo menú de estructuras. Por ejemplo, *Proc Mixed* permite ajustar y comparar modelos de simetría compuesta, de esfericidad, autorregresivos, de media móvil, autorregresivos e integrados de media móvil, antedependientes y no estructurados (para detalles concretos, véase Zimmerman y Núñez-Antón, 2009). *Proc Mixed* también permite especificar estructuras de covarianza heterogéneas dentro y a través de los grupos, lo cual evita tener que aceptar la equicorrelación de las observaciones y la homogeneidad de las matrices de dispersión.

Existen diversos criterios para determinar la bondad de ajuste del modelo elegido durante el proceso de modelado. Para comparar modelos anidados (uno se puede obtener a partir de otro manipulando parámetros), el criterio más usado es el test de razón de verosimilitudes (LRT) con la desvianza obtenida a partir de la función de máxima verosimilitud completa (FML) o de máxima verosimilitud restringida/residual (REML), según se trate de elegir entre modelos con idéntica estructura de covarianza o de medias (Kreft y de Leeuw, 1998). También se suelen emplear herramientas estadísticas menos formales, tales como el Criterio de Información (IC) de Akaike (AIC), el AIC Corregido (AICC), el AIC Consistente (CAIC), el Criterio de Información Bayesiano (BIC) y el Criterio de Información Hannan-Quinn (HQIC), así como diversas versiones surgidas a partir de estos. Especialmente, los criterios AIC y BIC, por hallarse ambos implementados en la mayor parte de los programas que ajustan modelos mixtos; los programas específicos HLM y MLwiN constituyen una excepción a lo dicho. El origen de los IC es diferente, pero su estructura es similar; de hecho, difieren en el peso que asignan al factor de penalización (Lee y Ghosh, 2009). En mayor o menor medida, todos ellos penalizan el logaritmo de la función de verosimilitud por el número de parámetros, la mayor parte de las veces desde la formulación marginal del modelo (se ignoran explícitamente los efectos aleatorios a la hora modelar la variación de los datos multinivel), y eligen aquel modelo que minimiza el valor de los mismos. Vaida y Blanchard (2005) y Liang, Wu y Zou (2008) ofrecen detalles del comportamiento del criterio AIC usando una formulación jerárquica del modelo.

Otros criterios de selección, tales como el coeficiente de determinación ajustado (R^2_{adj}) el coeficiente de correlación de concordancia (CCC) o la suma de cuadrados residual de predicción (PRESS), han recibido escasa atención. No obstante, en uno de los pocos estudios que han examinado el desempeño de los criterios predictivos (basados en el ajuste de los valores predichos) usando la formulación marginal y jerárquica del modelo, Wang y Schaalje (2009) informan que los criterios (R^2_{adj}) CCC y PRESS no se comportaban mejor que los criterios AIC y BIC. La comparación se hacía entre dos modelos anidados con idéntica estructura de covarianza. Detalles técnicos de los criterios predictivos los proporcionan Orelien y Edwards (2007), Schabenberger (2004) y Vonesh, Chinchilli y Pu (1996).

En función de sus propiedades asintóticas los IC pueden ser clasificados en dos categorías: (a) criterios eficientes, tales como AIC o AICC y (b) criterios consistentes, tales como BIC, CAIC o HQIC. Se dice que un criterio es eficiente si la discrepancia entre el verdadero PGD y el modelo especificado para aproximarlo disminuye conforme aumenta el tamaño muestra. A su vez, se dice que un criterio es consistente si la probabilidad de elegir el modelo correcto aumenta conforme lo hace el tamaño de muestra. Los criterios eficientes parten de la hipótesis de que el verdadero PGD es dimensión infinita y seleccionan el mejor modelo de dimensión finita. En cambio, los consistentes parten de la hipótesis de que el verdadero PGD es dimensión finita y tienden a elegirlo siempre que el tamaño de muestra tienda a infinito. Cuando se apela al concepto de eficiencia asintótica no se asume que el verdadero PGD esté incluido dentro de la familia de modelos investigados. Sin embargo, cuando se apela al concepto de consistencia asintótica está implícita la hipótesis de que verdadero PGD pertenece a la clase de modelos considerados, lo cual puede ser falso.

Los análisis de contenido ponen de relieve que los IC más usados para elegir modelos con idéntica estructura de medias son el AIC y el BIC (Littell et al., 2000). El desempeño de estos criterios ha sido examinado por diversos autores, incluyendo Ferron, Dailey y Yi (2002), Gomez, Schaalje y Fellingham (2005), Keselman, Algina, Kowalchuk y Wolfinger (1998) y Vallejo et al. (2008b). Exceptuando el estudio de Ferron et al. (2002), donde el AIC identificó el verdadero PDG en el 79 % de las veces y el BIC en el 66%, los estudios restantes avalan críticamente la sugerencia efectuada por Littell et al. (2000) de modelar la estructura de covarianza con estos criterios, sobre todo, mediante el BIC. En el estudio de Keselman et al. (1999) el AIC seleccionó la estructura correcta en el 47% de las veces y el BIC en el 35%, en el estudio de Vallejo et al. (2008b) el AIC lo hizo en el 68% de las veces y el BIC el 48%, mientras que en el de Gomez et al. (2005) ambos criterios lo hicieron en el 22% de las veces. Auque el desempeño dependía de las condiciones manipuladas, en todos los estudios se puso de relieve que la selección mejoraba conforme aumentaba el tamaño de muestra y disminuía la complejidad de la matriz.

Un estudio más completo es el llevado a cabo por Gurka (2006). Este investigador examinó el desempeño de los criterios AIC, AICC, BIC y CAIC en términos de seleccionar el modelo de curva de crecimiento correcto bajo diversas condiciones, incluyendo diferentes formas de calcular los criterios y diferentes métodos de estimación de parámetros. Los IC fueron evaluados bajo tres escenarios dife-

rentes en función de su habilidad para: (a) seleccionar la estructura de medias correcta entre tres posibles modelos, dada una matriz CS; (b) seleccionar la estructura de covarianza correcta entre tres efectos aleatorios posibles con la misma estructura de medias; y (c) seleccionar el modelo correcto entre seis modelos que resultaban de combinar tres estructuras de medias con dos de covarianza. Los resultados obtenidos por Gurka muestran, entre otras cosas, que los IC basados en el método REML elegían el verdadero modelo de medias tan bien o mejor que los IC basados en el método FML; lo cual no deja de ser chocante, teniendo en cuenta que en la literatura estadística especializada (Molenberghs y Verbeke, 2001; Littell et al., 2006; Singer y Willet, 2003) se defiende ajustar dicha estructura vía FML exclusivamente. Gurka también halla que el desempeño de los criterios eficientes basados en el estimador REML mejoraba cuando se excluía del mismo el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ (en adelante $REML_2$); en cambio, el desempeño los criterios consistentes mejoraba cuando se mantenía dicho término ($REML_1$). Los resultados globales revelan que los criterios consistentes seleccionaban el verdadero PGD más del 89% de las veces, frente a los eficientes que lo hacían en torno al 81%.

Los hallazgos de Gurka (2006) afectan de lleno al proceso de selección de modelos, dado que ponen de relieve diversas inconsistencias existentes en la literatura estadística y en la documentación de los programas comerciales, tanto en lo referido a los métodos de estimación de parámetros como a las fórmulas usadas para calcular los IC. Conviene resaltar, no obstante, que sus estudios están basados en escenarios excesivamente simples, lo cual limita el alcance de sus resultados. Antes de proceder a generalizar los resultados de Gurka, sería muy clarificador investigar el desempeño de los IC, contemplando las mejoras analíticas referidas, cuando se manipula la distribución del término de error, la complejidad de las estructuras usadas para generar los datos y el número modelos incluidos en el proceso de selección.

Es razonable pensar que buena parte del elevado desempeño encontrado en el estudio de Gurka (p.e., todas las versiones de los IC examinados elegían la verdadera estructura de covarianza más del 90% de las veces) se explique por la forma de las matrices manipuladas, completamente generales *versus* excesivamente parcas, y por el reducido número de alternativas implicadas en el proceso de selección. En principio, asumido conocido el verdadero PGD, cabe esperar que el número de modelos incluidos en la comparación afecte más a los criterios eficientes que a los consistentes, ya que los primeros asumen que el verdadero PGD es de dimensión infinita; no obstante, es evidente que para ningún criterio será lo mismo seleccionar un modelo entre dos alternativas que entre dos docenas. Además, también cabe preguntarse ¿hasta qué punto resulta realista asumir que la varianza de las observaciones se mantiene constante y/o que la correlación no decrece a lo largo del tiempo cuando se estudian las curvas de crecimiento?

Por consiguiente, el presente trabajo tiene como objetivo determinar cuán efectivos son los criterios AIC, AICC, BIC, CAIC y HQIC para descubrir el verdadero PGD en una familia de modelos anidados. Estos criterios serán evaluados bajo estimación FML y $REML_1$/$REML_2$ cuando se manipulan diversas estructuras de medias y/o de covarianzas. Además, para proporcionar un punto de referencia para la comparación también utilizaremos el criterio de ajuste condicional LRT. En aras de dar respuesta al objetivo planteado, usaremos sendos diseños crossover en los que se violan separada y conjuntamente los supuestos de normalidad de los datos y de homogeneidad de las matrices de dispersión.

Definición de las herramientas usadas para seleccionar el mejor modelo mixto

Usar la metodología del modelo mixto en el contexto longitudinal, implica tener que elegir entre modelos alternativos para explicar la variabilidad observada en los datos del modo más sencillo posible. Aunque no existe unanimidad acerca de cual es la mejor forma de seleccionar el modelo óptimo, herramientas tales como los IC y el LRT son usadas frecuentemente.

Criterios de Información (IC)

En la Tabla 1 se definen las versiones de los IC investigados, tanto bajo estimación FML como bajo estimación $REML_1$/$REML_2$. También se indican las fórmulas empleadas por el módulo *Proc Mixed* del SAS (versión 9.2, 2008) y por el comando *Mixed* del SPSS (versión 17, 2008).

Test de razón de verosimilitudes (LRT)

Como ya ha sido indicado, el estadístico de bondad de ajuste más usado para comparar modelos anidados es el LRT. Este contraste puede obtenerse a partir de la expresión siguiente:

$$\Delta = -2[\hat{l}_{reducido(H_0)} - \hat{l}_{completo(H_1)}],$$

donde Δ es el estadístico desvianza y $\hat{l}_{reducido(H_0)}$ y $\hat{l}_{completo(H_1)}$ los máximos de la función FML o REML, según se trate de elegir entre modelos con idéntica estructura de covarianza o de medias, bajo la hipótesis nula y alternativa, respectivamente. El estadístico Δ se distribuye bajo H_0 según χ^2_v donde v indica la diferencia entre el número de parámetros estimados en el modelo completo y en el modelo reducido.

A pesar de la amplia utilización del LRT, su uso conlleva ciertas limitaciones que es preciso tener en cuenta. Por ejemplo, únicamente está definido para comparar modelos anidados y tan sólo permite comparar dos al mismo tiempo. Cuando el número de modelos anidados sea superior a dos, la aplicación del LRT requiere proceder jerárquicamente (para más detalles, véase Dayton, 2003). Por el contrario, es importante destacar que los IC son válidos para comparar y seleccionar modelos anidados y no anidados. Además, permiten la comparación simultánea de un conjunto de modelos.

Método de la simulación

Para evaluar el desempeño de los métodos descritos realizamos tres estudios de simulación. En el primero mantuvimos constante la estructura de covarianza y modelamos la estructura de medias. En el segundo supusimos conocida la estructura de medias y modelamos la de covarianza. En el tercero modelamos ambas estructuras a la vez. En cada una de ellos utilizamos un diseño crossover con dos tratamientos, dos secuencias y seis (en el tercer estudio también doce) periodos de una semana, en el que se violaban separada y conjuntamente los supuestos de normalidad y esfericidad multimuestral. Los participantes del primer grupo recibieron la secuencia de tratamiento AAABBB, mientras que los del segundo grupo recibieron la secuencia inversa para contrarrestar los posibles efectos residuales. En base a lo expuesto, se planteó el modelo de la forma

Tabla 1
Definición de los criterios de información usados en la selección del modelo mixto

$AIC = -2\log l_{FML} + 2(p+q)^{SAS,SPSS}$	$AIC_1 = -2\log l_{REML1} + 2q$
	$AIC_2 = -2\log l_{REML2} + 2q^{SAS,SPSS}$
$AICC_1 = -2\log l_{FML} + 2(p+q)\left(\frac{N}{N-p-q-1}\right)^{SAS,SPSS}$	$AICC_1 = -2\log l_{REML1} + 2q\left(\frac{N-p}{N-p-q-1}\right)$
$AICC_2 = -2\log l_{FML} + 2(p+q)\left(\frac{n}{n-p-q-1}\right)$	$AICC_2 = -2\log l_{REML1} + 2q\left(\frac{n}{n-q-1}\right)$
	$AICC_1 = -2\log l_{REML2} + 2q\left(\frac{N-p}{N-p-q-1}\right)^{SAS,SPSS}$
	$AICC_2 = -2\log l_{REML2} + 2q\left(\frac{n}{n-q-1}\right)$
$BIC_1 = -2\log l_{FML} + (p+q)\log(N)^{SPSS}$	$BIC_1 = -2\log l_{REML1} + q\log(N-p)$
$BIC_2 = -2\log l_{FML} + (p+q)\log(n)^{SAS}$	$BIC_2 = -2\log l_{REML1} + q\log(n)$
	$BIC_1 = -2\log l_{REML2} + q\log(N-p)^{SPSS}$
	$BIC_2 = -2\log l_{REML2} + q\log(n)^{SAS}$
$CAIC_1 = -2\log l_{FML} + (p+q)[\log(N)+1]^{SPSS}$	$CAIC_1 = -2\log l_{REML1} + q[\log(N-p)+1]$
$CAIC_2 = -2\log l_{FML} + (p+q)[\log(n)+1]^{SAS}$	$CAIC_2 = -2\log l_{REML1} + q[\log(n)+1]$
	$CAIC_1 = -2\log l_{REML2} + q[\log(N-p)+1]^{SPSS}$
	$CAIC_2 = -2\log l_{REML2} + q[\log(n)+1]^{SAS}$
$HQIC_1 = -2\log l_{FML} + 2(p+q)\log[\log(N)]$	$HQIC_1 = -2\log l_{REML1} + 2q\log[\log(N-p)]$
$HQIC_2 = -2\log l_{FML} + 2(p+q)\log[\log(n)]^{SAS}$	$HQIC_2 = -2\log l_{REML1} + 2q\log[\log(n)]$
	$HQIC_1 = -2\log l_{REML2} + 2q\log[\log(N-p)]$
	$HQIC_2 = -2\log l_{REML2} + 2q\log[\log(n)]^{SAS}$

Nota: p = número de parámetros del modelo de medias; q = número de parámetros de la estructura de covarianza; n = número total de sujetos; N = número total de observaciones; FML = estimador de máxima verosimilitud completa; $REML_1$/ $REML_2$ = estimadores de máxima verosimilitud residual con y sin el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$

$y_{ij}=\beta_{00}+\beta_{01}G_j+u_{0j}+\beta_{10}T_{ij}+\beta_{11}G_j\times T_{ij}+u_{1j}+\beta_{20}CT_{ij}+\beta_{21}G_j\times CT_{ij}+u_{2j}CT_{ij}+e_{ij}$, donde $G_j=0$ si el *i*-ésimo participante era asignado a la secuencia AAABBB durante un periodo de seis semanas y $G_j=1$ si era asignado a la secuencia BBBAAA; T_{ij} denota la semana en la cual se registraba la respuesta y CT_{ij} denota el cambio de tendencia lineal entre los tres primeros periodos y los tres últimos. Las variables $T_{ij}\in\{1, 2, 3, 4, 5, 6\}$ y $CT_{ij}\in\{1, 2, 3, 1, 2, 3\}$ fueron centrada con respecto a sus respectivas medias, concretamente $T_{ij}^{\bullet}=(T_{ij}-3.5)$ y $CT_{ij}^{\bullet}=(CT_{ij}-2)$.

Variables manipuladas en el primer estudio

En el primer estudio se evaluó el desempeño de los criterios de selección para elegir de un conjunto de modelos candidatos la ver-

Table 2
Modelos usados para ajustar la estructura de medias y valor de los parámetros de efectos fijos

$M_1^{©}$ $y_{ij} = \beta_{00} + u_{0j} + e_{ij}$

M_2 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + e_{ij}$

M_3 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$

M_4 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$

M_5 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$

$M_6^{©}$ $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$

$\beta_{\text{completo}}^{©} = \left[\beta_{00} = 3.125\ \beta_{01} = 1.25\ \beta_{10} = -.20\ \beta_{11} = 0.45\ \beta_{20} = -.25\ \beta_{21} = 0.50\right]$

$\beta_{\text{reducido}}^{©} = \left[\beta_{00} = 3.125\ \beta_{01} = 0.00\ \beta_{10} = 0.00\ \beta_{11} = 0.00\ \beta_{20} = 0.00\ \beta_{21} = 0.00\right]$

$M_6^{©}, M_1^{©}$ = modelos completo y reducido usados para generar los datos

dadera estructura de medias. Dicha evaluación fue realizada bajo estimación FML y $REML_1$/$REML_2$ cuando se manipulaban las variables siguientes:

(a) *Tipo de modelo usado para generar los datos*. El ajuste de la estructura de medias implicaba seleccionar de un conjunto de seis modelos anidados el verdadero PGD. En la mitad de las condiciones manipuladas, dicho proceso requería ajustar un modelo completo y en la otra mitad uno reducido. En ambos casos, los efectos fijos del modelo fueron definidos combinando distintas matrices de diseño y distintos vectores de parámetros. Bajo la primera situación, además del intercepto, el modelo que generó los datos (M_6) incluía como efectos fijos los grupos de tratamiento (G), la tendencia lineal (T), el cambio de tendencia (CT), la interacción G × T y la interacción G × CT. Los cinco modelos restantes fueron especificados erróneamente eliminando de la matriz de diseño una o más covariadas. Por ejemplo, el M5 fue especificado erróneamente eliminando la covariada G × CT, mientras que el M1 lo fue eliminando las covariadas G, T, CT, G × T y G × CT, respectivamente. Bajo la segunda situación, el modelo usado para generar los datos fue el M_1. Los modelos restantes fueron especificados erróneamente siguiendo un proceso inverso al descrito, es decir, añadiendo variables a la matriz de diseño. En la Tabla 2 aparecen recogidos los modelos usados en el proceso de comparación, así como el valor de los parámetros de efectos fijos de los modelos que generaron los datos. El vector de coeficientes del modelo completo, ligeramente modificado, se corresponde con el de un experimento descrito por Hedeker y Gibbons (2006; páginas 122-126).

(b) *Tamaño de muestra total*. El desempeño fue investigado usando dos tamaños de muestra distintos: $n = 30$ y $n = 60$. Estos tamaños grupales fueron seleccionados por ser representativos de los encontrados frecuentemente en las investigaciones psicológicas. Dentro de cada tamaño de muestra, el valor del coeficiente de variación muestral Δ se fijó en 0.33, donde $\Delta = \frac{1}{\bar{n}}[S_j(n_j - \bar{n})^2 / J]^{1/2}$ siendo $\bar{n}$ el tamaño promedio de los grupos. Cuando el diseño está equilibrado, Δ=0 Para n=30 los tamaños grupales fueron: (10-20), (15-15) y (20-10), mientras que para $n = 60$ los tamaños grupales fueron: (20-40), (30-30) y (40-20).

(c) *Patrones de covarianza empleados para generar los datos*. Los patrones utilizados para generar los datos fueron tres, a saber: coeficientes aleatorios lineales (RCL), autorregresivo de primer orden heterogéneo [ARH(1)] y UN. El primer patrón es un ejemplo de modelo jerárquico que permite modelar un intercepto y una o más tendencias más para cada participante. Este patrón puede resultar muy útil para caracterizar los datos, dado que aúna flexibilidad y parquedad, de hecho, tan sólo requiere estimar parámetros. En este estudio q=3 el intercepto, la tendencia lineal y el cambio de tendencia. El segundo patrón permite que las varianzas sean heterogéneas y que las covarianzas decrezcan exponencialmente, pero asume que las observaciones se hallan igualmente espaciadas entre sí. Este modelo es típico de las series temporales cortas y precisa estimar (t+1) parámetros. Por su parte, el tercer patrón representa la estructura de covarianza que mejor se ajusta a los datos, además no exige que las observaciones se encuentren igualmente espaciadas. No obstante, requiere estimar t (t+1)/2 parámetros.

(d) *Desviación del supuesto de esfericidad*. Aunque el modelo mixto no asume que las varianzas de las diferencias entre pares de medidas repetidas sean iguales (supuesto de esfericidad), la investigación empírica ha puesto de relieve que las inferencias realizadas con este enfoque sí pueden verse afectadas por la falta de esfericidad (Vallejo, Fernández, Herrero y Conejo, 2004). Por este motivo, patrones de covarianza con valores de ε (índice de ausencia de esfericidad derivado por Box) de .47 y .70 fueron empleados para investigar sus efectos en el desempeño de los criterios de selección. Las estructuras de covarianza usadas están disponibles en la Web http://gip.uniovi.es/gdiyad/docume/Psicothema/.

(e) *Igualdad de las matrices de dispersión*. El desempeño de las herramientas de selección fue evaluado cuando las matrices de covarianza grupales eran homogéneas y también cuando eran heterogéneas. En el primer caso, los elementos de las dos matrices de dispersión fueron iguales entre sí (Σ_2=Σ_1) mientras que en el segundo caso, los elementos de una de las matrices fueron cinco veces mayores que los de la otra (Σ_2=5Σ_1).

(f) *Emparejamiento de las matrices de covarianza y el tamaño de los grupos*. La forma de relacionar el tamaño de los grupos y el tamaño de las matrices de dispersión pueden tener diferentes efectos en las pruebas estadísticas. Cuando el diseño está equilibrado, la relación entre el tamaño de las matrices de dispersión y el tamaño de los grupos es nula. Cuando el diseño está desequilibrado, la relación puede ser positiva o negativa. Una relación positiva implica que el grupo de menor tamaño se asocia con la matriz de dis-

persión menor, mientras que una relación negativa implica que el grupo de menor tamaño se asocia con la matriz de dispersión mayor.

(g) *Forma de la distribución de la variable de medida*. Aunque el enfoque del modelo mixto está basado en el cumplimiento del supuesto de normalidad, cuando se trabaja con datos reales es común que los índices de asimetría γ_1 y curtosis γ_2 se desvíen de cero (Micceri, 1989), lo cual puede inducirnos a interpretar incorrectamente los resultados. Para investigar el efecto que ejerce forma de la distribución en el desempeño de los criterios de selección, generamos datos desde distribuciones normales y no normales mediante las distribuciones *g* y *h* introducidas por Tukey (1977). Además de la distribución normal (g = h = 0; $\gamma_1 = \gamma_2 = 0$), también investigamos otras tres: (a) $g = 0$ y $h = .109$, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial doble o de Laplace ($\gamma_1 = 0$ & $\gamma_2 = 3$); (b) $g = .76$ y $h = -.098$, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial ($\gamma_1 = 2$ & $\gamma_2 = 6$) y (c) $g = 1$ y $h = 0$, una distribución que tiene el mismo grado de sesgo y de curtosis que la distribución lognormal ($\gamma_1 = 6.18$ & $\gamma_2 = 110.94$) Las distribuciones *g* y *h* fueron obtenidas utilizando la función RANNOR del SAS. Mediante ella generamos variables aleatorias normales estándar (Z_{ijk}) y transformamos cada una de ellas como $Z^*_{ijk} = g^{-1}[\exp(gZ_{ijk}) - 1]\exp(hZ^2_{ijk}/2)$ donde *g* y *h* son números reales que controlan el grado sesgo y de curtosis. Por último, para obtener una distribución con desviación estándar σ_{jk} cada una de las puntuaciones que conforman la variable dependiente fue creada utilizando el modelo lineal $Y_{ijk} = \sigma_{jk} \times (Z^*_{ijk} - \mu_{gh})$ donde $\mu_{gh} = \{\exp[g^2/(2-2h)] - 1\}/[g(1-h)^{1/2}]$ es la media de la de la distribución *g* y *h* (para detalles véase Kowalchuk y Headrick, 2009).

Variables manipuladas en el segundo estudio

En este estudio se evaluó el desempeño los criterios de selección para elegir de un conjunto de modelos candidatos la verdadera estructura de covarianza. Dicho ajuste implicaba seleccionar de un conjunto de seis patrones anidados el verdadero PGD. En la mitad de las condiciones manipuladas, se requería ajustar un modelo en el cual la varianza se mantenía constante a lo largo del tiempo y la covarianza decrecía exponencialmente (AR(1)) y en la otra mitad un modelo UN. Bajo la primera situación, además del modelo AR(1) usado para generar los datos, el conjunto de modelos candidatos incluía un modelo de independencia (IND), un modelo ARH(1), un modelo Toeplitz homogéneo (TOEP), un modelo TOEP heterogéneo (TOEPH) y un modelo UN. El modelo IND asume varianza constante y covarianza serial nula, mientras que los modelos TOEP y TOEPH generalizan, respectivamente, a los modelos AR(1) y ARH(1). Diversos investigadores ofrecen una descripción detallada de estos modelos, incluyendo Fitzmaurice et al. (2004), Littell et al. (2006) y Zimmerman y Núñez-Antón (2009). Repárese que las estructuras están anidadas unas dentro de otras, en el sentido que IND es un caso especial de AR(1), ésta lo es de TOPH, la cual a su vez lo es de TOEPH y ésta última lo es necesariamente de UN.

Bajo la segunda situación, el conjunto de modelos candidatos era idéntico al descrito, pero los datos fueron generados a partir del modelo UN. Además de los métodos de estimación y de los patrones de covarianza usados en la generación de los datos, también fueron manipuladas las variables tamaño de muestra, igualdad de las matrices de dispersión y forma de la distribución de la población. Las estructuras de covarianza usadas están disponibles en la citada Web.

Variables manipuladas en el tercer estudio

Para profundizar en el desempeño de las pruebas ajustamos simultáneamente la estructura de medias y la estructura de covarianza. Dicho ajuste implicaba seleccionar de un conjunto de nueve modelos candidatos el verdadero PGD. En la Tabla 3 aparecen recogidos los modelos utilizados en la comparación, así como el valor de los parámetros de efectos fijos usados para generar los datos. Examinando la Tabla 3 se aprecia que los modelos estaban anidados unos dentro de otros, en aquellos casos que el número de

Tabla 3
Conjunto de modelos de medias y de covarianza candidatos y valor de los parámetros de efectos fijos

M_1	$E(y_{ij}) = \beta_{00}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_2	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_3	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_4	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_5	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_6	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_7	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
$M^{©}_8$	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij} + \beta_{21}G_j \times CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
M_9	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}T^{2\bullet}_{ij} + \beta_{30}CT^{\bullet}_{ij} + \beta_{31}G_j \times CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
$\beta' = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -0.50\ \beta_{11} = 0.50\ \beta_{20} = -0.50\ \beta_{21} = 0.50]$		
$\beta' = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -1.00\ \beta_{11} = 1.00\ \beta_{20} = -1.00\ \beta_{21} = 1.00]$		

Nota: $M_1 \subset M_2 \subset M_3 \subset M_4 \subset M_5 \subset M_6 \subset M_7 \subset M^{©}_8 \subset M_9$; $M^{©}_8$ = modelo usado para generar los datos

efectos fijos era idéntico, como sucedía con los modelos $M_3 - M_4$ y $M_6 - M_7$, las estructuras de covarianza diferían y se hallaban anidadas entre sí, en el sentido que AR(1) es un caso especial de ARH(1) la cual es a su vez un caso especial de ANTE(1). Para una exhaustiva descripción de esta última estructura, véase Zimmerman y Núñez-Antón (2009).

En este tercer estudio, además de los métodos de estimación, también fueron manipuladas las variables tamaño de muestra, número de medida repetidas (t = 6 y t = 12), valor de los parámetros de efectos fijos, igualdad de las matrices de dispersión y forma de la distribución. Los valores de los parámetros de covarianza de la matriz ANTE(1) están disponibles en laWeb citada.

Resultados del primer estudio

La tabla 4 recoge el porcentaje de veces que los 29 criterios examinados, 10 vía FML y 19 vía REML, seleccionaban el verdadero modelo de medias cuando la estructura usada para generar los

Tabla 4
Porcentaje de veces que los criterios elegían el modelo de medias verdadero cuando el patrón de covarianza conocido era ARH (1)

		Modelo Completo					Modelo Reducido				
ME	Criterio	Norm	Lapla	Expon	Logn	Media	Norm	Lapla	Expon	Lognor	Media
FML	$AIC^{(SAS, SPSS)}$	59.24	49.31	38.69	30.26	**44.37** [1]	68.73	71.66	65.08	49.23	**63.68**
FML	$AICC_1$	26.54	17.37	12.33	08.21	**16.11**	93.61	94.69	91.54	61.78	**85.40**
FML	$AICC_2^{(SAS, SPSS)}$	54.23	44.43	33.85	25.34	**39.46** [2]	75.76	76.60	70.13	52.28	**68.69**
FML	$HQIC_1$	41.96	32.65	22.44	16.72	**28.44**	88.29	90.43	86.18	66.17	**82.77**
FML	$HQIC_2^{(SAS)}$	50.70	38.37	30.27	21.50	**35.21** [3]	78.63	82.12	77.64	63.09	**50.37**
FML	$BIC_1^{(SPSS)}$	23.46	15.72	07.69	05.15	**13.00**	95.97	97.38	95.15	82.23	**92.68** [2]
FML	$BIC_2^{(SAS)}$	38.82	29.10	19.28	13.02	**25.06**	89.55	91.74	87.44	69.50	**84.56**
FML	$CAIC_1^{(SPSS)}$	17.28	10.61	05.01	02.49	**08.85**	97.94	98.73	96.79	85.37	**94.70** [1]
FML	$CAIC_2^{(SAS)}$	30.45	21.30	11.96	07.12	**17.69**	93.99	95.93	92.75	76.83	**89.87** [3]
FML	LRT	38.53	29.71	20.94	14.74	**25.98**	85.23	87.09	82.10	62.79	**79.30**
$REML_1$	AIC_1	99.99	99.97	96.71	87.36	**96.01**	91.87	93.08	90.81	78.09	**88.46**
$REML_2$	$AIC_2^{(SAS, SPSS)}$	85.30	88.60	84.75	74.54	**83.30**	61.52	65.74	58.98	38.70	**56.23**
$REML_1$	$AICC_1$	99.99	99.49	99.04	94.12	**98.16**	93.59	94.12	**92.51**	80.02	**90.06**
$REML_1$	$AICC_2$	75.00	75.01	75.26	75.38	**75.16**	98.50	98.46	97.78	90.12	**96.22**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	85.28	88.54	89.98	80.29	**86.02**	69.69	71.58	62.53	41.49	**61.32**
$REML_2$	$AICC_2$	63.52	65.61	68.92	70.84	**67.23**	85.85	85.61	58.07	47.73	**69.31**
$REML_1$	$HQIC_1$	99.99	99.99	99.88	98.50	**99.59**	97.87	96.41	96.69	87.40	**94.59**
$REML_1$	$HQIC_2$	99.99	99.98	99.27	94.67	**98.48**	95.25	96.55	93.98	82.78	**92.14**
$REML_2$	$HQIC_1$	85.33	88.86	92.49	94.06	**90.18**	88.42	88.95	76.04	51.95	**76.34**
$REML_2$	$HQIC_2^{(SAS)}$	85.42	88.74	89.49	84.50	**87.04**	79.42	80.81	67.84	45.45	**68.38**
$REML_1$	BIC_1	99.99	99.99	99.99	99.74	**99.93** [3]	99.68	98.68	98.39	93.25	**97.50** [2]
$REML_1$	BIC_2	99.99	99.99	99.93	98.15	**99.52**	98.26	98.61	97.10	88.41	**95.60**
$REML_2$	$BIC_1^{(SPSS)}$	87.24	89.59	92.69	94.28	**90.95**	90.15	88.29	73.49	49.84	**75.44**
$REML_2$	$BIC_2^{(SAS)}$	85.38	88.72	91.59	91.91	**89.40**	87.98	86.39	71.79	40.39	**71.64**
$REML_1$	$CAIC_1$	99.99	99.99	99.99	99.96	**99.99** [1]	**99.82**	99.79	99.01	93.90	**98.13** [1]
$REML_1$	$CAIC_2$	99.99	99.99	99.99	99.75	**99.93** [2]	95.46	99.40	98.32	91.76	**96.23** [3]
$REML_2$	$CAIC_1^{(SPSS)}$	86.83	89.72	92.77	94.42	**90.94**	90.16	88.51	73.40	49.79	**70.46**
$REML_2$	$CAIC_2^{(SAS)}$	85.40	88.73	92.40	93.20	**89.93**	89.50	86.97	72.15	48.47	**74.27**
$REML_2$	LRT	15.39	12.05	10.65	09.94	**12.01**	41.68	55.08	58.68	56.97	**53.10**

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables naturaleza del emparejamiento, tamaño de muestra, desviación de la esfericidad e igualdad de las matrices de dispersión; ME = método de estimación; FML = estimación por máxima verosimilitud completa; $REML_1$ = estimación por máxima verosimilitud residual incorporando el término aditivo; $REML_2$ = estimación por máxima verosimilitud residual eliminando el término aditivo

datos era ARH(1). El patrón de resultados correspondiente a las matrices RCL y UN no aparece recogido. Observando la Web citada se aprecia que dicho patrón era cualitativa y cuantitativamente similar al descrito. Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables igualdad de las matrices de dispersión y tamaño de muestra. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía de la forma de la distribución de la variable de medida, tipo de modelo a seleccionar y procedimiento de estimación utilizado. Aunque no aparece recogido en la tabla 4, las variables naturaleza del emparejamiento, desviación de la esfericidad e igualdad de las matrices de dispersión afectaron ligeramente al porcentaje de identificaciones correctas y el tamaño de muestra sustancialmente.

2. El desempeño de los IC era más elevado bajo estimación REML que bajo estimación FML. Promediando a través de las 2592 (144 × 18) condiciones manipuladas, el porcentaje de aciertos obtenidos vía REML fue del 86.5%, mientras que el obtenido vía FML promediando a través de 1296 (144 × 9) condiciones fue del 52.3%. Con el LRT sucedió lo contrario.

Tabla 5
Porcentaje de veces que los criterios elegían la estructura de covarianza verdadera bajo estimación FML y REML

		Patrón Autorregresivo [AR(1)]					Patrón General (UN)				
ME	Criterio	Norm	Lapla	Expon	Logn	Media	Norm	Lapla	Expon	Lognor	Media
FML	$AIC^{(SAS, SPSS)}$	77.91	53.77	12.60	04.21	**37.12**	56.37	59.18	86.06	92.92	**73.63** [2]
FML	$AICC_1$	72.22	67.06	33.82	17.00	**47.53**	25.79	26.13	35.19	39.19	**31.60**
FML	$AICC_2^{(SAS, SPSS)}$	84.33	64.80	17.80	06.74	**43.42**	32.61	34.07	63.59	77.57	**51.96** [3]
FML	$HQIC_1$	96.23	87.53	38.39	17.92	**60.02**	07.35	08.54	38.57	49.19	**25.92**
FML	$HQIC_2^{(SAS)}$	90.51	74.25	24.47	08.89	**49.53**	22.96	25.92	64.47	74.09	**46.86**
FML	$BIC_1^{(SPSS)}$	99.52	97.20	63.09	35.31	**73.78** [2]	00.05	00.13	08.02	14.25	**05.61**
FML	$BIC_2^{(SAS)}$	96.34	89.70	41.54	18.94	**61.63**	02.53	04.13	30.20	37.92	**18.70**
FML	$CAIC_1^{(SPSS)}$	99.89	98.73	72.42	43.71	**78.69** [1]	00.00	00.06	04.03	08.05	**03.04**
FML	$CAIC_2^{(SAS)}$	99.43	95.24	54.98	28.39	**69.51** [3]	00.18	00.76	14.01	20.17	**08.78**
FML	LRT	67.77	42.12	09.12	03.45	**30.61**	84.12	82.17	94.18	97.05	**89.38** [1]
$REML_1$	AIC_1	83.69	34.41	34.13	15.79	**42.00**	50.74	58.73	82.63	89.32	**70.36** [2]
$REML_2$	$AIC_2^{(SAS, SPSS)}$	83.69	34.41	34.13	15.79	**42.00**	50.74	58.78	82.62	89.31	**70.37** [3]
$REML_1$	$AICC_1$	86.44	39.60	45.58	25.24	**49.21**	32.85	38.22	65.83	76.65	**53.39**
$REML_1$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.14	41.29	44.16	**37.02**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	86.44	39.59	45.57	25.24	**49.21**	32.85	38.27	65.83	76.65	**53.40**
$REML_2$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.13	41.29	44.36	**37.07**
$REML_1$	$HQIC_1$	97.93	68.99	77.26	54.75	**74.73**	04.97	09.48	36.21	53.47	**26.03**
$REML_1$	$HQIC_2$	93.68	52.42	58.38	33.53	**59.51**	18.76	25.73	61.52	75.25	**45.32**
$REML_2$	$HQIC_1$	97.93	68.66	77.22	54.73	**74.64**	04.47	08.14	35.88	53.28	**25.44**
$REML_2$	$HQIC_2^{(SAS)}$	93.68	52.42	58.38	33.53	**59.50**	18.76	25.69	61.48	75.23	**45.29**
$REML_1$	BIC_1	99.94	88.44	93.95	81.83	**91.04** [3]	00.00	00.04	06.87	17.28	**06.05**
$REML_1$	BIC_2	98.45	73.13	82.08	61.17	**78.73**	00.99	03.90	27.31	44.94	**19.29**
$REML_2$	$BIC_1^{(SPSS)}$	99.94	88.39	93.43	81.66	**90.86**	00.00	00.04	06.87	17.17	**06.02**
$REML_2$	$BIC_2^{(SAS)}$	98.45	73.21	82.08	61.16	**78.71**	00.99	03.89	27.31	44.91	**19.28**
$REML_1$	$CAIC_1$	99.99	92.51	96.20	87.41	**94.03** [1]	00.00	00.08	03.37	09.93	**03.35**
$REML_1$	$CAIC_2$	99.67	83.80	90.38	75.17	**87.50**	00.00	00.44	12.19	25.54	**09.55**
$REML_2$	$CAIC_1^{(SPSS)}$	98.98	92.49	96.18	87.40	**94.01** [2]	00.00	00.08	03.36	09.93	**03.35**
$REML_2$	$CAIC_2^{(SAS)}$	99.66	83.79	90.38	75.16	**87.25**	00.00	00.44	12.18	25.54	**09.55**
$REML_2$	LRT	78.11	25.50	17.60	07.87	**32.28**	81.45	82.78	92.57	95.46	**88.06** [1]

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables tamaño de muestra e igualdad de las matrices de dispersión

3. En promedio, las diferencias existentes entre los IC y el LRT bajo estimación FML fueron mínimas. En concreto, los IC seleccionaron correctamente el modelo completo en el 25.4% de las veces y el reducido en el 79.2%, mientras que el LRT lo hizo en el 26% y 79.3% de las veces. Sin embargo, bajo estimación REML los IC seleccionaron correctamente el modelo completo en el 91.2% de las veces y el reducido en el 81.8%, mientras que el LRT lo hizo en el 12% y 53.1% de las veces, respectivamente.

4. Con relación al desempeño de los IC bajo estimación REML, los criterios consistentes elegían el verdadero modelo de medias en el 89.4% de las veces y los eficientes el 80.7%. Además, el desempeño de ambas clases de criterios mejoraba cuando se incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ en la ecuación. El porcentaje de aciertos de los criterios consistentes era del 98.1% bajo $REML_1$ y del 81.2% bajo $REML_2$. Por su parte, los criterios eficientes elegían el verdadero PGD el 90% de las veces bajo $REML_1$ y el 75.5% bajo $REML_2$.

Tabla 6
Porcentaje de veces que los criterios elegían correctamente el modelo de efectos fijos y aleatorios bajo estimación FML y REML

		t = 6		*t* = 12		
ME	Criterio	$\beta'_{(1)}$	$\beta'_{(2}$	$\beta'_{(1)}$	$\beta'_{(2)}$	Global
FML	$AIC^{(SAS, SPSS)}$	38.94[1]	67.01[2]	73.92	79.44	**64.8[2]**
FML	$AICC_1$	08.62	24.53	20.75	24.70	**19.7**
FML	$AICC_2^{(SAS, SPSS)}$	33.59[2]	64.31	74.74[3]	83.22[3]	**64.0**
FML	$HQIC_1$	20.74	55.14	69.62	89.10	**58.7**
FML	$HQIC_2^{(SAS)}$	30.64	64.23[3]	77.03[2]	84.56	**64.1[3]**
FML	$BIC_1^{(SSPSS)}$	05.54	25.84	33.32	70.72	**33.9**
FML	$BIC_2^{(SAS)}$	17.17	50.51	70.91	89.84[2]	**57.1**
FML	$CAIC_1^{(SPSS)}$	02.74	16.98	22.82	57.16	**24.9**
FML	$CAIC_2^{(SAS)}$	09.52	34.71	55.95	85.54	**46.4**
FML	LRT	31.89[3]	73.75[1]	87.93[1]	99.16[1]	**73.2[1]**
$REML_1$	AIC_1	81.51[1]	82.87	88.59	89.36	**85.6[1]**
$REML_2$	$AIC_2^{(SAS, SPSS)}$	74.21[2]	87.27[1]	92.68[1]	98.93[1]	**88.3[2]**
$REML_1$	$AICC_1$	77.34[3]	81.38	88.25	89.39	**84.1**
$REML_1$	$AICC_2$	42.99	54.64	44.14	44.48	**46.6**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	63.58	85.08[2]	90.91[2]	98.91[2]	**84.6[3]**
$REML_2$	$AICC_2$	37.22	56.31	46.10	49.43	**47.3**
$REML_1$	$HQIC_1$	59.98	74.79	79.13	89.59	**75.9**
$REML_1$	$HQIC_2$	73.21	81.04	87.47	89.22	**82.7**
$REML_2$	$HQIC_2$	51.50	76.42	81.90	98.32	**77.1**
$REML_2$	$HQIC_2^{(SAS)}$	60.76	84.18[3]	91.03[3]	98.87[3]	**83.7**
$REML_1$	BIC_1	32.28	47.20	42.69	80.73	**50.7**
$REML_1$	BIC_2	54.90	73.73	80.09	87.12	**73.9**
$REML_2$	$BIC_1^{(SPSS)}$	27.13	47.99	42.52	87.77	**51.4**
$REML_2$	$BIC_2^{(SAS)}$	46.73	73.11	80.98	98.79	**74.9**
$REML_1$	$CAIC_1$	23.35	43.77	32.04	73.29	**43.1**
$REML_1$	$CAIC_2$	40.34	59.66	65.73	88.20	**63.5**
$REML_2$	$CAIC_1^{(SPSS)}$	20.16	35.62	32.11	79.41	**41.8**
$REML_2$	$CAIC_2^{(SAS)}$	34.51	58.58	63.99	97.62	**63.7**
$REML_2$	LRT	28.33	71.11	79.59	96.89	**68.9**

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través del tamaño de muestra, forma de la distribución e igualdad de las matrices de dispersión. $\beta'_{(1)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -0.50\ \beta_{11} = 0.50\ \beta_{21} = -0.50\ \beta_{21} = 0.50]$; $\beta'_{(2)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -1.00\ \beta_{11} = 1.00\ \beta_{10} = -1.00\ \beta_{21} = 1.00]$, *t* = número de periodos de observación. La última columna representa el porcentaje promedio a través de los tres experimentos.

Resultados del segundo estudio

En la tabla 5 aparece el porcentaje de veces que los criterios examinados elegían la estructura de covarianza verdadera, cuando los datos fueron generados desde sendas matrices AR(1) y UN. Los datos tabulados denotan el porcentaje promedio de elecciones correctas a través de las variables igualdad de las matrices y tamaño de muestra. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía del patrón de covarianza usado para generar los datos y de la forma de la distribución. La influencia del los procedimientos de estimación era menor. Aunque no se recoge en la Tabla 5, los detalles se encuentran en la Web antes citada, las variables igualdad de las matrices de dispersión y tamaño de muestra afectaban sustancialmente al porcentaje de identificaciones correctas cuando el modelo usado para generar los datos había sido el UN, pero escasamente cuando se usaba el AR(1).

2. El desempeño del LRT fue superior al de los IC, tanto bajo estimación FML como REML. En promedio el LRT seleccionó correctamente el verdadero PGD en el 60% de las veces (el 31% bajo ARH(1) y el 89 % bajo UN), mientras que los IC lo hicieron en el 47.2%.

3. Cuando el modelo utilizado para generar los datos fue el AR(1), el porcentaje de aciertos disminuía conforme los datos se desviaban de la normalidad. El fenómeno contrario se producía con los datos generados bajo el modelo UN.

4. Con independencia del procedimiento de estimación utilizado, los criterios consistentes se comportaban mejor que sus homólogos eficientes cuando el modelo utilizado para generar los datos fue el AR(1). El porcentaje de aciertos de los criterios consistentes fue del 73.2% y el de los eficientes del 44.9%. Por el contrario, la situación se invertía cuando el modelo utilizado para generar los datos fue el UN. Los criterios consistentes elegían el verdadero PGD en el 18.1% de las veces y los eficientes en el 52.7%. A diferencia de lo encontrado en el estudio anterior, el desempeño de ambas clases de criterios no mejoraba cuando el estimador REML incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$.

Resultados del tercer estudio

En la tabla 6 aparece tabulado el porcentaje de veces que los criterios examinados elegían correctamente la estructura de medias y de covarianza, tanto bajo estimación FML como REML. Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables tamaño de muestra, igualdad de las matrices de dispersión y forma de la distribución. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía del método de estimación, valor de los parámetros y número de medidas repetidas. Aunque no aparece recogido en la Tabla 6, los detalles pueden consultarse en la Web, la variable tamaño de muestra afectaba sustancialmente a la selección del verdadero PGD y las variables forma de la distribución e igualdad de las matrices de dispersión moderadamente.

2. El desempeño de los IC fue mejor bajo estimación REML que bajo estimación FML. Promediando a través de las 1052 (64 × 18) condiciones manipuladas, el porcentaje de aciertos obtenidos vía REML fue del 69.7% (del cual el 65.2% corresponde a los consistentes y el 74.2% a los eficientes), mientras que el obtenido vía FML promediando a través de 576 (64 × 9) condiciones fue del 55.9% (del cual el 47.5% corresponde a los consistentes y el 64.3% a los eficientes). Por su parte, el LRT eligió el verdadero PGD en el 73.1% de las veces bajo estimación FML y en el 68.9% bajo estimación REML.

3. Cuando el método de estimación usado era FML y $t = 6$, las diferencias existentes entre los IC y el LRT no excedían los 2 puntos porcentuales. Sin embargo, bajo estimación REML las diferencias favorecían a los IC y eran superiores a los 20 puntos. Sorprendentemente, la situación se invertía cuando $t = 12$. En este caso las diferencias excedían los 10 puntos porcentuales.

4. Por último, hay que destacar que el desempeño de los IC mejoraba si el estimador REML incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$, pero sólo cuando $t = 6$. También cabe resaltar que el desempeño de los criterios consistentes era superior cuando se usaban las fórmulas implementadas en el módulo *Proc Mixed* del SAS, en lugar de las implementadas en el comando *Mixed* del SPSS. Observando la Tabla 1 se puede comprobar que los algoritmos usados por SAS y SPSS para calcular los criterios eficientes son idénticos.

Resultados globales

El desempeño global fue del 56.5% usando FML y del 62.9% usando REML. Los IC seleccionaron el verdadero PGD el 48.1% de las veces vía FML y el 68.7% vía REML, mientras que el LRT lo hizo el 64.8% y 57.7% de las veces, respectivamente. Bajo el método de estimación FML, los criterios eficientes seleccionaron el verdadero PGD el 51.3% de las veces - 58.1% el AIC y 44.5% el AICC - y los criterios consistente el 47.1% - 46.4% el BIC, 42.9% el CAIC y 52.1% el HQIC. A su vez, bajo el método REML, los criterios eficientes seleccionaron el verdadero PGD el 69.6% de las veces -74.7% el AIC y 64.5% el AICC - y los criterios consistente el 68.1% de las veces - 67.2% el BIC, 63.9% el CAIC y 73.2% el HQIC.

Conclusiones, recomendaciones y limitaciones

Son muchas las conclusiones se pueden extraer del estudio actual, no obstante, conviene destacar las cinco siguientes:

• En primer lugar, los datos pusieron de relieve que ninguno de los procedimientos examinados elegía consistentemente el verdadero PGD; sin embargo, su ejecución mejoraba al aumentar el tamaño de muestra, el número de medidas repetidas y la magnitud de los parámetros.

• En segundo lugar, independientemente de la parte del modelo que se ajustase, los IC basados en el estimador REML seleccionaban el verdadero PGD un mayor número de veces que los IC basados en el estimador FML. Este resultado no sólo confirma y extiende los hallazgos de Gurka (2006), sino que cuestiona la recomendación recogida en la literatura estadística especializada de usar en exclusiva los IC con REML para comparar modelos con idéntica estructura de medias (Orelien y Edwards, 2007).

• En tercer lugar, el desempeño de los IC mejoraba si el estimador REML incluía el término constante, especialmente cuando el número de medidas repetidas era moderado. A diferencia de lo que sucedía en el trabajo de Gurka (2006), donde el estimador REML1 sólo mejoraba el desempeño de los criterios consistentes, en el trabajo actual también mejoraba la ejecución de los criterios eficientes. Este resultado coincide con el encontrado por Wang y Schaalje (2009) al comparar el desempeño de los criterios AIC y BIC con el de los criterios predictivos R^2_{adj} CCC y PRESS.

• En cuarto lugar, el desempeño de los IC era superior cuando dichos criterios se calculaban usando el número total de sujetos (nivel 2), en vez del número total de observaciones (nivel 1). Este hallazgo, además de corroborar los resultados de Gurka (2006), también sirve de soporte empírico a la estrategia seguida en el módulo *Proc*

Mixed del SAS (2008), como opuesta a la seguida en el comando *Mixed* del SPSS (2008), de calcular los criterios consistentes usando el número de participantes en el nivel 2 del modelo jerárquico.

• En quinto lugar, a pesar de que los criterios AIC (78%), AICC (76.5%) y HQIC(76.8%) basados en el estimador REML1 elegían el verdadero PGD el mayor número de veces, cuando se requería ajustar la estructura de covarianza y el modelo completo el desempeño del LRT era tan bueno o mejor que el de los criterios reseñados.

Para concluir queremos efectuar una recomendación, una advertencia y una sugerencia. Globalmente, los criterios eficientes trabajaban mejor que los criterios consistentes cuando la estructura de covarianza era compleja y, viceversa, cuando era sencilla. Los criterios consistentes tendían a seleccionar modelos más parcos, generalmente de carácter estacionario, que los criterios eficientes. Ahora bien, en los estudios longitudinales de carácter aplicado suele ser habitual que la varianza de las observaciones sea heterogénea y que la correlación entre las mismas decrezca a lo largo del tiempo, de ahí que nos decantemos por el empleo de los criterios eficientes, en particular del AIC basado en el estimador $REML_1$. A nuestro juicio es el que cumple mejor el objetivo de encontrar un equilibrio entre un modelo complejo y otro parco. Hecha esta recomendación, debemos advertir que los resultados son limitados a las condiciones examinadas, si bien conjeturamos que pueden ser generalizadas a un rango más amplio de condiciones; por ejemplo, a situaciones donde los modelos no se hallen anidados unos dentro de otros. Finalmente, en la investigación realizada el verdadero PGD siempre pertenecía a la familia de modelos investigados. Sin embargo, cuando se trabaja con datos reales desconocemos si el verdadero PGD pertenece a la clase de modelos considerados. Por este motivo, sería deseable realizar una investigación donde el objetivo fuese comparar los IC en términos de seleccionar el modelo más próximo al verdadero PGD, dado que éste no se haya incluido en el conjunto de modelos presentes en la comparación.

Agradecimientos

Este trabajo ha sido financiado mediante sendos Proyectos de Investigación concedidos por el Ministerio de Ciencia e Innovación. (Ref.: PSI-2008-03624/PSI2009-11136/PSIC).

Referencias

Ato, M., & Vallejo, G. (2007). *Diseños experimentales en Psicología*. Madrid: Pirámide.

Claeskens, G., y Hjort, N.L. (2008). *Model Selection and Model Averaging*. New York: Cambridge University Press.

Dayton, C.M. (2003). Model comparisons using information measures. *Journal of Modern Applied Statistical Methods, 2*, 281-292.

Feng, R., Zhou, G., Zhang, M., y Zhang, H. (2009). Analysis of Twin Data Using SAS. *Biometrics*, Epub 2008 July 21 [PMID: 18647295].

Ferron, J., Dailey, R., y Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*, 379-403.

Fitzmaurice, G.M., Laird, N.M., y Ware, J.H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley.

Gómez, V.E., Schaalje, G.B., y Fellingham, G.W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation, 34*, 377-392.

Gurka, M.J., y Edwards, L.J. (2008). Mixed models. En C.R. Rao, J.P. Miller y D.C. Rao (Eds.): *Handbook of Statistics, vol. 27, Epidemiological and Medical Statistics* (pp. 253-280). New York: Elsevier Science.

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

Kowalchuk, R.K., y Headrick, T.C. (2009). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*. DOI:10.1348/000711009X42 3067.

Hedeker, D., y Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: John Wiley.

Keselman, H.J., Algina, J., Kowalchuk, R.K., y Wolfinger, R.D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics- Simulation and Computation, 27*, 591-604.

Kreft, I.G., y de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.

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

Liang, H., Wu, H., y Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika, 95*, 773- 778.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., y Schabenberger, O. (2006). SAS *System for Mixed Models*. 2nd edition. Cary, NC: SAS Institute Inc.

Littell, R.C., Pendergast, J., y Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793-1819.

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

Molenberghs, G., y Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling, 1*, 235-269.

Orelien, J.G. y Edwards, L.J. (2007). Fixed-effect variable selection in linear mixed models using R^2 statistics. *Computational Statistics & Data Analysis, 52*, 1896-1907.

SAS Institute Inc. (2008). *SAS/STAT® Software: Version 9.2*. SAS Institute Inc., Cary, NC.

Schabenberger, O. (2004). Mixed model influence diagnostics. *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., Paper 189-29.

Singer, D.J., y Willet, J.B. (2003). *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

SPSS for Windows (2008). Version 17, SPSS Inc., IL: Chicago.

Tukey, J.W. (1977). Modern techniques in data analysis. NSF-sponsored regional research conference at Southern Massachusetts University (North Dartmouth, MA).

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

Vallejo, G., Arnau, J., y Bono, R. (2008a). Construcción de modelos jerárquicos en contextos aplicados. *Psicothema, 20*, 830-838.

Vallejo, G., Ato, M., y Valdés, T. (2008b). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology, 4*, 10-21.

Vallejo, G., Fernández, P., Herrero, J., y Conejo, N. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema, 16*, 498-508.

Vonesh, E.F., Chinchilli, V.M., y Pu, K. (1996). Goodness-of-Fit in generalized nonlinear mixed-effects models. *Biometrics, 52*, 575-587.

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

Zimmerman, D.L., y Núñez-Antón, V. (2009). *Antedependence Models for Longitudinal Data*. London: Chapman & Hall/CRC.

Nested model selection for longitudinal data using information criteria and the conditional adjustment strategy

Guillermo Vallejo[*], Jaime Arnau[**], Roser Bono[**], Paula Fernández[*] & Ellián Tuero-Herrero[*]

[*]*Universidad de Oviedo*
[**]*Universidad de Barcelona*
gvallejo@uniovi.es

Abstract

Knowledge of the subject matter plays a vital role when attempting to choose the best possible linear mixed model to analyze longitudinal data. To date, in the absence of strong theory, much of the work has focused on modeling the covariance matrix by comparing non-nested models using selection criteria. In this paper, we compare the performance of conditional likelihood ratio test (LRT) and several versions of information criteria for selecting nested mean structures and/or nested covariance structures, assuming that the true data generating processes are known. Simulation results indicate that the efficient criteria performed better than their consistent counterparts when covariance structures used in data generation were complex, and worse when structures were simple. The conditional LRT under full maximum likelihood (FML) estimation was better overall than the other criteria in terms of selection performance. However, under restricted maximum likelihood (REML), estimation was inferior. We also find that the strategy of using REML for covariance structure selection and FML for mean structure selection suggested in the statistical literature may be misleading.

Keywords: Mixed linear model; Repeated measures; Model selection criteria.

Resumen

Un marco teórico potente resulta clave para especificar el modelo mixto que explica mejor la variabilidad de datos longitudinales. A falta de teoría, la mayoría de las investigaciones realizadas hasta la fecha, se ha centrado en ajustar la matriz de dispersión usando criterios de selección de modelos para elegir entre estructuras de covarianza no anidadas. En este trabajo, comparamos el desempeño del estadístico razón de verosimilitud (LRT) condicional & de varias versiones de los criterios de información para seleccionar estructuras de medias y/o de covarianzas anidadas, asumiendo conocido el verdadero proceso generador de datos. Los resultados numéricos indican que los criterios de información eficientes funcionaban mejor que sus homólogos consistentes cuando las matrices de dispersión usadas en la generación eran complejas y peor cuando eran simples. Globalmente, el desempeño del LRT condicional basado en el estimador de máxima verosimilitud completa (FML) era superior al resto de los criterios examinados. Sin embargo, el desempeño era inferior cuando se basaba en el estimador máxima verosimilitud restringida (REML). También encontramos que la estrategia sugerida en la literatura estadística de usar el estimador REML para seleccionar la estructura de covarianza y el estimador FML para seleccionar la estructura de medias debería ser evitada.

Palabras clave: Modelo lineal mixto; Medidas repetidas; Criterios de selección de modelos.

Currently, disciplines using approaches based on the mixed linear model theory to analyze data with a hierarchical structure are increasingly frequent (Vallejo, Arnau, & Bono, 2008a). Moreover, in the past three decades, the important theoretical development of these models and their definitive establishment was promoted by the incorporation of specific analytical procedures in the main professional statistical packages, including the *Proc Mixed* in the SAS model, the *lme* function in the S-PLUS/R or the *Mixed* and *xtmixed* commands in the SPSS and STATA, respectively. Mixed linear models allow the analysis of longitudinal and cross-sectional data, although they are especially useful to deal with temporal data, because they permit adjusting and making inferences about the mean structure (as in the classical univariate and multivariate approaches of repeated measures) and modeling the covariance structure (in terms of random effects and pure error). Rather than assuming an overly simple dispersion matrix (e.g., the compound symmetry matrix -CS- typical of the univariate approach) or a completely general one (e.g., the unstructured matrix -UN- typical of the multivariate approach), this approach seeks balance between the criteria of flexibility and parsimony or scientific simplicity (Ato & Vallejo, 2007). In this regard, it is noted that if a researcher specifies an excessively simple model, there is a risk of making erroneous inferences, due to underestimation of the standard errors. If, in contrast, an excessively complex model is formulated, there is a risk of making inefficient inferences.

In order to improve the quality of the inferences obtained with the mixed model approach, it is essential to model two different aspects of the data (Littell, Pendergast, & Natarajan, 2000). On the one hand, the fixed effects used to describe the mean responses as a function of time (hereafter, mean structure). And, on the other, the random effects used to describe the variation among the within-subject repeated measures (hereafter, covariance structure). As the form of the covariance structure depends on the selection of the mean structure (Fitzmaurice, Laird, & Ware, 2004), when both of these structures are modeled effectively, more exact (with less bias) and precise (with less variance) estimations are obtained from the parameters. Vallejo, Ato, and Valdés (2008b) confirm the importance of identifying the true process of generating data (PGD). In this study, the errors rates based on the true PGD never exceeded its nominal value. However, the standard errors were biased when the true PGD was erroneously specified.

Selection of the best model is central to interpret the data adequately but this goal is difficult to achieve because, for the same sample evidence, there are multiple candidate models (Claeskens & Hjort, 2008). To facilitate modeling the covariance matrix, SAS, , probably the most popular and versatile of all the currently existing programs (Feng, Zhou,

Zhang, & Zhang, 2009), and other statistical programs, incorporate a complete structure menu. For example, *Proc Mixed* allows matching and comparing compound symmetry models, sphericity models, autoregressive models, moving-average models, autoregressive and integrated moving-average models, antedependent and unstructured models (for specific details, see Zimmerman & Núñez-Antón, 2009). *Proc Mixed* also allows specifying heterogeneous covariance structures within and across groups, which forestalls having to accept the equicorrelation of the observations and the homogeneity of the dispersion matrixes.

There are diverse criteria to determine the goodness of fit of the model selected during the modeling process. To compare nested models (one model is obtained from another model by manipulating parameters), the most frequent criterion is the likelihood-ratio test (LRT) with a deviance obtained from the full maximum likelihood function (FML) or restricted/residual maximum likelihood function (REML), depending on whether one chooses between models with an identical covariance structure or means structure (Kreft & de Leeuw, 1998). Less formal statistical tools are also habitually used, such as the Information Criterion (IC) of Akaike (AIC), the Corrected AIC (AICC), the Consistent AIC (CAIC), the Bayesian Information Criterion (BIC) and the Hannan-Quinn Information Criterion (HQIC), as well as diverse versions derived from them. Particularly, the AIC and BIC criteria, as they are both implemented in most of the programs that adjust mixed models; the specific HLM and MLwiN programs are an exception to this. The origin of the ICs is different, but their structure is similar; in fact, they differ in the weight they assign to the penalty factor (Lee & Ghosh, 2009). To some extent, they all penalize the logarithm of the likelihood function because of the number of parameters, most of the times from the marginal formulation of the model (random effects are explicitly ignored when modeling the variation of multilevel data), and they select the model that minimizes their value. Vaida and Blanchard (2005) and Liang, Wu, and Zou (2008) provide details of the performance of the AIC criterion using a hierarchical formulation of the model.

Other selection criteria, such as the adjusted determination coefficient, (R^2_{adj}), the concordance correlation coefficient (CCC) or the prediction residual sum of squares (PRESS) have received scarce attention. Nevertheless, in one of the few studies that has examined the performance of the predictive criteria (based on the fit of the predicted values) using the marginal and hierarchical formulation of the model, Wang and Schaalje (2009) report that the R^2_{adj}, CCC, and PRESS criteria do not perform better than the AIC and BIC criteria. The comparison was between two nested models with identical covariance structure. Technical

details of the predictive criteria are provided by Orelien and Edwards (2007), Schabenberger (2004), and Vonesh, Chinchilli, and Pu (1996).

As a function of their asymptotic properties, ICs can be classified in two categories: (a) efficient criteria, such as AIC or AICC and (b) consistent criteria, such as BIC, CAIC or HQIC. A criterion is said to be efficient if the discrepancy between the true PGD and the model specified to approximate it decreases as sample size increases. A criterion is said to be consistent if the probability of choosing the correct model increases with sample size. Efficient criteria are based on the hypothesis that the true PGD is an infinite dimension, and they select the best finite dimension model. In contrast, consistent criteria are based on the hypothesis that the true PGD is a finite dimension, and they tend to select it when sample size tends to infinitity. When using the concept of asymptotic efficiency, it is not assumed that the true PGD is included in the family of investigated models. However, when using the concept of asymptotic consistency, this implies the hypothesis that the true PGD belongs to the class of models under consideration, which may be false.

Content analyses show that the most frequently used ICs used to select models with identical mean structures are the AIC and BIC (Littell et al., 2000). The performance of these criteria has been examined by diverse authors, including Ferron, Dailey, and Yi (2002), Gomez, Schaalje, and Fellingham (2005), Keselman, Algina, Kowalchuk, and Wolfinger (1998), and Vallejo et al. (2008b). Except for the study of Ferron et al. (2002), where the AIC identified the true PDG 79% of the times and the BIC did so 66% of the times, the remaining studies critically support the suggestion of Littell et al. (2000) of modeling the covariance structure with these criteria, , especially by means of the BIC. In the study of Keselman et al. (1999), the AIC selected the correct structure 47% of the times and the BIC, 35% of the times; in the study of Vallejo et al. (2008b), the AIC made the correct selection 68% of the times and the BIC, 48% of the times, whereas in the study of Gomez et al. (2005), both criteria were correct 22% of the times. Although the performance depended on the manipulated conditions, in all the studies, it was shown that selection improved as sample size increased and matrix complexity decreased.

A more complete study is that carried out by Gurka (2006). This researcher examined the performance of the AIC, AICC, BIC, and CAIC criteria in terms of selecting the model with a correct growth curve under diverse conditions, including different forms of calculating the criteria and different methods of parameter estimation. The ICs were assessed under three different scenarios as a function of their ability to: (a) select the correct mean structure from three possible models, given a CS matrix; (b) select the correct covariance structure from

three possible random effects with the same mean structure; and (c) select the correct model from six models resulting from the combination of three mean structures with two covariance structures. The results obtained by Gurka show that, among other things, the IC based on the REML method selected the true mean model as well as or better than the IC based on the FML method, which is surprising when taking into account that the specialized statistical literature (Molenberghs & Verbeke, 2001; Littell et al., 2006; Singer & Willet, 2003) defends adjusting that structure exclusively via FML. Gurka also found that the performance of the efficient criteria based on the REML estimator improved when excluding the $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ term (hereafter $REML_2$) from it; in contrast, the performance of the consistent criteria improved when this term was retained ($REML_1$). The global results reveal that the consistent criteria selected the true PGD more than 89% of the times, versus the efficient criteria, at 81%.

The findings of Gurka (2006) fully affect the process of model selection, as they reveal diverse inconsistencies in the statistical literature and in the documentation of commercial programs, both with reference to the methods of parameter estimation and in the formulas used to calculate the ICs. Nevertheless, we note that their studies are based on excessively simple scenarios, which limits the scope of their results. Before proceeding to generalize Gurka's results, it would clarify matters if we investigate the performance of the ICs, contemplating the above-mentioned analytical improvements, when manipulating the distribution of the error term, the complexity of the structures used to generate data, and the number of models included in the selection process.

It is reasonable to think that a large part of the good performance found in the study of Gurka (i.e., all the versions of the IC examined selected the true covariance structure more than 90% of the times) is explained by the form of the manipulated matrixes, completely general *versus* excessively simple ones, and by the reduced number of alternatives involved in the selection process. In principle, assuming we know the true PGD, we can expect that the number of models included in the comparison will affect the efficient criteria more than the consistent criteria, because the former assume that the true PGD is an infinite dimension; nevertheless, it is evident that it is not the same for any criterion to select a model from two alternatives as from two dozen alternatives. Moreover, to what extent is it realistic to assume that the variance of the observations remains constant and/or that the correlation does not decrease over time when the growth curves are studied?

Therefore, the present work has the goal to determine how effective the AIC, AICC, BIC, CAIC and HQIC criteria are to discover the true PGD in a family of nested models. These criteria will be assessed under FML and $REML_1$/$REML_2$ estimation when manipulating diverse mean and/or covariance structures. Further, to provide a reference point for the comparison, we shall also use the LRT conditional adjustment criterion. In order to meet the proposed goal, we shall use various crossover designs in which the assumptions of data normality and homogeneity of the dispersion matrixes are violated separately and concurrently.

Definition of the tools used to select the best mixed model

To use the methodology of the mixed model in the longitudinal context implies having to select from alternative models to explain the variability of the data in the simplest possible way. Although there is no agreement about the best way to select the best model, tools such as the IC and LRT are frequently used.

Information Criteria (ICs)

In Table 1 are defined the versions of the ICs investigated, both under FML estimation and under $REML_1$/$REML_2$ estimation. The formulas used by the *Proc Mixed* of the SAS module (version 9.2, 2008) and by the *Mixed* of the SPSS command (version 17, 2008) are also indicated.

Table 1. Definition of the information criteria used to select the mixed model

FML estimation method	REML estimation method
$\text{AIC} = -2\log l_{FML} + 2(p+q)^{\text{SAS,SPSS}}$	$\text{AIC}_1 = -2\log l_{REML1} + 2q$
	$\text{AIC}_2 = -2\log l_{REML2} + 2q^{\text{SAS,SPSS}}$
$\text{AICC}_1 = -2\log l_{FML} + 2(p+q)\left(\frac{N}{N-p-q-1}\right)^{\text{SAS,SPSS}}$	$\text{AICC}_1 = -2\log l_{REML1} + 2q\left(\frac{N-p}{N-p-q-1}\right)$
$\text{AICC}_2 = -2\log l_{FML} + 2(p+q)\left(\frac{n}{n-p-q-1}\right)$	$\text{AICC}_2 = -2\log l_{REML1} + 2q\left(\frac{n}{n-q-1}\right)$
	$\text{AICC}_1 = -2\log l_{REML2} + 2q\left(\frac{N-p}{N-p-q-1}\right)^{\text{SAS,SPSS}}$
	$\text{AICC}_2 = -2\log l_{REML2} + 2q\left(\frac{n}{n-q-1}\right)$
$\text{BIC}_1 = -2\log l_{FML} + (p+q)\log(N)^{\text{SPSS}}$	$\text{BIC}_1 = -2\log l_{REML1} + q\log(N-p)$
$\text{BIC}_2 = -2\log l_{FML} + (p+q)\log(n)^{\text{SAS}}$	$\text{BIC}_2 = -2\log l_{REML1} + q\log(n)$
	$\text{BIC}_1 = -2\log l_{REML2} + q\log(N-p)^{\text{SPSS}}$
	$\text{BIC}_2 = -2\log l_{REML2} + q\log(n)^{\text{SAS}}$
$\text{CAIC}_1 = -2\log l_{FML} + (p+q)[\log(N)+1]^{\text{SPSS}}$	$\text{CAIC}_1 = -2\log l_{REML1} + q[\log(N-p)+1]$
$\text{CAIC}_2 = -2\log l_{FML} + (p+q)[\log(n)+1]^{\text{SAS}}$	$\text{CAIC}_2 = -2\log l_{REML1} + q[\log(n)+1]$
	$\text{CAIC}_1 = -2\log l_{REML2} + q[\log(N-p)+1]^{\text{SPSS}}$
	$\text{CAIC}_2 = -2\log l_{REML2} + q[\log(n)+1]^{\text{SAS}}$
$\text{HQIC}_1 = -2\log l_{FML} + 2(p+q)\log[\log(N)]$	$\text{HQIC}_1 = -2\log l_{REML1} + 2q\log[\log(N-p)]$
$\text{HQIC}_2 = -2\log l_{FML} + 2(p+q)\log[\log(n)]^{\text{SAS}}$	$\text{HQIC}_2 = -2\log l_{REML1} + 2q\log[\log(n)]$
	$\text{HQIC}_1 = -2\log l_{REML2} + 2q\log[\log(N-p)]$
	$\text{HQIC}_2 = -2\log l_{REML2} + 2q\log[\log(n)]^{\text{SAS}}$

Note: p = number of parameters of the mean model; q = number of parameters of the covariance structure; n = total number of subjects; N = total number of observations; FML = full maximum likelihood estimator; REML_1/ REML_2 = residual maximum likelihood estimators with and without the term $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$.

Likelihood-ratio test (LRT)

As noted, the most frequently used goodness-of-fit statistic to compare nested models is the LRT. This contrast can be obtained from the following expression:

$$\Delta = -2[\hat{l}_{\text{reduced}(H_0)} - \hat{l}_{\text{full}(H_1)}],$$

where Δ is the deviance statistic and $\hat{l}_{\text{reduced}(H_0)}$ and $\hat{l}_{\text{full}(H_1)}$ the maximum values of the FML or REML function, depending on whether one is selecting between models with identical covariance or means structure, under the null and alternative hypothesis, respectively. Δ is distributed under H_0 as a function of χ^2_ν, where ν indicates the difference between the number of estimated parameters in the full model and in the reduced model.

In spite of the extended use of LRT, its use involves taking certain limitations into account. For example, it is only defined to compare nested models and only allows comparing two models at once. When there are more than two nested models, hierarchically procedures must be used to apply LRT (for more details, see Dayton, 2003). In contrast, it is important to note that ICs are valid to compare and select nested and non-nested models. Furthermore, they allow simultaneous comparison of a set of models.

Simulation Method

To assess the performance of the described methods, we carried out three simulation studies. In the first one, the covariance structure was kept constant and we modeled the mean structure. In the second one, we assumed that the mean structure was known and we modeled the covariance structure. In the third one, we modeled both structures at once. In each one, we used a crossover design with two treatments, two sequences and six (in the third study, also 12) one-week periods, in which the normality and multisample sphericity assumptions were violated separately and concurrently. The participants in the first group received the AAABBB treatment sequence, whereas those of the second group received the inverse sequence to counteract possible residual effects. On the basis of the above, we proposed the following model

$$y_{ij} = \beta_{00} + \beta_{01} G_j + u_{0j} + \beta_{10} T_{ij} + \beta_{11} G_j \times T_{ij} + u_{1j} T_{ij} + \beta_{20} CT_{ij} + \beta_{21} G_j \times CT_{ij} + u_{2j} CT_{ij} + e_{ij},$$

where $G_j = 0$ if the *i*-th participant was assigned to the AAABBB sequence for a six-week period and $G_j = 1$ if he was assigned to the BBBAAA sequence; T_{ij} indicates the week in which the response was recorded, and CT_{ij} indicates the change of linear tendency among the first three periods and the last three periods. The variables $T_{ij} \in \{1, 2, 3, 4, 5, 6\}$ and $CT_{ij} \in \{1, 2, 3, 1, 2, 3\}$ were centered with regard to their respective means, specifically $T_{ij}^{\bullet} = (T_{ij} - 3.5)$ and $CT_{ij}^{\bullet} = (CT_{ij} - 2)$.

Variables manipulated in the first study

In the first study, we assessed the performance of the selection criteria to select the true mean structure from a set of candidate models. This assessment was carried out under FML and $REML_1$/$REML_2$ estimation when the following variables were manipulated:

(a) *Type of model used to generate the data*. The adjustment of the mean structure implied selecting the true PGD from a set of six nested models. In one half of the manipulated conditions, this process required adjusting a full model and in the other half, a reduced model. In both cases, the fixed effects of the model were defined by combining diverse design matrixes and different parameter vectors. In the first situation, besides the intercept, the model that generated the data (M_6) included as fixed effects the treatment groups (G), the linear tendency (T), the change in tendency (CT), the G × T interaction and the G × CT interaction. The remaining five models were erroneously specified by removing one or more covariates from the design matrix For example, the M_5 was erroneously specified by removing the covariate G × CT, whereas the M_1 was erroneously specified by removing the covariates G, T, CT, G × T, and G × CT, respectively. In the second situation, the model used to generate the data was the M_1. The remaining models were erroneously specified following the inverse process, that is, by adding variables to the design matrix. Table 2 shows the models used in the comparison process, as well as the value of the fixed effect parameters of the models that generate data. The coefficient vector of the full model, slightly modified, corresponds to that of an experiment described by Hedeker and Gibbons (2006; pp. 122-126).

Table 2. Models used to adjust the mean structure and value of the fixed effect parameters

$M_1^{©}$	$y_{ij} = \beta_{00} + u_{0j} + e_{ij}$
M_2	$y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + e_{ij}$
M_3	$y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$
M_4	$y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$
M_5	$y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$
$M_6^{©}$	$y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$

$\boldsymbol{\beta}'_{\text{completo}} = [\beta_{00} = 3.125 \;\; \beta_{01} = 1.25 \; \beta_{10} = -.20 \; \beta_{11} = 0.45 \;\; \beta_{20} = -.25 \; \beta_{21} = 0.50]$

$\boldsymbol{\beta}'_{\text{reducido}} = [\beta_{00} = 3.125 \;\; \beta_{01} = 0.00 \; \beta_{10} = 0.00 \; \beta_{11} = 0.00 \;\; \beta_{20} = 0.00 \; \beta_{21} = 0.00]$

$M_6^{©}$, $M_1^{©}$ = full and reduced models used to generate the data.

(b) *Size of the total sample*. Performance was investigated using two sample sizes: n = 30 and n = 60. These group sizes were selected because they are representative of those frequently found in psychological research. Within each sample size, the value of the sample coefficient of variation (Δ) was fixed at 0.33, where $\Delta = \frac{1}{\bar{n}}[\Sigma_j (n_j - \bar{n})^2 / J]^{1/2}$, and $\bar{n}$ was the mean size of the groups. When the design is balanced, $\Delta = 0$. For $n = 30$, the group sizes were: (10-20), (15-15) and (20-10) whereas for $n = 60$, the group sizes were: (20-40), (30-30) and (40-20).

(c) *Covariance patterns employed to generate the data*. Three patterns were used to generate the data: random linear coefficients (RLC), first-order heterogeneous autoregressive [ARH(1)] and UN. The first pattern is an example of a hierarchical model that allows modeling an intercept and one or more additional tendencies for each participant. This pattern can be very useful to characterize the data, as it unites flexibility and paucity, in fact, it is only necessary to estimate $1 + q(q+1)/2$ parameters. In this study, $q = 3$, the intercept, the linear tendency, and the change in tendency. The second pattern allows heterogeneous variances and exponentially decreasing covariances, but it assumes that the observations are equally spaced. This model is typical of short time series and requires estimating $(t+1)$ parameters. The third pattern represents the covariance structure that best fits the data, and moreover, it does not require the observations to be equally spaced. Nevertheless, it requires estimating $t(t+1)/2$ parameters.

(d) *Deviation from the sphericity assumption*. Although the mixed model does not assume equal variances of the differences between pairs of repeated measures (sphericity

assumption), empirical research has shown that the inferences made with this approach can be affected by the lack of sphericity (Vallejo, Fernández, Herrero, & Conejo, 2004). Therefore, covariance patterns with values of ε (index of absence of sphericity derived by Box) of .47 and .70 were used to investigate their effects on the performance of the selection criteria. The covariance structures used are available on the Website http://gip.uniovi.es/gdiyad/docume/Psicothema/.

(e) *Equality of the dispersion matrixes*. The performance of the selection tools was assessed both when the group covariance matrixes were homogeneous and heterogeneous. In the first case, the elements of the two dispersion matrixes were equal $(\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1)$, whereas in the second one, the elements of one of the matrixes were five times larger than those of the other matrix $(\mathbf{\Sigma}_2 = 5\mathbf{\Sigma}_1)$.

(f) *Pairing covariance matrixes and group size*. The way one relates the size of the groups and the size of the dispersion matrixes can have different effects on the statistical tests. When the design is balanced, the relation between the size of the dispersion matrixes and the size of the groups is null. When the design is unbalanced, the relation can be positive or negative. A positive relation implies that the smaller group is associated with a smaller dispersion matrix, whereas a negative relation implies that the smaller group is associated with a larger dispersion matrix.

(g) *Form of distribution of the measurement variable*. Although the approach of the mixed model is based on meeting the normality assumption, when dealing with real data, the skewness (γ_1) and kurtosis (γ_2) indexes are frequently different from zero (Micceri, 1989), which can lead one to interpret the results incorrectly. To investigate the effect of the distribution form on the performance of the selection criteria, we generated data from normal and nonnormal distributions by means of the distributions *g* and *h* introduced by Tukey (1977). In addition to the normal distribution $(g = h = 0; \gamma_1 = \gamma_2 = 0)$, we investigated three further distributions: (a) $g = 0$ and $h = .109$, a distribution with the same degree of bias and kurtosis as the double exponential or Laplace distribution or $(\gamma_1 = 0 \ \& \ \gamma_2 = 3)$; (b) $g = .76$ and $h = -.098$, a distribution with the same degree of bias and kurtosis as the exponential distribution $(\gamma_1 = 2 \ \& \ \gamma_2 = 6)$; and (c) $g = 1$ and $h = 0$, a distribution with the same degree of bias and kurtosis as the lognormal distribution $(\gamma_1 = 6.18 \ \& \ \gamma_2 = 110.94)$. The distributions *g* and *h* were obtained using the RANNOR function of the SAS. Thereby, we generated randomized normal standard variables (Z_{ijk}) and we transformed each one of them as

$Z^*_{ijk} = g^{-1}[\exp(gZ_{ijk}) - 1]\exp(hZ^2_{ijk}/2)$, where g and h are real numbers that control the degree of bias and of kurtosis. Lastly, to obtain a distribution with standard deviation σ_{jk}, each one of the scores that make up the dependent variable was created using the linear model $Y_{ijk} = \sigma_{jk} \times (Z^*_{ijk} - \mu_{gh})$, where $\mu_{gh} = \{\exp[g^2/(2-2h)] - 1\}/[g(1-h)^{1/2}]$ is the mean of the distribution g and h (for details, see Kowalchuk & Headrick, 2009).

Variables manipulated in the second study

In this study, we assessed the performance of the selection criteria when selecting the true covariance structure from a set of candidate models. This adjustment implied selecting the true PGD from a set of six nested patterns. In one half of the manipulated conditions, the variance of the model was kept constant over time, and the covariance decreased exponentially (AR(1)) and in the other half, a UN model was used. In the first situation, in addition to the AR(1) model used to generate the data, the set of candidate models included an independence model (IND), an ARH(1) model, a homogeneous Toeplitz model (TOEP), a heterogeneous TOEP model (TOEPH) and a UN model. The IND model assumes a constant variance and null serial covariance, whereas the TOEP and TOEPH models generalize, respectively, to the AR(1) and ARH(1) models. Diverse investigators, including Fitzmaurice et al. (2004), Littell et al. (2006), and Zimmerman and Núñez-Antón (2009), provide a detailed description of these models. Note that the structures are nested within each other, in the sense that IND is a special case of AR(1), which is a special case of TOPH, which, in turn, is a special case of TOEPH, and the last one is necessarily a special case of UN.

In the second situation, the set of candidate models was identical to the one described above, but the data were generated from the UN model. In addition to the estimation methods and covariance patterns used to generate the data, we also manipulated the variables sample size, equality of the dispersion matrixes, and the distribution form of the population. The covariance structures used are available in the above-mentioned Website.

Variables manipulated in the third study

To study the performance of the tests in more depth, we adjusted the mean structure and covariance structure simultaneously. This adjustment involved selecting the true PGD from a set of nine models. Table 3 shows the models used in the comparison and the value of the

fixed effect parameters used to generate the data. Upon examination, Table 3 shows that the models were nested within each other in the cases in which the number of fixed effects was identical, like in the $M_3 - M_4$ and $M_6 - M_7$ models, the covariance structures differed and were nested within each other, in the sense that AR(1) is a special case of ARH(1), which, in turn, is a special case of ANTE(1). For an exhaustive description of this last structure, see Zimmerman and Núñez-Antón (2009).

Table 3. Set of candidate mean and covariance models and value of fixed effect parameters

M_1	$E(y_{ij}) = \beta_{00}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR(1)}]$
M_2	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR(1)}]$
M_3	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR(1)}]$
M_4	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH(1)}]$
M_5	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH(1)}]$
M_6	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH(1)}]$
M_7	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE(1)}]$
$M_8^{©}$	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE(1)}]$
M_9	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}T_{ij}^{2\bullet} + \beta_{30}CT_{ij}^{\bullet} + \beta_{31}G_j \times CT_{ij}^{\bullet}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE(1)}]$

$\boldsymbol{\beta}' = [\beta_{00} = 1.00 \;\; \beta_{01} = 1.25 \;\; \beta_{10} = -0.50 \;\; \beta_{11} = 0.50 \;\; \beta_{20} = -0.50 \;\; \beta_{21} = 0.50]$

$\boldsymbol{\beta}' = [\beta_{00} = 1.00 \;\; \beta_{01} = 1.25 \;\; \beta_{10} = -1.00 \;\; \beta_{11} = 1.00 \;\; \beta_{20} = -1.00 \;\; \beta_{21} = 1.00]$

Note: $M_1 \subset M_2 \subset M_3 \subset M_4 \subset M_5 \subset M_6 \subset M_7 \subset M_8^{©} \subset M_9$; $M_8^{©}$ = model used to generate the data.

In this third study, in addition to the estimation methods, the variables sample size, number of repeated measures (t = 6 and t = 12), value of the fixed effect parameters, equality of the dispersion matrixes, and form of distribution were also manipulated. The values of the covariance parameters of the ANTE(1) matrix are available in the above-mentioned Website .

Results of the first study

Table 4 shows the percentage of times that the 29 criteria examined, 10 via FML and 19 via REML, selected the true mean model when the structure used to generate the data was ARH(1). The pattern of results corresponding to the RLC and UN matrixes is not presented. In the above-mentioned website, the pattern can be seen to be qualitatively and quantitatively similar to the one described. The data denote the mean percentage of correct selections across

the variables equality of the dispersion matrixes and sample size. Globally, the results indicate that:

1. The performance of the criteria examined depended on the form of the distribution of the measurement variable, the type of model to be selected, and the estimation procedure used. Although not presented in Table 4, the variables type of pairing, deviation from sphericity, and equality of the dispersion matrixes slightly affected the percentage of correct identifications and substantially affected sample size.
2. The performance of the ICs was superior under REML estimation than under FML estimation. When averaging across the 2592 (144 × 18) manipulated conditions, the percentage of correct selections obtained via REML was 86.5%, whereas that obtained via FML when averaging across 1296 (144 × 9) conditions was 52.3%. The opposite occurred with LRT.
3. On average, the differences between the ICs and the LRT under FML estimation were minimal. Specifically, the ICs correctly selected the full model 25.4% of the times, and the reduced model 79.2%, whereas the LRT selected the full model 26% and the reduced model 79.3% of the times. However, under REML estimation, the ICs correctly selected the full model 91.2% of the times, and the reduced model 81.8%, whereas the LRT selected the full model 12% and the reduced model 53.1% of the times, respectively.
4. With regard to the performance of the ICs under REML estimation, the consistent criteria selected the true mean model 89.4% of the times, and the efficient criteria 80.7%. Further, the performance of both types of criteria improved when the $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ term was included in the equation. The percentage of correct selections of the consistent criteria was 98.1% under $REML_1$ and of 81.2% under $REML_2$. The efficient criteria selected the true PGD 90% of the times under $REML_1$ and of 75.5% under $REML_2$.

Table 4. Percentage of times that the criteria selected the true mean model when the covariance pattern was ARH (1)

ME	Criterion	Full Model					Reduced Model				
		Norm	Lapla	Expon	Logn	Mean	Norm	Lapla	Expon	Lognor	Mean
FML	$AIC^{(SAS, SPSS)}$	59.24	49.31	38.69	30.26	**44.37**[1]	68.73	71.66	65.08	49.23	**63.68**
FML	$AICC_1$	26.54	17.37	12.33	08.21	**16.11**	93.61	94.69	91.54	61.78	**85.40**
FML	$AICC_2^{(SAS, SPSS)}$	54.23	44.43	33.85	25.34	**39.46**[2]	75.76	76.60	70.13	52.28	**68.69**
FML	$HQIC_1$	41.96	32.65	22.44	16.72	**28.44**	88.29	90.43	86.18	66.17	**82.77**
FML	$HQIC_2^{(SAS)}$	50.70	38.37	30.27	21.50	**35.21**[3]	78.63	82.12	77.64	63.09	**50.37**
FML	$BIC_1^{(SPSS)}$	23.46	15.72	07.69	05.15	**13.00**	95.97	97.38	95.15	82.23	**92.68**[2]
FML	$BIC_2^{(SAS)}$	38.82	29.10	19.28	13.02	**25.06**	89.55	91.74	87.44	69.50	**84.56**
FML	$CAIC_1^{(SPSS)}$	17.28	10.61	05.01	02.49	**08.85**	97.94	98.73	96.79	85.37	**94.70**[1]
FML	$CAIC_2^{(SAS)}$	30.45	21.30	11.96	07.12	**17.69**	93.99	95.93	92.75	76.83	**89.87**[3]
FML	LRT	38.53	29.71	20.94	14.74	**25.98**	85.23	87.09	82.10	62.79	**79.30**
$REML_1$	AIC_1	99.99	99.97	96.71	87.36	**96.01**	91.87	93.08	90.81	78.09	**88.46**
$REML_2$	$AIC_2^{(SAS, SPSS)}$	85.30	88.60	84.75	74.54	**83.30**	61.52	65.74	58.98	38.70	**56.23**
$REML_1$	$AICC_1$	99.99	99.49	99.04	94.12	**98.16**	93.59	94.12	92.51	80.02	**90.06**
$REML_1$	$AICC_2$	75.00	75.01	75.26	75.38	**75.16**	98.50	98.46	97.78	90.12	**96.22**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	85.28	88.54	89.98	80.29	**86.02**	69.69	71.58	62.53	41.49	**61.32**
$REML_2$	$AICC_2$	63.52	65.61	68.92	70.84	**67.23**	85.85	85.61	58.07	47.73	**69.31**
$REML_1$	$HQIC_1$	99.99	99.99	99.88	98.50	**99.59**	97.87	96.41	96.69	87.40	**94.59**
$REML_1$	$HQIC_2$	99.99	99.98	99.27	94.67	**98.48**	95.25	96.55	93.98	82.78	**92.14**
$REML_2$	$HQIC_1$	85.33	88.86	92.49	94.06	**90.18**	88.42	88.95	76.04	51.95	**76.34**
$REML_2$	$HQIC_2^{(SAS)}$	85.42	88.74	89.49	84.50	**87.04**	79.42	80.81	67.84	45.45	**68.38**
$REML_1$	BIC_1	99.99	99.99	99.99	99.74	**99.93**[3]	99.68	98.68	98.39	93.25	**97.50**[2]
$REML_1$	BIC_2	99.99	99.99	99.93	98.15	**99.52**	98.26	98.61	97.10	88.41	**95.60**
$REML_2$	$BIC_1^{(SPSS)}$	87.24	89.59	92.69	94.28	**90.95**	90.15	88.29	73.49	49.84	**75.44**
$REML_2$	$BIC_2^{(SAS)}$	85.38	88.72	91.59	91.91	**89.40**	87.98	86.39	71.79	40.39	**71.64**
$REML_1$	$CAIC_1$	99.99	99.99	99.99	99.96	**99.99**[1]	99.82	99.79	99.01	93.90	**98.13**[1]
$REML_1$	$CAIC_2$	99.99	99.99	99.99	99.75	**99.93**[2]	95.46	99.40	98.32	91.76	**96.23**[3]
$REML_2$	$CAIC_1^{(SPSS)}$	86.83	89.72	92.77	94.42	**90.94**	90.16	88.51	73.40	49.79	**70.46**
$REML_2$	$CAIC_2^{(SAS)}$	85.40	88.73	92.40	93.20	**89.93**	89.50	86.97	72.15	48.47	**74.27**
$REML_2$	LRT	15.39	12.05	10.65	09.94	**12.01**	41.68	55.08	58.68	56.97	**53.10**

Note: The data indicate the mean percentage of correct selections across the variables nature of pairing, sample size, deviation from sphericity, and dispersion matrix equality; ME= estimation method; FML = full maximum likelihood estimation; $REML_1$ = residual maximum likelihood estimation; $REML_2$ = residual maximum likelihood estimation removing the additive term; Norm = normal distribution; Lapla =Laplace distribution; Expon = exponential distribution; Logn = lognormal distribution.

Results of the second study

Table 5 shows the percentage of times that the criteria examined chose the true covariance structure when the data were generated from the AR(1) and the UN matrixes. The tabulated data show the mean percentage of correct choices across the variables matrix equality and sample size. Globally, the results indicate that:

1. The performance of the criteria examined depended on the covariance pattern used to generate the data and the form of distribution. The influence of the estimation procedures was lower. Although not shown in Table 5, the details of the variables dispersion matrix equality and sample size substantially affected the percentage of correct identifications when the UN model was used to generate the data, but they scarcely affected it when the AR(1) was used (see the afore-mentioned Website).
2. The performance of LRT was superior to that of the ICs, either under FML or REML estimation. On average, LRT selected the true PGD correctly 60% of the times (31% under ARH(1) and 89% under UN), whereas the ICs selected correctly 47.2% of the times.
3. When the AR(1) model was used to generate the data, the percentage of correct selections decreased as the data deviated from normality. The opposite phenomenon was produced with data generated under the UN model.
4. Independently of the estimation procedure used, the consistent criteria performed better than their efficient counterparts when the AR(1) model was used to generate the data. The percentage of correct selections of the consistent criteria was 73.2%, and that of the efficient criteria, 44.9%. In contrast, the opposite situation occurred when the UN model was used to generate the data. The consistent criteria chose the true PGD 18.1% of the times, and the efficient criteria, 52.7%. In contrast to the findings of the previous study, the performance of both types of criteria did not improve when the REML estimator included the $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ term.

Table 5. Percentage of times the criteria selected the true covariance structure under FML and REML estimation

		Autoregressive pattern [AR(1)]					Unstructured pattern (UN)				
ME	Criterion	Norm	Lapla	Expon	Logn	Mean	Norm	Lapla	Expon	Logn	Mean
FML	$AIC^{(SAS,\ SPSS)}$	77.91	53.77	12.60	04.21	**37.12**	56.37	59.18	86.06	92.92	**73.63**[2]
FML	$AICC_1$	72.22	67.06	33.82	17.00	**47.53**	25.79	26.13	35.19	39.19	**31.60**
FML	$AICC_2^{(SAS,\ SPSS)}$	84.33	64.80	17.80	06.74	**43.42**	32.61	34.07	63.59	77.57	**51.96**[3]
FML	$HQIC_1$	96.23	87.53	38.39	17.92	**60.02**	07.35	08.54	38.57	49.19	**25.92**
FML	$HQIC_2^{(SAS)}$	90.51	74.25	24.47	08.89	**49.53**	22.96	25.92	64.47	74.09	**46.86**
FML	$BIC_1^{(SPSS)}$	99.52	97.20	63.09	35.31	**73.78**[2]	00.05	00.13	08.02	14.25	**05.61**
FML	$BIC_2^{(SAS)}$	96.34	89.70	41.54	18.94	**61.63**	02.53	04.13	30.20	37.92	**18.70**
FML	$CAIC_1^{(SPSS)}$	99.89	98.73	72.42	43.71	**78.69**[1]	00.00	00.06	04.03	08.05	**03.04**
FML	$CAIC_2^{(SAS)}$	99.43	95.24	54.98	28.39	**69.51**[3]	00.18	00.76	14.01	20.17	**08.78**
FML	LRT	67.77	42.12	09.12	03.45	**30.61**	84.12	82.17	94.18	97.05	**89.38**[1]
$REML_1$	AIC_1	83.69	34.41	34.13	15.79	**42.00**	50.74	58.73	82.63	89.32	**70.36**[2]
$REML_2$	$AIC_2^{(SAS,\ SPSS)}$	83.69	34.41	34.13	15.79	**42.00**	50.74	58.78	82.62	89.31	**70.37**[3]
$REML_1$	$AICC_1$	86.44	39.60	45.58	25.24	**49.21**	32.85	38.22	65.83	76.65	**53.39**
$REML_1$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.14	41.29	44.16	**37.02**
$REML_2$	$AICC_1^{(SAS,\ SPSS)}$	86.44	39.59	45.57	25.24	**49.21**	32.85	38.27	65.83	76.65	**53.40**
$REML_2$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.13	41.29	44.36	**37.07**
$REML_1$	$HQIC_1$	97.93	68.99	77.26	54.75	**74.73**	04.97	09.48	36.21	53.47	**26.03**
$REML_1$	$HQIC_2$	93.68	52.42	58.38	33.53	**59.51**	18.76	25.73	61.52	75.25	**45.32**
$REML_2$	$HQIC_1$	97.93	68.66	77.22	54.73	**74.64**	04.47	08.14	35.88	53.28	**25.44**
$REML_2$	$HQIC_2^{(SAS)}$	93.68	52.42	58.38	33.53	**59.50**	18.76	25.69	61.48	75.23	**45.29**
$REML_1$	BIC_1	99.94	88.44	93.95	81.83	**91.04**[3]	00.00	00.04	06.87	17.28	**06.05**
$REML_1$	BIC_2	98.45	73.13	82.08	61.17	**78.73**	00.99	03.90	27.31	44.94	**19.29**
$REML_2$	$BIC_1^{(SPSS)}$	99.94	88.39	93.43	81.66	**90.86**	00.00	00.04	06.87	17.17	**06.02**
$REML_2$	$BIC_2^{(SAS)}$	98.45	73.21	82.08	61.16	**78.71**	00.99	03.89	27.31	44.91	**19.28**
$REML_1$	$CAIC_1$	99.99	92.51	96.20	87.41	**94.03**[1]	00.00	00.08	03.37	09.93	**03.35**
$REML_1$	$CAIC_2$	99.67	83.80	90.38	75.17	**87.50**	00.00	00.44	12.19	25.54	**09.55**
$REML_2$	$CAIC_1^{(SPSS)}$	98.98	92.49	96.18	87.40	**94.01**[2]	00.00	00.08	03.36	09.93	**03.35**
$REML_2$	$CAIC_2^{(SAS)}$	99.66	83.79	90.38	75.16	**87.25**	00.00	00.44	12.18	25.54	**09.55**
$REML_2$	LRT	78.11	25.50	17.60	07.87	**32.28**	81.45	82.78	92.57	95.46	**88.06**[1]

Note: The data indicate the mean percentage of correct selections across the variables sample size and dispersion matrix equality; Norm = normal distribution; Lapla =Laplace distribution; Expon = exponential distribution; Logn = lognormal distribution.

Results of the third study

Table 6 presents the tabulated percentage of times that the criteria examined correctly selected the mean and covariance structure, both under FML and REML estimation. The data show the mean percentage of correct selections across the variables sample size, dispersion matrix equality, and form of distribution. Globally, the results indicate that:

1. The performance of the criteria examined depended on the estimation method, the value of the parameters, and the number of repeated measures. Although not shown in Table 6, the details can be seen on the Website. The variable sample size substantially affected the selection of the true PGD and moderately affected the variables form of distribution, and dispersion matrix equality.
2. The performance of the ICs was better under REML estimation than under FML estimation. Averaging across the 1052 (64 × 18) manipulated conditions, the percentage of correct selections obtained via REML was 69.7% (of which 65.2% corresponds to the consistent criteria and 74.2% to the efficient criteria), whereas that obtained by FML averaging across 576 (64 × 9) conditions was 55.9% (of which 47.5% corresponds to the consistent criteria and 64.3% to the efficient criteria). The LRT selected the true PGD 73.1% of the times under FML estimation and 68.9% under REML estimation.
3. When the estimation method used was FML and t = 6, the differences between the ICs and the LRT did not exceed 2 percentual points. However, under REML estimation, the differences favored the ICs and were higher than 20 points. Surprisingly, the situation reversed when t = 12. In this case, the differences exceeded 10 percentual points.
4. Lastly, the performance of the ICs improved if the REML estimator included the $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ term but only when t = 6. It is noted that the performance of the consistent criteria was superior when using the formulas implemented in the *Proc Mixed* of the SAS module, instead of those implemented in the *Mixed* command of the SPSS. In Table 1, it is confirmed that the algorithms used by SAS and SPSS to calculate the efficient criteria are identical.

Table 6. Percentage of times the criteria correctly selected the fixed effect and random models under FML and REML estimation

		$t = 6$		$t = 12$		
ME	Criterion	$\boldsymbol{\beta}'_{(1)}$	$\boldsymbol{\beta}'_{(2)}$	$\boldsymbol{\beta}'_{(1)}$	$\boldsymbol{\beta}'_{(2)}$	Global
FML	$AIC^{(SAS,\ SPSS)}$	38.94[1]	67.01[2]	73.92	79.44	**58.1**[2]
FML	$AICC_1$	08.62	24.53	20.75	24.70	**36.7**
FML	$AICC_2^{(SAS,\ SPSS)}$	33.59[2]	64.31	74.74[3]	83.22[3]	**55.2**[3]
FML	$HQIC_1$	20.74	55.14	69.62	89.10	**52.4**
FML	$HQIC_2^{(SAS)}$	30.64	64.23[3]	77.03[2]	84.56	**51.7**
FML	$BIC_1^{(SSPSS)}$	05.54	25.84	33.32	70.72	**42.1**
FML	$BIC_2^{(SAS)}$	17.17	50.51	70.91	89.84[2]	**50.7**
FML	$CAIC_1^{(SPSS)}$	02.74	16.98	22.82	57.16	**39.2**
FML	$CAIC_2^{(SAS)}$	09.52	34.71	55.95	85.54	**46.5**
FML	LRT	31.89[3]	73.75[1]	87.93[1]	99.16[1]	**64.8**[1]
$REML_1$	AIC_1	81.51[1]	82.87	88.59	89.36	**78.0**[1]
$REML_2$	$AIC_2^{(SAS,\ SPSS)}$	74.21[2]	87.27[1]	92.68[1]	98.93[1]	**71.4**
$REML_1$	$AICC_1$	77.34[3]	81.38	88.25	89.39	**76.5**[3]
$REML_1$	$AICC_2$	42.99	54.64	44.14	44.48	**58.6**
$REML_2$	$AICC_1^{(SAS,\ SPSS)}$	63.58	85.08[2]	90.91[2]	98.91[2]	**69.9**
$REML_2$	$AICC_2$	37.22	56.31	46.10	49.43	**53.0**
$REML_1$	$HQIC_1$	59.98	74.79	79.13	89.59	**74.4**
$REML_1$	$HQIC_2$	73.21	81.04	87.47	89.22	**76.8**[2]
$REML_2$	$HQIC_2$	51.50	76.42	81.90	98.32	**70.1**
$REML_2$	$HQIC_2^{(SAS)}$	60.76	84.18[3]	91.03[3]	98.87[3]	**71.3**
$REML_1$	BIC_1	32.28	47.20	42.69	80.73	**66.0**
$REML_1$	BIC_2	54.90	73.73	80.09	87.12	**73.5**
$REML_2$	$BIC_1^{(SPSS)}$	27.13	47.99	42.52	87.77	**61.0**
$REML_2$	$BIC_2^{(SAS)}$	46.73	73.11	80.98	98.79	**68.1**
$REML_1$	$CAIC_1$	23.35	43.77	32.04	73.29	**63.6**
$REML_1$	$CAIC_2$	40.34	59.66	65.73	88.20	**70.0**
$REML_2$	$CAIC_1^{(SPSS)}$	20.16	35.62	32.11	79.41	**57.1**
$REML_2$	$CAIC_2^{(SAS)}$	34.51	58.58	63.99	97.62	**64.7**
$REML_2$	LRT	28.33	71.11	79.59	96.89	**57.7**

Note: The data indicate the mean percentage of correct selections across sample size, distribution form, and dispersion matrix equality. $\boldsymbol{\beta}'_{(1)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -0.50\ \beta_{11} = 0.50\ \beta_{21} = -0.50\ \beta_{21} = 0.50]$; $\boldsymbol{\beta}'_{(2)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -1.00\ \beta_{11} = 1.00\ \beta_{21} = -1.00\ \beta_{21} = 1.00]$, t = number of observation periods. The last column represents the mean percentage across the three experiments.

Global results

The global performance was 56.5% when using FML and 62.9% when using REML. The ICs selected the true PGD 48.1% of the times via FML and 68.7% via REML, whereas the LRT selected correctly 64.8% and 57.7% of the times, respectively. Under the FML estimation method, the efficient criteria selected the true PGD 51.3% of the times—58.1% the AIC and 44.5% the AICC—and the consistent criteria 47.1% of the times—46.4% the BIC, 42.9% the CAIC, and 52.1% the HQIC. In turn, under the REML method, the efficient criteria selected the true PGD 69.6% of the times—74.7% the AIC and 64.5% the AICC—and the consistent criteria 68.1% of the times—67.2% the BIC, 63.9% the CAIC, and 73.2% the HQIC.

Conclusions, recommendations and limitations

Many conclusions can be extracted from this study, but we consider it appropriate to note the following five conclusions:

- Firstly, the data showed that none of the procedures examined consistently selected the true PGD; however, their performance improved when increasing sample size, the number of repeated measures, and the magnitude of the parameters.
- Secondly, independently of the part of the model that was adjusted, the ICs based on the REML estimator selected the true PGD more times than the ICs based on the FML estimator. This result not only confirms and extends the findings of Gurka (2006), but it also questions the recommendation made in the specialized statistical literature of exclusively using ICs with REML to compare models with identical mean structures (Orelien & Edwards, 2007).
- Thirdly, the performance of the ICs improved if the REML estimator included the constant term, especially when the number of repeated measures was moderate. In contrast to the findings of the work of Gurka (2006), where the $REML_1$ estimator only improved the performance of the consistent criteria, in the current work, the performance of the efficient criteria also improved. This result coincides with that found by Wang and Schaalje (2009) when comparing the performance of the AIC and BIC criteria with that of the predictive criteria R^2_{adj}, CCC, and PRESS.
- Fourthly, the performance of the ICs was better when these criteria were calculated using the total number of subjects (level 2), instead of the total number of observations (level 1).

This finding, in addition to corroborating the results of Gurka (2006), also offers empirical support to the strategy followed in the *Proc Mixed* of the SAS (2008) module, as opposed to that followed in the *Mixed* of the SPSS (2008) command, of calculating the consistent criteria using the number of participants in level 2 of the hierarchical model.

- Fifthly, despite that the AIC (78%), AICC (76.5%), and HQIC (76.8%) criteria based on the $REML_1$ estimator selected the true PGD more frequently, when the covariance structure and the full model had to be adjusted, the performance of the LRT was as good as or better than that of the mentioned criteria.

To conclude, we wish to make one recommendation, one warning, and one suggestion. Globally, the efficient criteria performed better than the consistent criteria when the covariance structure was complex and vice versa when it was simple. The consistent criteria tended to select simpler models—generally of a stationary nature—than the efficient criteria. However, in applied longitudinal studies, the variance of the observations is habitually heterogeneous, and their correlations usually decrease over time; therefore, we decided to use the efficient criteria, in particular, the AIC based on the $REML_1$ estimator. In our opinion, this one is the best to meet the goal of finding a balance between a complex and a simple model. Having made this recommendation, we warn readers that the results are limited to the conditions examined, although we conjecture that they could be generalized to a broader range of conditions; for example, to situations where the models are not nested within each other. Lastly, in the investigation carried out, the true PGD always belonged to the family of models investigated. However, when dealing with real data, we do not know whether the true PGD belongs to the type of models considered. Therefore, it would be advisable to carry out an investigation with the goal of comparing the ICs in terms of selecting the model closest to the true PGD, as this model is not included in the set of models present in the comparison.

Acknowledgements

This work was financed by various Research Projects granted by the National I+D+I Plan of the Ministerio de Ciencia e Innovación. Ref.: PSI-2008-03624/PSIC and PSI2009-11136/PSIC.

References

Ato, M., & Vallejo, G. (2007). *Diseños Experimentales en Psicología*. Madrid: Pirámide.

Claeskens, G., & Hjort, N. L. (2008). *Model Selection and Model Averaging*. New York: Cambridge University Press.

Dayton, C. M. (2003). Model comparisons using information measures. *Journal of Modern Applied Statistical Methods, 2*, 281-292.

Feng, R., Zhou, G., Zhang, M., and Zhang, H. (2009). Analysis of Twin Data Using SAS. *Biometrics, 65*, 584-589.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*, 379-403.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley.

Gomez, V. E., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation, 34*, 377–392.

Gurka, M. J., & Edwards, L. J. (2008). Mixed models. En C. R. Rao, J. P. Miller, & D. C. Rao (Eds.): *Handbook of Statistics, Vol 27, Epidemiological and Medical Statistics* (pp. 253-280). New York: Elsevier Science.

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

Kowalchuk, R. K., & Headrick, T. C. (2009). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*. DOI:10.1348/000711009X42 3067.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: John Wiley.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics- Simulation and Computation, 27*, 591-604.

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

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

Liang, H., Wu, H., & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika 95,* 773- 778.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS System for Mixed Models*. 2nd edition. Cary, NC: SAS Institute Inc.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19,* 1793-1819.

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

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling, 1,* 235–269.

Orelien, J. G., & Edwards, L. J**.** (2007). Fixed-effect variable selection in linear mixed models using R^2 statistics**.** *Computational Statistics & Data Analysis, 52,* 1896-1907.

SAS Institute Inc. (2008). *SAS/STAT® Software: Version 9.2.* SAS Institute Inc., Cary, NC.

Schabenberger, O. (2004). Mixed model influence diagnostics. *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., Paper 189-29.

Singer, D. J, & Willet, J. B. (2003). *Applied Longitudinal Data Analysis.* New York: Oxford University Press.

SPSS for Windows (2008). Version 17, SPSS Inc., IL: Chicago.

Tukey, J.W. (1977). Modern techniques in data analysis. NSF-sponsored regional research conference at Southern Massachusetts University (North Dartmouth, MA).

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

Vallejo, G., Arnau, J., & Bono, R. (2008a). Construcción de modelos jerárquicos en contextos aplicados. *Psicothema, 20,* 830-838.

Vallejo, G., Ato, M., & Valdés, T. (2008b). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology, 4,* 10-21.

Vallejo, G., Fernández, P., Herrero, J., & Conejo, N. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema, 16,* 498-508.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-Fit in generalized nonlinear mixed-effects models. *Biometrics, 52,* 575-587.

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

Zimmerman, D. L., & Núñez-Antón, V. (2009). *Antedependence Models for Longitudinal Data*. London: Chapman & Hall/CRC.

7.1.2. Objetivo 2

En el ámbito de la Educación son habituales las intervenciones con Programas (para mejorar el rendimiento en determinadas materias, para desarrollar buenas técnicas de estudio, para lograr una óptima autorregulación del aprendizaje...) lo que requiere muchos datos de los estudiantes. Sin embargo dichos programas, al no ser obligatorios, suelen dar lugar a la presencia de registros incompletos, lo que produce problemas de datos ausentes en el análisis. De hecho, la pérdida de datos en las investigaciones es siempre una realidad y un problema a considerar. Así las cosas, se pueden distinguir dos patrones de datos faltantes, según la manera en la cual aparecen en la serie de medición (Fermín, Galindo, & Martín, 2007). Datos faltantes intermitentes, que se refieren a la situación en la cual una o varias unidades de investigación participan en algunas, pero no en todas las ocasiones del estudio. Y el otro tipo de datos faltantes es el abandono, el cual se refiere a la situación cuando una o varias unidades dejan el estudio prematuramente, después de haber participado en una o más ocasiones.

Los abandonos son considerados mediante una variable indicadora, R, que hace alusión a la ausencia y se denomina *mecanismo de abandono*. Para el j-ésimo estudiante, toma el valor cero cuando éste completa el estudio o r, con $2 \leq r \leq n$, si el estudiante es visto por última vez en la $(r\text{-}1)$-ésima ocasión de medición. Diggle y Kenward (1994) clasificaron el abandono en estudios longitudinales, siguiendo la terminología de Little y Rubin (1987), como sigue: (1) Abandono Completamente Aleatorio (MCAR), en el cual el mecanismo de abandono y el proceso de medición son independientes, esto es, la ausencia es independiente de cualquier covariable, medida o no medida relacionada al proceso de medición; (2) Abandono Aleatorio (MAR), también denominado Ignorable o No Informativo, en el cual el mecanismo de datos faltantes depende de covariables y al menos de una de las respuestas anteriores a la ausencia; cuando la ausencia depende exclusivamente de las covariables consideradas y no de las respuestas anteriores a la ausencia; Little (1995), acuñó el término "abandono dependiente de covariable" y, (3) Abandono No Aleatorio (MNAR), también llamado Abandono Informativo o No Ignorable, es este caso la ausencia está relacionada con las respuestas no observadas, esto es, esas que pudieron haber sido observadas si el estudiante no se ausenta.

Aunque en realidad parece imposible especificar correctamente el mecanismo de pérdida, diversos estudios (DeSouza, Legedza, & Sankoh, 2009; Verbeke & Molenberghs, 2000), han concluido que los MAR son un mecanismo realista para la

mayoría de las aplicaciones prácticas dado que los abandonos de los estudios se relacionan con bastante frecuencia con anteriores resultados en los estudios. Así las cosas, existen varios métodos para tratar los abandonos (Hogan, Jason, & Korkontzelou, 2004) y uno de ellos, ampliamente estudiado, es el de los modelos de selección. En éstos se especifica una distribución para los datos completos y luego se modela el mecanismo de abandono condicionado sobre los hipotéticos datos completos.

Ahora bien, cuando se realizan estudios longitudinales los datos faltantes no son el único inconveniente. Otro problema que puede estar presente en el análisis de los datos es la existencia de correlación, tanto entre las medidas tomadas en puntos diferentes de tiempo como entre las variables de respuestas de los estudiantes. Para poder paliar ambas problemáticas (los datos desequilibrados y la dependencia de las observaciones) y analizar de forma adecuada los datos presentes en un diseño de medidas repetidas, los modelos más apropiados son los Modelos Lineales Mixtos (Laird & Ware 1982).

Así, la contribución del estudio que a continuación se presenta descansa en el aporte de datos numéricos del rendimiento de los Criterios de Información en la selección del modelo de generación de los datos cuando el supuesto de normalidad es violado. De esta manera, y en consonancia con la investigación realizada previamente, el **Segundo Objetivo** de esta Tesis consiste en proporcionar una mayor comprensión de los Criterios de Información (AIC, AICC, BIC, CAIC y HQIC) cuando los tamaños de muestra y los modelos de covarianza son más complejos, ambos generando datos incompletos y comparando el desempeño de los Criterios de Información mediante un estudio de simulación Monte Carlo.

Paper II

Selecting the best unbalanced repeated measures model

Behavior Research Methods

e-ISSN 1554-3528
Volume 43
Number 1

Behav Res (2010) 43:18-36
DOI 10.3758/
s13428-010-0040-1

Springer

Behav Res (2011) 43:18–36
DOI 10.3758/s13428-010-0040-1

Selecting the best unbalanced repeated measures model

Guillermo Vallejo · M. Paula Fernández · Pablo E. Livacic-Rojas · Ellián Tuero-Herrero

Published online: 7 December 2010

Abstract This study examined the performance of selection criteria available in the major statistical packages for both mean model and covariance structure. Unbalanced designs due to missing data involving both a moderate and large number of repeated measurements and varying total sample sizes were investigated. The study also investigated the impact of using different estimation strategies for information criteria, the impact of different adjustments for calculating the criteria, and the impact of different distribution shapes. Overall, we found that the ability of consistent criteria in any of the their examined forms to select the correct model was superior under simple covariance patterns than under complex covariance patterns, and vice versa for the efficient criteria. The simulation studies covered in this paper also revealed that, regardless of method of estimation used, the consistent criteria based on number of subjects were more effective than the consistent criteria based on total number of observations, and vice versa for the efficient criteria. Furthermore, results indicated that, given a dataset with missing values, the efficient criteria were more affected than the consistent criteria by the lack of normality.

Keywords Information criteria · Longitudinal data · Maximum likelihood · Missing at random · Model selection · Restricted likelihood

This paper was accepted for publication in ***Behavior Research Methods*** on October 25, 2010.

G. Vallejo (✉) · M. P. Fernández
Department of Psychology, University of Oviedo,
Plaza de Benito Feijóo, s/n 33003,
Oviedo, Spain
e-mail: gvallejo@uniovi.es

P. E. Livacic-Rojas
Department of Psychology, University of Santiago of Chile,
Avenida Ecuador 3650, Estación Central,
Santiago, Chile

E. Tuero-Herrero
Department of Psychology University of Oviedo,
Plaza de Benito Feijóo, s/n 33003,
Oviedo, Spain

Introduction

Repeated measures data arise when an outcome variable of interest is measured repeatedly for each individual in the study over a period of time. A key characteristic of repeated measures is the missing data problem. In many applications of psychology in the real world; missing values are inevitable, often taking the dropout form (i.e., a subject has observations up to a time point and none after that). Little and Rubin (2002) defined three types of missingness mechanisms for dropouts: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). A dropout process is said to be MCAR when the missing data are independent both of observed and unobserved outcomes of the variable being analyzed (e.g., missing data occurring in follow-ups made after the treatment phase has ended), and said to be MAR if, conditional on the observed data (e.g., previous performance, lack of efficacy of assigned treatment or other subject characteristics), the missingness is independent of the unobserved data. A dropout process that is neither MCAR nor MAR is termed MNAR. Although in reality it is virtually impossible to properly specify the missingness mechanism, several studies, including those of Verbeke and Molenberghs (2000), Little and Rubin (2002), and DeSouza, Legedza, and Sankoh (2009), have concluded that MAR is a realistic mechanism for most

practical applications because the study dropouts are frequently related to previous outcomes in the study.

Another characteristic feature of repeated measures studies is that observations within the same subject may be correlated, and for this reason, methods that account for the correlation among responses are often used. The widely used models include linear mixed effects models described by Laird and Ware (1982) and regression models with structured covariance proposed by Jennrich and Schluchter (1986).

In order to improve the quality of the inferences obtained with these approaches (i.e., to obtain valid and powerful tests of the fixed-effects parameter), it is important to choose a model with an appropriate mean and covariance structure (Littell, Pendergast, & Natarajan, 2000; Fitzmaurice, Laird, & Ware, 2004). When an appropriate model is selected, the precision and accuracy of parameter estimates improve, particularly when missing data are present. However, it is not always easy to find the model generating the data, because there are multiple candidate models for the same sample evidence (Claeskens & Hjort, 2008).

Several general statistical software packages are available for the analysis of these models, including R/S-Plus, SAS, SPSS (now known as PASW), and STATA. Among them, SAS is the most widely used statistical package (Feng, Zhou, Zhang, & Zhang, 2009). It is very useful for fitting different covariance patterns and uses various criteria to select the best fitting model. When two or more models are nested within each other, the likelihood ratio test can be used to discriminate between them. When models are not nested, information criteria are the most-used tool. An important advantage of using model selection criteria is that they can be used for nested and non-nested comparisons.

A variety of information criteria have been proposed for modeling longitudinal data with correlated errors. Selection tools such as the Akaike Information Criteria (AIC; Akaike, 1974), the corrected AIC (AICc; Hurvich & Tsai, 1989), the Bayesian Information Criteria (BIC; Schwarz, 1978), the consistent AIC (CAIC; Bozdogan, 1987), and the Hannan-Quinn Information Criteria (HQIC; Hannan & Quinn, 1979), are computed automatically when using *Proc Mixed* in SAS. The *mixed* command in SPSS/PASW provides four criteria (HQIC is not available), while the *lme*() function in R/Splus and xt*mixed* command in STATA only provide the AIC and BIC. Other software packages, such as HLM and MLwiN, do not provide any information criteria for selecting the most parsimonious correct model.

The performance of information criteria when interest lies in mixed model selection has been examined empirically under three different scenarios: (a) with respect to their ability to select the correct mean model given a particular covariance structure (Gurka, 2006; Wang & Schaalje, 2009), (b) with respect to their ability to select the correct covariance structure when the mean model is known (Ferron, Dailey, & Yi, 2002; Gomez, Schaalje, & Fellingham, 2005; Gurka, 2006; Keselman, Algina, Kowalchuk, & Wolfinger, 1998; Vallejo, Ato, & Valdés, 2008), and (c) with respect to their ability to simultaneously select the correct mean and covariance structure (Gurka, 2006). All the studies cited compared the effectiveness of the AIC and BIC. Additionally, Gurka (2006) also compared the performance of the AICc and CAIC, and Vallejo et al. (2008) the performance of the HQIC. Although no consensus exists over what constitutes a good criterion for selecting the best model, given that the performance of the criteria depended on the conditions investigated, the results indicate that the selection improved as the sample size increased and the complexity of model decreased.

Among the studies mentioned above, Gurka's (2006) study is particularly relevant, since it offers a complete comparative review of performance criteria. Several findings from this study should be highlighted. First, the criteria based on the restricted/residual maximum likelihood (REML) log-likelihood function selected the true mean structure as well as or better than the criteria based on the maximum likelihood (ML) log-likelihood function, which contradicts the contention that REML-based information criteria are not appropriate for selecting the mean structure of the mixed model (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; Molenberghs & Verbeke, 2001; Singer & Willet, 2003). Second, the performance of information criteria in selecting the proper mean structure depended on the version of the REML function used. Third, the study provided valuable information with respect to discrepancies in the formulas for calculating the information criteria. It should be pointed out, however, that this simulation study was based on very simple scenarios (e.g., providing candidate models with few alternatives or using highly parsimonious covariance patterns to generate datasets). In words of Gurka (2006, p. 25) "the discussed simulation was admittedly limited, as a more sophisticated covariance model selection evaluation is worth an article by itself".

Therefore, the purpose of this article is to provide further insight into information criteria obtained using SAS's *Proc Mixed* when realistic sample sizes and more complex covariance models are used, both to generate incomplete data and compare the performance of criteria. A further contribution of this article is to provide some numerical evidence of the performance of information criteria in selecting the model generating the data when the normality assumption is violated. Although real data rarely conform to normality and missing data are common in longitudinal studies (Schafer & Graham, 2002), these topics have not been thoroughly considered. Because most of the information criteria are derived using asymptotic normal theory,

this study allowed their assessment under conditions that are often the norm rather than the exception in longitudinal data.

Model and notation

Let y_{ijk} represent the response for the ith subject ($i = 1,...,n_j$), from the jth group ($j = 1,...,g$) at the kth time point ($k = 1,...t_i$). Also, let $n = \sum_j^g n_j$ denote the total number of subjects enrolled in the study and $m = \sum_i^n t_i$ denote the total number of observations in the dataset. The subscript i means that no assumption of the complete data is being made.

In the context of repeated measures data, assume the data arise from the covariance pattern regression model

$$\mathbf{y}_i = \mathbf{X}_i\,\boldsymbol{\beta} + \mathbf{e}_i, \tag{1}$$

where $\mathbf{y}_i$ is a $(t_i \times 1)$ ector of observations for ith subject, $\mathbf{X}_i$ is a $(t_i \times 1)$ matrix of known explanatory variables for ith subject with rank p, $\boldsymbol{\beta}$ is a $(p \times 1)$vector of unknown regression parameters, and $\mathbf{e}_i$ is an $(t_i \times 1)$ error vector. For the (i, j)th subject, it is assumed that the vector $\mathbf{e}_i$ is normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}_i$ which is a function of a q-dimensional vector of unknown parameters $\boldsymbol{\theta}$. Jennrich and Schluchter (1986) and Littell et al. (2006) consider several possible forms for $\boldsymbol{\Sigma}_i$

Both ML and REML estimation are commonly used to obtain estimators for the parameters of model (1). Building on the normality assumption of the $\mathbf{e}_i$, minus 2 times the log-likelihood function based on the full data, apart from a constant, is

$$-2l_{ML}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log\,|\boldsymbol{\Sigma}_i| + \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2}$$

where $\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\left(\mathbf{X}_i'\,\widehat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i'\,\widehat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{y}_i\right)$ is the generalized least squares (GLS) estimator of vector $\boldsymbol{\beta}$, assuming $\widehat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta})$ is known. In practice, the ML estimators must be found from the data available by solving the Hessian matrix and score function using an iterative procedure. For details see Jennrich and Schluchter (1986) and Lindstrom and Bates (1988).

The ML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ have desirable large sample properties. However, in finite samples, the ML estimator $\boldsymbol{\theta}$ is biased. To circumvent this problem, Harville (1974) recommended the use of REML originally developed by Patterson and Thompson (1971) as a method of estimating variance components. Minus 2 times the restricted log-likelihood function, apart from a constant, takes the form

$$-2l_{REML}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\,|\boldsymbol{\Sigma}_i| + \sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'\,\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right) + \log|\sum_{i=1}^{n}\mathbf{X}_i'\,\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i| - \log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|, \tag{3}$$

where $\widehat{\boldsymbol{\beta}}$ is the GLS estimator. This is identical to the REML function given by Harville (1974) and Cooper and Thompson (1977), among others.

When discussing model selection criteria, it is natural to ask: should one use ML or REML? The major software mixed models fitting procedures, such as R/Splus *lme*(), SAS *Proc Mixed*, SPSS/PASW *mixed*, and Stata *xtmixed*, allow a choice between both methods of estimation, with the default option usually being REML. For a further discussion of this problem, the reader is referred to West, Welch, and Galecki (2007) and McCulloch, Searle, and Neuhaus (2008). It should be pointed out, however, that many current mixed model- fitting procedures, such as SAS's *Proc Mixed* (SAS, 2008), omit the last term on the right side of (3) in the computation of the REML likelihood, which could impact the computing of an information criterion for comparing models having different mean structures. In fact, Gurka (2006) suggested using the REML function which has the term $\log|\sum_i^n \mathbf{X}_i'\mathbf{X}_i|$, denoted as l_{REML_1}, for the consistent criteria in model selection, and using the REML function without the term $\log|\sum_i^n \mathbf{X}_i'\mathbf{X}_i|$, denoted as l_{REML_2}, for the efficient criteria.

Information criteria for model selection

There has been extensive research in model selection criteria over the last decade (see Azari, Li, & Tsai, 2006; Bozdogan, 2000; Jiang & Rao, 2003; Kitagawa & Konishi, 2010; Shang & Cavanaugh, 2008, among others). Many of these developments have focused on using intensive computing techniques and on imposing a penalty on the likelihood that not only is related to sample size but to the dimension of estimated parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ (e.g., with the use of the Fisher information matrix or generalized degrees of freedom). Despite their practical importance, they are not implemented in the widely used software programs, so they were not the focus of our attention. As indicated earlier, the purpose of this paper is to investigate the performance of information criteria currently available in statistical pack-

ages for choosing a model with an appropriate mean and covariance structure, instead of examining the effectiveness of new proposed criteria.

When applying model selection criteria to obtain the best linear mixed model, one has to decide on the likelihood and correct number of parameters to use. In the linear mixed model the interest is either in the unknown regression (fixed effects) parameters, which are assumed to be shared by all individuals, or in the subject-specific deviation from the population estimates (random effects), which are assumed to be unique to a particular individual. For example, suppose we are testing the efficacy of a new therapy to reduce anxiety disorders. When the focus is on the fixed effects we want to know whether the average rate of change over time is different between the two groups, however, when the focus is on the random effects we want to know how individual profiles change over time in each group.

In the fixed-effects focus, the random effect is a device for modeling the correlation of the responses for *i*th subject, and the model is often referred to as the marginal model with correlated errors or the population-averaged model without random effect (Gurka & Edwards, 2008). Then, the problem is closely related to a regression model selection problem with correlated errors, and use of the marginal model does not involve defining information criteria different from those derived in the context of linear regression models. When the random effects are themselves of interest, then the conditional or hierarchical model should be used, and the conventional information criteria may be inappropriate (Liang, Wu, & Zou, 2008; Vaida & Blanchard, 2005).

In this study, all considered criteria have a form that consists of two basic elements. According to Lee and Ghosh (2009), one term measures the goodness-of-fit (log-likelihood) of a model and the other term penalizes the log-likelihood for the model complexity. In the following subsections, we describe the information criteria family examined in this paper, which are summarized in Table 1.

Akaike's Information Criteria (AIC)

Akaike (1974) suggested that the Kullback-Leibler (K-L) discrepancy measure provides a natural criterion for ordering candidate data models. The basic idea behind this criterion is to find within a predefined family of candidate models the one that minimizes the information loss occurring when the fitted candidate model is used as an approximation of the true model. The AIC in "smaller-is-better" form is defined as

$$AIC = -2\, l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}}, \widehat{\theta}\right) + 2s \tag{4}$$

where $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}}, \widehat{\theta}\right)$ is the maximized log-likelihood function and $s = p + q$, with p and q representing the dimension of estimated parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\theta}$ under the given fitted candidate model. Here, the goodness-of-fit term, $-2\, l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}}, \widehat{\theta}\right)$, gauges how well the candidate model fits the data, and the

Table 1 True parameter values of the fixed effects used to generate the data

ML estimation		REML estimation	
Criteria	Definition	Criteria	Definition
AIC	$\widehat{l}_{ML} + 2s$	AIC_1	$\widehat{l}_{REML_1} + 2s$
AIC_2	$\widehat{l}_{REML_2} + 2s$		
$AICc_1$	$\widehat{l}_{ML} + 2s\left(\frac{m}{m-s-1}\right)$	$AICc_1$	$\widehat{l}_{REML_1} + 2s\left(\frac{m-p}{m-s-1}\right)$
$AICc_2$	$\widehat{l}_{ML} + 2s\left(\frac{n}{n-s-1}\right)$	$AICc_2$	$\widehat{l}_{REML_1} + 2s\left(\frac{n}{n-s-1}\right)$
		$AICc_3$	$\widehat{l}_{REML_2} + 2s\left(\frac{m-p}{m-s-1}\right)$
		$AICc_4$	$\widehat{l}_{REML_2} + 2s\left(\frac{n}{n-s-1}\right)$
BIC_1	$\widehat{l}_{ML} + s\log(m)$	BIC_1	$\widehat{l}_{REML_1} + s\log(m-p)$
BIC_2	$\widehat{l}_{ML} + s\log(n)$	BIC_2	$\widehat{l}_{REML_1} + s\log(n)$
		BIC_3	$\widehat{l}_{REML_2} + s\log(m-p)$
		BIC_4	$\widehat{l}_{REML_2} + s\log(n)$
$CAIC_1$	$\widehat{l}_{ML} + s[\log(m) + 1]$	$CAIC_1$	$\widehat{l}_{REML_1} + s[\log(m-p) + 1]$
$CAIC_2$	$\widehat{l}_{ML} + s[\log(n) + 1]$	$CAIC_2$	$\widehat{l}_{REML_1} + s[\log(n) + 1]$
		$CAIC_3$	$\widehat{l}_{REML_2} + s[\log(m-p) + 1]$
		$CAIC_4$	$\widehat{l}_{REML_2} + s[\log(n) + 1]$
$HQIC_1$	$\widehat{l}_{ML} + 2s\log[\log(m)]$	$HQIC_1$	$\widehat{l}_{REML_1} + 2s\log[\log(m-p)]$
$HQIC_2$	$\widehat{l}_{ML} + 2s\log[\log(n)]$	$HQIC_2$	$\widehat{l}_{REML_1} + 2s\log[\log(n)]$
		$HQIC_3$	$\widehat{l}_{REML_2} + 2s\log[\log(m-p)]$
		$HQIC_4$	$\widehat{l}_{REML_2} + 2s\log[\log(n)]$

For purposes of comparison, the formulas listed in the table use $s = p + q$. $REML_1$ is the residual maximum likelihood estimation including the last term of Eq. 3; $REML_2$ is the residual maximum likelihood estimation excluding the last term of Eq. 3; n is the total number of subjects; m is the total number of observations; p is the rank of known design matrix; q is the number of covariance parameters

penalty term, $2s$, measures the complexity that compensates for bias in the lack of fit. When REML estimation is used, $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right)$ in (4) is replaced by the maximized REML log-likelihood $l_{REML}\left(\mathbf{y}|\widehat{\theta}\right)$. Therefore, the number of parameters in (4) should be $s = q$, as used by *Proc Mixed* in SAS and *mixed* command in SPSS/PASW. However, *xtmixed* command in STATA and *lme*() function in R/Splus retain $s = p + q$, but for comparing between two models with the same fixed effects, this does not make any difference.

Corrected Akaike's Information Criteria (AICc)

The AIC provides us with an approximately unbiased estimator of the expected K-L discrepancy in settings where the sample size is large and the dimension of the model is comparatively small. However, in settings in which the sample size is small, it is known that the AIC is biased (Burnham & Anderson, 2002). To account for this potential source of bias, Hurvich and Tsai (1989), based on the initial work of Sugiura (1978), proposed an efficient criterion, the corrected AIC (AICc), that outperforms the AIC in small samples, while converging with the AIC in larger samples. The AICc in "smaller-is-better" form is defined as

$$AICc = -2\,l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right) + 2s\left(\frac{m^*}{m^* - s - 1}\right) \tag{5}$$

where $s = p + q$ and m^* is the total number of observations (m), as used by *Proc Mixed* in SAS and *mixed* command in SPSS/PASW. When REML estimation is used, $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right)$ in (5) is replaced by the maximized REML log-likelihood $l_{REML}\left(\mathbf{y}|\widehat{\theta}\right)$. Thus, the number of parameters in (5) should be $s = q$ and $m = m - p$ given that REML is based on $m - p$ observations. The AICc is sometimes defined where m^* is the number of subjects (n), under both ML and REML.

Bayesian Information Criteria (BIC)

Along with the AIC, another well-known and widely used tool in statistical model selection is the BIC. The criterion was derived by Schwarz (1978) from Bayesian theory as the solution to the identification problem. Unlike the AIC, which is designed to find the best approximating model to the unknown true data-generating model, the BIC is designed to find the most probable model given the data. The BIC in "smaller-is-better" form is defined as

$$BIC = -2\,l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right) + s\log\left(m^*\right) \tag{6}$$

where $s = p + q$. An issue in using the BIC is that m^* represents n, as used by SAS *Proc Mixed* or m, as used by R/Splus *lme*(), Stata *xtmixed*, and SPSS/PASW *mixed*. When REML estimation is used, $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right)$ in (6) is replaced by the maximized REML log-likelihood $l_{REML}\left(\mathbf{y}|\widehat{\theta}\right)$. In the case of REML, $s = q$ and $m^* = n$ in SAS *Proc Mixed,* and $s = q$ and $m = m - p$ in SPSS/PASW *mixed* procedure, while $s = p + q$ and $m = m - p$ in R/Splus *lme*() and STATA *xtmixed*, respectively.

Consistent Akaike's Information Criterion (CAIC)

It is known that the AIC tends to select overly complex models, and does not produce an asymptotically consistent estimate of model order. Bozdogan (1987) proposed another consistent version of the AIC that takes sample size into account, and corrects the liberal tendency of the AIC by penalizing over-parameterization. Just like the BIC, the CAIC points to the right model with probability of unity as the sample size increases. The CAIC in "smaller-is-better" form is defined as

$$CAIC = -2\,l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right) + s\left[\log\left(m^*\right) + 1\right] \tag{7}$$

where $s = p + q$. As with the BIC, an issue in using the CAIC is that m^* represents n, as used by SAS *Proc Mixed* or m, as used by SPSS/PASW *mixed*. When REML estimation is used, $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right)$ in (7) is replaced by the maximized REML log-likelihood $l_{REML}\left(\mathbf{y}|\widehat{\theta}\right)$. In the case of REML, $s = q$ and $m^* = n$ in SAS *Proc Mixed,* and $s = q$ and $m = m - p$ in SPSS/PASW *mixed* procedure. R/Splus *lme*() and Stata *xtmixed* do not compute this criterion.

Hannan-Quinn's Information Criterion (HQIC)

Since the AIC tends to identify values of p and q that are too large while the BIC tends to underestimate the optimal model, Hannan and Quinn (1979) developed a consistent criterion that slows the changes occurring in the penalty term upon varying the sample size. The HQIC in "smaller-is-better" form is defined as

$$HQIC = -2\,l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right) + 2s\log\left[\log\left(m^*\right)\right] \tag{8}$$

where $s = p + q$ and $m^* = m$. When REML estimation is used, $l_{ML}\left(\mathbf{y}|\widehat{\boldsymbol{\beta}},\widehat{\theta}\right)$ in (8) is replaced by the maximized REML log-likelihood $l_{REML}\left(\mathbf{y}|\widehat{\theta}\right)$. In the case of REML, $s = q$ and the reported value of m^* in SAS *Proc Mixed* is m–p R/Splus *lme*(), Stata *xtmixed*, and SPSS/PASW *mixed* do not compute this criterion.

Simulation study

To establish the relative merits of information criteria in selecting the best regression model with patterned covariance structure, we conducted a simulation study using a completely randomized design in which n subjects were randomly distributed in g treatment groups with t equally spaced measurements from each subject. As indicated by Algina and Keselman (2004), the design used in this work has appeared in the literature under a variety of names (i.e., group randomized trials design, randomized parallel-groups design, split-plot repeated measures design), however, in all situations the hypothesis of interest is whether there are differential rates of change over time. To investigate the performance of these criteria in a situation with missing data we manipulated the following variables: types of mean and covariance structures, number of repeated measurements per experimental unit, total sample size, population distribution shape, and estimation methods used to determine the appropriateness of criteria. Variables concerning the number of groups, pattern of missing data, and homogeneity of covariance matrices were held constant in this study.

Study variables

To compare the procedures, two different mean models were used to generate data: (1) a full model with a common intercept, a dummy variable indicating membership in one of two groups, a continuous covariate that took equally spaced values in each group, and an additional group × slope interaction; and (2) a reduced model with a common intercept, a linear trend, and an additional group covariate. Specifically, we used the following two equations to generate data for the response y_{ijk} measured at time k for subject i in group j:

$$y_{ijk} = \beta_0 + \beta_1 Trt_{ij} + \beta_2 Time_{ik} + \beta_3 Trt_{ij} \times Time_{ik} + e_{ijk},$$
$$y_{ijk} = \beta_0 + \beta_1 Trt_{ij} + \beta_2 Time_{ik} + e_{ijk}. \quad (9)$$

where Trt_{ij} denotes an indicator variable for subject i in treatment group j (i.e., $Trt = 0$ for placebo control; $Trt = 1$ for active treatment), and $Time_{ik}$ was coded 0 to t time points.

Each of the two regression models was paired with one of the following covariance structures: (a) first-order autoregressive covariance pattern (AR), (b) first-order autoregressive covariance pattern with variance heterogeneity within subjects (ARH), (c) Toeplitz covariance pattern with variance heterogeneity within subjects (TOEPH), and (d) unstructured covariance pattern (UN). The former structure is extensively used for modeling covariance in longitudinal data. However, the latter are potentially more realistic because the assumption of constant variance across time is relaxed. Because published research reports contain no large empirical studies in which experienced judges show the covariance patterns likely to be encountered in practice, it is difficult to known how typical or pervasive the chosen covariance structures are. Thus, though the structures considered in the simulation might be regarded as arbitrarily chosen they nonetheless coincide with conditions previously investigated by several authors, including Keselman, Algina et al. (1998), Gomez et al. (2005), and Vallejo et al. (2008), and they are broad enough to represent data sets that may be encountered in applied settings. In all cases, population values for covariance matrices were chosen so that the variances at each time point were relatively small at the beginning of the study, the traces were identical, and the deviations from sphericity were similar. As indicated above, covariance structures were assumed to be the same for all subjects in the two treatment groups ($\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}$) Details on these structures when the number of repeated measurements was eight ($t = 8$) are in Table 2.

A two-group parallel design containing either $t = 4$ or $t = 8$ repeated measures per subject was considered. In behavioral sciences research, more than four observations per subject are not typical; in fact, the mode is equal to 4, at least according to the survey results provided by Kowalchuk, Lix, and Keselman (1996), while a large number of repeated observations per subject would not be unusual in clinical trials. We initially planned to investigate $t = 10$, however, preliminary simulations indicated that using SAS Proc Mixed took an inordinate amount of time when $t = 10$. It is doubtful that this change substantially affected the results.

For each value of t, three total sample sizes were considered, $n = 30$ ($n_1 = n_2 = 15$), $n = 60$ ($n_1 = n_2 = 30$), and $n = 120$ ($n_1 = n_2 = 60$). These sample sizes were selected, in part, because they are typical of what is encountered in applied research (see. e.g., Keselman, Huberty et al. 1998). The small sample size yielded *a priori* statistical power of approximately 40% for the time main effect and the interaction between time and group at the 5% Type I error rate. For the same design effects, the moderate and large sample sizes provided power values of approximately 60% and 80%, respectively. Since our work assumed there were missing data with repeated measures due to dropout, analytic power calculations may not be readily available (Kleinman & Horton, 2010). For this reason, the values of betas given in Table 3 for the .40, .60, and .80 target powers were calculated using numerical techniques and SAS Proc Mixed. The selected set of regression coefficients under the full model represents a situation in which the rate of change is greater in the treatment group

Table 2 Parameter values of the covariance patterns used to generate the data

(a) First-Order Autoregressive (AR), 2 parameters

$\mathbf{\Sigma}_1 = Cov\left(e_{ijk}, e_{ijk'}\right) = \sigma^2$ if $k = k'$ and $\sigma^2\rho^{|k-k'|}$ if $k \neq k', \sigma^2 = 136.5; \rho = .8;$

(b) Heterogeneous First-Order Autoregressive (ARH), t+1 parameters

$\mathbf{\Sigma}_2 = Cov\left(e_{ijk}, e_{ijk'}\right) = \sigma_k\sigma_{k'}$ if $k = k'$ and $\sigma_k\sigma_{k'}\rho^{|k-k'|}$ if $k \neq k$

$\sigma_1^2 = 47; \sigma_2^2 = 48.5; \sigma_3^2 = 62; \sigma_4^2 = 87.5; \sigma_5^2 = 125; \sigma_6^2 = 174.5; \sigma_7^2 = 236; \sigma_8^2 = 309.5; \rho = .8;$

(c) Heterogeneous Toeplitz (TOEPH), 2t-1 parameters

$\mathbf{\Sigma}_3 = Cov\left(e_{ijk}, e_{ijk'}\right) = \sigma_k\sigma_{k'}$ if $k = k'$ and $\sigma_k\sigma_{k'}\rho_{|k-k'|}$ if $k \neq k$

$\sigma_1^2 = 47; \sigma_2^2 = 48.5; \sigma_3^2 = 62; \sigma_4^2 = 87.5; \sigma_5^2 = 125; \sigma_6^2 = 174.5; \sigma_7^2 = 236; \sigma_8^2 = 309.5;$

$\rho_1 = .8; \rho_2 = .7; \rho_3 = .6; \rho_4 = .5; \rho_5 = .4; \rho_6 = .3; \rho_7 = .2.$

(d) Unstructured (UN), $t(t+1)/2$ parameters

$\mathbf{\Sigma}_4 = Cov\left(e_{ijk}, e_{ijk'}\right) = \sigma_k^2$ if $k = k$ and $\sigma_{kk'}$ if $k \neq k'$,

$\sigma_1^2 = 47; \sigma_2^2 = 48.5; \sigma_3^2 = 62; \sigma_4^2 = 87.5; \sigma_5^2 = 125; \sigma_6^2 = 174.5; \sigma_7^2 = 236; \sigma_8^2 = 309.5;$

$\sigma_{12} = 43.0; \sigma_{13} = 45.9; \sigma_{14} = 51.3; \sigma_{15} = 57.5; \sigma_{16} = 63.4; \sigma_{17} = 68.5; \sigma_{18} = 72.4;$

$\sigma_{23} = 46.6; \sigma_{24} = 52.1; \sigma_{25} = 58.4; \sigma_{26} = 64.4; \sigma_{27} = 69.5; \sigma_{28} = 73.5; \sigma_{34} = 58.9;$

$\sigma_{35} = 66.0; \sigma_{36} = 72.8; \sigma_{37} = 78.6; \sigma_{38} = 83.1; \sigma_{45} = 78.4; \sigma_{46} = 86.5; \sigma_{47} = 93.4;$

$\sigma_{48} = 98.7; \sigma_{56} = 103.4; \sigma_{57} = 111.6; \sigma_{58} = 1118; \sigma_{67} = 131.9; \sigma_{68} = 139.4; \sigma_{78} = 162.2.$

The upper left 4×4 matrix was used when the number of repeated measures was four. For $t = 4$, $\sigma = 61.25$ in the case (a)

than in the control group, but at the first time point the groups do no differ. On the other hand, the selected set of regression coefficients under the reduced model represents a situation in the which the patterns of change in the mean response over time is the same in both groups, but the population mean response profiles are parallel. In other words, the regression coefficients of these two models were selected to obtain an ordinal interaction and equal slopes in

Table 3 True parameter values of the fixed effects used to generate the data

	Full model								Reduced model					
	$t = 4$				$t = 8$				$t = 4$			$t = 8$		
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_0	β_1	β_2
1–β≈.40 (n_1 = 15; n_2 = 15)														
AR	25	0.00	–1.45	–2.15	25	0.00	–1.20	–1.60	25	5.45	–1.45	25	4.95	–1.20
ARH	25	0.00	–1.40	–2.00	25	0.00	–1.20	–1.55	25	4.85	–1.40	25	4.20	–1.20
TOEPH	25	0.00	–1.30	–1.40	25	0.00	–1.10	–1.40	25	4.57	–1.30	25	4.20	–1.10
UN	25	0.00	–1.50	–2.15	25	0.00	–0.60	–0.75	25	4.15	–1.50	25	3.45	–0.60
1–β≈.60 (n_1 = 30; n_2 = 30)														
AR	25	0.00	–1.35	–1.90	25	0.00	–1.00	–1.50	25	5.10	–1.35	25	4.65	–1.00
ARH	25	0.00	–1.30	–1.80	25	0.00	–1.00	–1.45	25	4.35	–1.30	25	3.95	–1.00
TOEPH	25	0.00	–1.10	–1.60	25	0.00	–0.90	–1.30	25	4.20	–1.10	25	4.00	–0.90
UN	25	0.00	–1.30	–1.80	25	0.00	–0.50	–0.70	25	3.75	–1.30	25	3.00	–0.50
1–β≈.80 (n_1 = 60; n_2 = 60)														
AR	25	0.00	–1.20	–1.65	25	0.00	–0.90	–1.25	25	4.45	–1.20	25	3.85	–0.90
ARH	25	0.00	–1.15	–1.60	25	0.00	–0.90	–1.25	25	3.70	–1.15	25	3.55	–0.90
TOEPH	25	0.00	–1.00	–1.40	25	0.00	–0.80	–1.05	25	3.55	–1.00	25	3.50	–0.80
UN	25	0.00	–1.10	–1.50	25	0.00	–0.40	–0.60	25	3.35	–1.10	25	2.65	–0.40

the different groups. Both scenarios are of high scientific interest for investigating treatment effects in a repeated measures design.

Two population distributions were considered using the methods described in the subsequent section. The multivariate normal distribution with univariate skew (γ_1) and kurtosis (γ_2)equal to zero ($\gamma_1 = \gamma_2 = 0$) was the first. A moderately skewed distribution with shape parameters equivalent to those of an exponential distribution ($\gamma_1 = 2$ and $\gamma_2 = 6$) was the second. Level of non-normality examined was representative of that often encountered in educational and psychological research (Micceri, 1989).

Finally, the performance of the AIC, AICc, BIC, CAIC, and the HQIC for each of 96 data types was evaluated first under ML and then under REML estimation, using both the $REML_1$ and $REML_2$ function, again using the terminology in Gurka (2006). Therefore, the simulation plan used a $2^4 \times 3 \times 4$ factorial design, with 192 different cells or conditions.

Data generation

In each treatment group, Gaussian continuous longitudinal data were simulated using the method of Ripley (1987). This procedure involves the following two steps:

1) Generate pseudorandom observation vectors $\mathbf{z}_{ij}$ with $E(\mathbf{z}_{ij}) = \mathbf{0}$ and $Cov(\mathbf{z}_{ij}) = \mathbf{I}$ from a t-variate normal distribution, where $\mathbf{I}$ is the identity matrix. These vectors were obtained using the RANNOR function in SAS.
2) Create complete data sets $\mathbf{y}_{ij}$ by multiplying the vector $\mathbf{z}_{ij}$ by the Cholesky decomposition $\mathbf{L}_l$, that is, $\mathbf{y}_{ij} = \boldsymbol{\beta}_j + \mathbf{L}_l\mathbf{z}_{ij}$, where $\mathbf{y}_{ij}$ is a vector of length t for the (i, j) th subject, $\boldsymbol{\beta}j$ is a p-dimensional vector containing the population fixed effects, and $\mathbf{L}_l$ is a lower triangular matrix of dimension t satisfying $\boldsymbol{\Sigma}_l = \mathbf{L}_l\mathbf{L}_l^{'}, l = 1, \ldots 4$, as presented in Table 2.

Then, longitudinal data with a monotone missing pattern were generated according to a specific type of MAR dropout model, namely, after the first time point, nonresponse at a given time point is a function of the response at the previous occasion, which was always observed. Specifically, if the value of the dependent variable at occasion k was lower than a cutoff value λ (observed values of outcome variable measure at occasion k-1) and if U_k (a uniform random variable) was less than a probability determined for γ (an expected proportion of missing data), then the subject dropped out at the occasion k and the subsequent occasions. The constants λ and γ were chosen to yield approximate average cumulative dropout rates of 0, 10, 20, and 30% in the $t = 4$ condition, and average cumulative dropout rates of 0, 5, 10, 15, 20, 25, 35, and 50% in the $t = 8$ condition. These two dropout rates were assumed to be constant across time points and distributed evenly between the two groups.

Multivariate non-normal data were generated through the power method developed by Fleishman (1978) and extended to the multivariate situation by Vale and Maurelli (1983). The steps used to implement this procedure can be summarized as follows:

1) Calculate a weight vector $\mathbf{w} = [abcd]^{'}$ with the desired skew and kurtosis for each distribution using Fleishman's power transformation method.
2) Compute an appropriate intermediate correlation matrix $\mathbf{R}_l$ by solving for all possible pairs of repeated measurements with the following third-order polynomial equation: $R_{X_kX_{k'}} = \rho_{Z_kZ_{k'}}(b^2 + 6bd + 9d^2) + \rho^2_{Z_kZ_{k'}}2c^2 + \rho^3_{Z_kZ_{k'}}6d^2$, where $\rho_{Z_kZ_{k'}}$ is the correlation coefficient between any two standard normal variables and $X_k(= a + bZ_k + cZ_k^2 + dZ_k^3)$ and $X_{k'}(= a + bZ_{k'} + cZ_{k'}^2 + dZ_{k'}^3)$ are the two correlated non-normal variables.
3) Factorize the matrix $\mathbf{R}_l$ to generate a vector of multivariate normal random variates with the prescribed $\rho_{Z_kZ_{k'}}$, that is, $\mathbf{x}_{ij} = \mathbf{M}_l\mathbf{z}_{ij}$, where $\mathbf{x}_{ij}$ denotes the vector of transformed variates with $E(\mathbf{x}_{ij}) = \mathbf{0}$ and $Cov(\mathbf{x}_{ij}) = \mathbf{R}_l$, and $\mathbf{M}_l$ is the lower triangular matrix in the Cholesky decomposition with the property $\mathbf{R}_l = \mathbf{M}_l\mathbf{M}_l^{'}$.
4) Transform the generated multivariate normal variates to the desired distributional shape and the desired population fixed effects and variances, that is, $\mathbf{y}_{ij} = \boldsymbol{\beta}_j + \mathbf{D}_l\left(\mathbf{X}_{ij}^{\bullet}\mathbf{w}\right)$, where $\mathbf{D}_l$ is a diagonal matrix containing the standard deviations of covariance matrix $\boldsymbol{\Sigma}_l$ defined in Table 2 and $\mathbf{X}_{ij}^{\bullet} = \left[\mathbf{1}_K\mathbf{x}_{ij}\mathbf{x}_{ij}^2\mathbf{x}_{ij}^3\right]$.

Simulated data were generated under each of the 192 conditions manipulated in the study, and a range of 36 models was fitted to each simulated condition. In particular, the two models referred to above for generating data, and a null model, were combined with twelve different covariance structures. These structures included: compound symmetry (CS), heterogeneous compound symmetry (CSH), linear random coefficients (RCL), first-order autoregressive (AR), heterogeneous first-order autoregressive (ARH), first-order factor analytic (FA), first-order autoregressive moving-average (ARMA), Huynh-Feldt (HF), Toeplitz (TOEP), heterogeneous Toeplitz (TOEPH), first-order antedependence (ANTE), and unstructured (UN). For all 36 possible models, an optimal model was selected computing the five information criteria and their examined variations. As shown in Table 1, nine used ML estimation and eighteen used REML estimation.

This process was repeated 5,000 times for each condition using a SAS macro. To evaluate the performance of these

information criteria we recorded the percentage of times each one of these criteria detected the model generating the data. When the estimation failed to converge, it was assumed that none of the criteria selected that model.

Simulation results

In order to assess the performance of the consistent criteria (BIC, CAIC, and HQIC) and their efficient counterparts (AIC and AICc) in selecting the best regression model with patterned covariance, we computed the percentage of times these criteria selected the model generating the data under ML and under REML, using both REML estimation with the term $\log|\sum_i^n \mathbf{X}_i'\mathbf{X}_i|$ (denoted as REML_1) as well as without it (denoted as REML_2). For comparison, we also considered the variations of five criteria based on the total sample size (n) and the total number of observations (m). Tables 4, 5, 6 and 7 display the numerical results obtained from Monte Carlo simulations. Tables 4 and 5 report the results when data were both multivariate normal and non-normal (i.e., exponential-type data) in form and the full model was used to generate the data, whereas Tables 6 and 7 present the results when data were both multivariate normal and non-normal in form and the reduced model was adopted.

Performances associated with normal data for full model

The percentages of correct decisions when the data were normally distributed and the full model was adopted are found in Table 4. These results can be summarized as follows:

1) Simulation results show that no one criterion selected the model generating the data consistently nor performed uniformly better than the others. Averaging across the covariance structures, total sample size, number of repeated measurements, and methods of estimation, the overall success was 56.4% for the AIC, 48.3% for the AICc, 39.1% for the BIC, 32.6% for the CAIC, and 52.1% for the HQIC.
2) In general, the performance of the consistent criteria (BIC, CAIC, and HQIC) was superior to efficient criteria (AIC and AICc) for relatively simple covariance patterns (i.e., AR and ARH), whereas efficient criteria were superior for more complex covariance patterns (i.e., TOEPH and UN). It was also found that the performance of the AIC and AICc was fairly comparable, however, the AICC did not perform better than the AIC when sample sizes were moderate to small. On the other hand, the CAIC gave results almost identical to those of the BIC, although overall the BIC performed slightly better than the CAIC.
3) All versions of the five criteria performed better for larger numbers of subjects and performed much better for designs in which the number of repeated measurements was large. For moderate and large sample sizes, it is very apparent that the effect of sample size was greater when the length of repeated measures was short. In particular, the performance at $n = 120$ and $t = 8$ was slightly better than at $n = 60$ and $t = 8$, better than at $n = 120$ and $t = 4$, and substantially better than at $n = 60$ and $t = 4$.
4) When comparing the consistent criteria based on n and the consistent criteria based on m, the former led to a considerably larger percentage of correct decisions, regardless of whether ML or REML estimation was used. On the other hand, the performance of the AICc did not show this behavior. In fact, the simulation results pointed towards the use of m under REML and n under ML.
5) Finally, it is important to highlight that the five criteria and their examined variations performed better under REML than under ML estimation. The average differences between the two methods of estimation were approximately 9 percentage points for the efficient criteria and approximately 13 percentage points for the consistent criteria. It is also noteworthy that the ability of criteria to select the correct model was superior under REML_1 than under REML_2. The average differences between the two versions of REML estimation were approximately nine percentage points for the efficient criteria and approximately 19 percentage points for the consistent criteria.

Performances associated with non-normal data and full model

The percentages of correct decisions when the data were obtained from a moderately skewed distribution and the full model was adopted are found in Table 5. These results can be summarized as follows:

1) The pattern of results found under departures from normality was qualitatively similar to the pattern described for the normal distribution; however, from a quantitative point of view the shape of the distribution substantially affected the percentage of correct decisions. Averaging across the covariance structures, total sample size, number of repeated measurements, and methods of estimation, the overall success was 35.2% for the AIC, 33.3% for the AICc, 32.7% for the BIC, 28.4% for the CAIC, and 39.5% for the HQIC.

Table 4 Percentages of correct identifications for normal data and full model (t = 4, 8)

n = 30 ($n_1 = n_2$ = 15)				n = 60 ($n_1 = n_2$ = 30)				n = 120 ($n_1 = n_2$ = 60)				
CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	
REML estimation (t = 4)												
AIC_1	57.8	35.2	18.7	14.5	67.1	51.6	31.9	24.2	66.9	70.0	65.9	42.5
AIC_2	56.5	34.1	15.3	13.8	61.2	47.0	27.4	22.0	64.9	65.0	62.3	41.2
$AICc_1$	63.7	33.5	11.2	9.1	69.1	53.7	25.5	20.9	69.9	71.7	60.4	41.6
$AICc_2$	74.0	12.5	5.4	0.4	80.0	55.2	11.7	5.4	87.1	77.2	65.5	34.8
$AICc_3$	61.6	26.8	12.9	10.6	62.8	48.5	26.4	19.6	67.1	66.5	61.8	40.1
$AICc_4$	67.0	10.9	3.1	0.4	68.6	44.1	10.3	3.9	71.9	70.1	59.9	31.4
$HQIC_1$	82.3	24.4	9.3	3.8	87.2	45.0	8.2	9.5	92.2	77.0	38.9	13.8
$HQIC_2$	70.8	30.2	12.5	7.9	81.4	51.1	17.7	12.6	88.7	80.1	50.6	25.3
$HQIC_3$	64.9	17.6	5.8	2.5	68.9	35.9	7.4	4.2	81.7	63.9	33.6	11.7
$HQIC_4$	62.6	27.4	8.2	6.9	69.9	42.9	15.6	10.9	81.3	69.9	44.7	21.1
BIC_1	87.4	9.2	0.3	0.1	93.2	20.7	0.7	0.2	97.3	51.5	7.6	1.5
BIC_2	84.1	20.9	3.4	2.3	90.8	45.8	3.6	2.4	96.8	77.7	19.1	8.6
BIC_3	51.6	5.5	0.2	0.0	56.1	11.7	0.3	0.1	69.8	40.7	4.8	1.6
BIC_4	62.8	16.0	1.5	1.8	67.7	24.2	2.4	1.1	78.4	68.0	15.5	6.7
$CAIC_1$	85.2	4.9	0.3	0.0	90.8	14.0	0.0	0.1	92.8	38.4	3.3	0.6
$CAIC_2$	85.6	13.0	0.9	0.2	92.6	30.8	1.2	0.4	95.6	66.8	10.2	2.5
$CAIC_3$	41.8	1.5	0.0	0.0	49.3	4.7	0.0	0.0	62.3	32.3	1.7	0.4
$CAIC_4$	54.0	9.0	0.4	0.0	59.3	13.7	0.5	0.2	72.3	57.4	6.6	1.7
ML estimation (t = 4)												
AIC	34.0	22.1	11.7	10.6	50.7	40.4	29.6	22.5	64.3	62.5	60.6	31.2
$AICc_1$	40.8	12.2	5.3	0.0	58.0	36.1	12.7	3.5	69.6	66.1	55.7	19.0
$AICc_2$	37.9	16.6	12.5	5.2	51.7	41.9	28.9	19.9	65.9	63.8	59.1	28.1
$HQIC_1$	37.9	13.3	4.2	2.3	57.3	30.0	7.3	5.1	78.4	61.2	30.6	9.7
$HQIC_2$	38.1	20.3	8.0	5.8	57.7	37.3	13.9	10.4	77.1	64.6	39.0	16.2
BIC_1	31.2	9.1	0.3	0.2	44.2	12.4	0.5	0.2	66.7	31.3	3.6	0.8
BIC_2	37.5	17.4	1.0	1.6	54.3	21.3	3.7	1.0	73.4	48.0	13.9	4.3
$CAIC_1$	26.7	6.3	0.2	0.0	35.9	7.9	0.1	0.1	58.9	21.7	1.4	0.2
$CAIC_2$	34.1	11.5	0.3	0.2	47.2	15.8	0.5	0.4	68.0	36.2	5.5	0.9
REML estimation (t = 8)												
AIC_1	69.5	71.8	74.2	92.8	74.8	83.9	95.8	99.0	77.3	87.7	99.4	99.8
AIC_2	60.8	56.9	61.0	64.1	64.7	69.8	76.4	66.1	73.4	82.7	92.8	88.1
$AICc_1$	72.6	81.6	68.7	86.2	77.0	87.2	95.0	99.2	78.7	88.8	99.8	99.8
$AICc_2$	0.0	0.2	0.0	15.0	82.7	94.6	73.7	0.0	84.3	97.1	99.8	99.9
$AICc_3$	63.7	64.5	54.6	52.4	67.9	72.1	72.5	64.2	75.5	83.6	92.4	86.3
$AICc_4$	0.0	0.2	0.0	5.4	72.4	66.6	38.5	0.0	80.4	89.0	88.5	72.9
$HQIC_1$	86.9	92.4	34.2	45.5	92.2	97.3	77.8	99.0	93.2	99.6	99.2	99.9
$HQIC_2$	79.6	85.3	63.0	90.4	88.6	96.8	91.3	99.7	91.7	99.0	99.8	99.9
$HQIC_3$	57.3	60.8	26.9	25.3	65.2	66.4	55.0	50.6	88.0	82.6	80.6	76.0
$HQIC_4$	64.8	63.9	48.6	54.0	66.5	70.8	61.6	57.3	82.0	87.8	86.4	81.0
BIC_1	91.6	75.8	0.0	0.2	98.2	95.2	22.4	25.5	99.5	99.8	84.1	97.2
BIC_2	86.8	92.6	34.8	44.8	94.8	96.6	69.4	97.6	96.3	99.9	98.0	99.8
BIC_3	38.4	31.3	0.0	0.0	45.4	41.0	9.1	10.6	66.7	61.8	52.9	55.2
BIC_4	57.1	60.4	26.8	25.0	60.9	65.5	40.0	45.0	77.4	78.5	72.0	70.1
$CAIC_1$	88.8	59.5	0.0	0.7	96.3	94.6	9.0	3.8	97.1	99.0	69.5	92.0

Table 4 (continued)

$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)				
CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	
$CAIC_2$	90.0	89.0	16.9	5.2	96.6	98.0	47.7	71.4	98.0	99.7	94.1	99.6
$CAIC_3$	31.2	21.4	0.0	0.2	37.4	32.6	2.8	1.8	59.3	54.4	37.5	47.2
$CAIC_4$	47.6	47.2	9.1	2.4	54.6	52.9	20.4	27.2	72.0	70.8	63.2	63.4
ML estimation ($t = 8$)												
AIC	40.5	43.9	45.6	64.8	57.1	60.3	69.0	75.8	70.7	80.6	95.7	91.6
$AICc_1$	0.0	0.0	0.0	0.0	68.7	57.0	36.5	0.0	78.8	87.1	82.9	88.8
$AICc_2$	46.2	49.4	39.3	52.1	64.4	62.1	67.7	72.8	74.3	81.8	95.3	91.4
$HQIC_1$	43.7	45.5	13.6	30.6	70.8	52.9	42.3	58.8	92.6	81.5	86.3	81.8
$HQIC_2$	48.6	49.2	35.9	58.1	71.6	60.0	58.4	67.6	91.1	85.6	90.4	86.3
BIC_1	37.3	21.5	1.2	0.2	59.4	29.5	8.1	13.2	93.6	60.7	66.8	52.1
BIC_2	53.6	45.6	13.7	30.1	69.8	48.0	33.9	52.1	97.4	74.9	80.3	74.4
$CAIC_1$	26.0	13.0	0.4	0.0	52.7	23.6	3.2	2.5	85.5	53.7	60.8	38.8
$CAIC_2$	38.0	34.1	3.9	11.0	63.2	37.9	19.3	36.0	94.6	67.1	76.0	64.5

CP_1 = first-order autoregressive covariance pattern; CP_2 = heterogeneous first-order autoregressive covariance pattern; CP_3 = heterogeneous Toeplitz covariance pattern; CP_4 = unstructured covariance pattern

Table 5 Percentages of correct identifications for non-normal data and full model ($t = 4, 8$)

	$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
REML estimation ($t = 4$)												
AIC_1	21.5	31.8	22.0	23.2	19.5	37.6	34.2	30.6	18.5	42.8	54.2	44.7
AIC_2	21.3	30.1	20.6	14.2	18.2	34.7	31.1	27.8	17.0	39.1	49.7	41.5
$AICc_1$	26.3	35.7	19.7	15.3	21.9	40.4	34.0	25.2	19.5	44.2	55.3	44.8
$AICc_2$	56.1	37.1	12.0	2.0	33.6	52.2	28.9	10.3	24.7	54.4	60.4	35.2
$AICc_3$	26.0	33.3	18.5	14.5	20.6	36.9	31.1	23.7	18.1	40.9	54.5	37.9
$AICc_4$	41.9	23.0	9.0	1.9	29.4	43.8	23.6	7.2	22.0	45.6	51.3	29.0
$HQIC_1$	45.9	33.9	10.7	6.3	47.1	52.3	32.2	10.7	49.1	67.4	52.2	16.6
$HQIC_2$	35.4	35.4	16.8	13.7	36.4	49.8	28.1	19.5	40.2	62.8	57.4	35.8
$HQIC_3$	35.2	27.3	9.0	4.9	36.3	40.7	17.7	8.8	40.1	52.3	42.3	13.9
$HQIC_4$	29.1	30.3	13.3	10.5	30.7	41.1	23.5	14.9	33.8	50.2	47.5	23.4
BIC_1	63.5	30.2	4.9	0.5	68.3	49.3	5.7	1.1	72.0	67.7	20.6	3.3
BIC_2	54.9	35.9	11.5	4.1	65.4	54.5	14.2	3.9	68.4	78.3	41.8	7.7
BIC_3	36.2	35.5	3.0	0.3	40.1	25.7	3.6	0.7	45.2	41.4	15.1	1.6
BIC_4	36.5	23.6	9.1	3.0	43.6	36.0	11.7	2.9	45.3	50.3	28.0	5.5
$CAIC_1$	70.0	23.8	1.2	0.2	75.4	38.8	1.9	0.5	76.2	62.1	15.7	0.9
$CAIC_2$	68.4	31.4	6.9	0.9	71.0	51.9	10.5	1.4	74.1	74.3	28.8	3.7
$CAIC_3$	31.6	10.3	0.9	0.1	38.5	18.4	1.8	0.5	41.5	31.9	9.7	0.5
$CAIC_4$	36.2	18.1	3.1	0.7	43.1	29.0	6.7	1.0	48.3	45.8	18.9	1.5
ML estimation ($t = 4$)												
AIC	6.6	19.4	10.8	14.4	10.3	29.4	28.6	23.6	12.4	39.3	42.3	29.1
$AICc_1$	12.1	14.3	2.9	0.2	24.9	34.1	20.8	7.8	23.9	45.4	44.5	27.8
$AICc_2$	7.9	10.9	9.4	8.0	15.1	31.9	29.5	20.5	18.1	39.9	42.9	25.7

Table 5 (continued)

	$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
$HQIC_1$	10.2	18.2	4.3	3.9	20.9	33.0	15.5	5.6	28.0	49.5	35.0	7.7
$HQIC_2$	8.3	19.8	6.9	8.5	16.6	34.3	26.0	16.8	23.6	40.0	39.9	18.4
BIC_1	10.3	9.7	0.9	0.4	20.7	20.4	2.2	0.7	25.7	30.1	10.8	0.5
BIC_2	16.2	16.4	3.7	2.6	26.2	28.3	10.8	2.1	32.2	47.2	20.4	5.8
$CAIC_1$	9.6	6.9	0.2	0.1	17.3	15.9	1.4	0.4	27.3	31.8	5.3	0.1
$CAIC_2$	13.4	11.3	1.5	0.9	24.6	24.8	4.9	1.9	31.6	44.5	12.2	1.9
REML estimation ($t = 8$)												
AIC_1	14.9	37.0	40.3	89.2	19.1	45.0	63.3	98.6	20.5	45.9	67.5	99.8
AIC_2	12.8	30.9	32.2	58.3	16.0	38.9	49.0	64.5	19.8	42.5	60.4	83.1
$AICc_1$	19.4	52.0	50.7	84.5	21.5	52.2	73.4	98.3	20.2	48.8	73.3	99.9
$AICc_2$	0.0	0.0	0.0	18.2	40.2	62.3	65.3	0.0	39.4	71.2	97.1	99.9
$AICc_3$	17.7	42.4	38.6	51.3	18.5	41.0	56.9	62.5	19.2	44.8	65.2	82.2
$AICc_4$	0.0	0.0	0.0	9.2	32.3	60.2	32.3	0.0	37.4	62.2	81.1	66.1
$HQIC_1$	50.8	75.8	29.5	53.9	58.1	88.4	67.3	97.6	53.6	90.3	96.0	99.9
$HQIC_2$	25.3	55.9	44.0	87.0	35.6	75.2	78.1	98.5	38.0	81.0	95.2	99.9
$HQIC_3$	31.0	50.6	17.2	27.6	36.0	55.1	37.2	50.1	41.6	69.5	70.7	70.2
$HQIC_4$	21.1	42.0	32.8	50.9	27.3	53.2	53.0	57.7	31.0	67.0	77.0	75.2
BIC_1	82.9	71.7	2.0	1.0	84.0	97.5	25.6	29.5	87.4	99.2	74.8	94.4
BIC_2	50.7	75.8	24.9	53.8	60.8	92.5	59.9	95.6	70.5	96.1	91.6	99.8
BIC_3	31.8	33.3	0.9	0.6	36.0	38.4	8.8	11.2	47.3	53.2	38.9	51.0
BIC_4	30.8	50.5	16.8	27.3	33.6	52.6	30.8	45.9	48.3	67.0	61.3	64.0
$CAIC_1$	87.6	71.0	1.0	0.0	89.6	96.5	14.7	7.6	90.2	97.6	60.8	88.2
$CAIC_2$	70.0	77.5	8.9	14.0	75.6	96.7	42.1	74.8	80.1	98.0	84.2	99.5
$CAIC_3$	27.8	23.3	0.5	0.0	32.6	29.4	4.6	02.9	42.3	46.5	27.6	44.8
$CAIC_4$	34.0	42.2	5.7	6.9	37.9	46.6	17.0	30.4	49.0	61.9	51.0	57.0
ML estimation ($t = 8$)												
AIC	6.5	21.8	27.0	62.5	8.9	33.6	44.5	74.0	10.9	39.8	59.7	92.4
$AICc_1$	0.0	0.0	0.0	0.0	21.5	52.1	31.0	0.0	26.4	60.5	81.1	77.6
$AICc_2$	10.4	31.4	33.1	57.3	12.4	38.6	51.1	70.5	20.4	41.9	66.2	91.6
$HQIC_1$	19.6	35.9	17.7	39.6	28.6	46.4	36.6	57.4	39.3	67.8	72.2	80.3
$HQIC_2$	12.8	32.6	30.1	61.5	22.0	47.1	50.0	65.1	34.0	64.9	76.1	85.2
BIC_1	20.6	23.8	1.5	1.9	27.8	31.9	11.3	15.2	43.9	51.5	36.7	61.2
BIC_2	19.5	35.7	17.8	39.6	29.4	43.8	31.1	53.3	46.6	65.2	62.5	75.5
$CAIC_1$	16.4	18.0	0.4	0.2	23.6	26.1	6.7	4.0	38.6	46.2	27.1	53.3
$CAIC_2$	22.5	36.6	6.2	11.1	30.7	38.4	22.9	37.2	47.2	58.3	50.6	68.9

See the note from Table 4

2) A comparison of the results for both normal and non-normal distributions shows that the average differences were approximately 19 and 9 percentage points for the efficient and consistent criteria, respectively. As a consequence, when the data were obtained from a moderately skewed distribution the differences between the two classes of criteria with respect to selecting the correct model among a set of competing models were small.
3) Regardless of the total sample size, length of the repeated measures, and method of estimation used, the performance of the consistent criteria was better under simple covariance patterns than under complex covariance patterns, and vice versa for the efficient criteria.
4) In general, the consistent criteria based on n were more effective when selecting the best model than the consistent criteria based on m, particularly for more complex covariance patterns. In contrast, the simula-

Table 6 Percentages of correct identifications for normal data and reduced model (t = 4, 8)

	n = 30 ($n_1 = n_2$ = 15)				n = 60 ($n_1 = n_2$ = 30)				n = 120 ($n_1 = n_2$ = 60)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
REML estimation (t = 4)												
AIC_1	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AIC_2	13.4	14.2	6.9	5.8	34.5	35.6	20.9	24.6	50.5	53.4	55.2	26.0
$AICc_1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$AICc_2$ 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
$AICc_3$	16.5	10.8	6.4	5.0	36.4	32.2	19.9	20.9	51.1	55.2	54.2	23.5
$AICc_4$	52.4	25.4	1.7	0.2	58.9	40.8	15.7	19.7	63.2	66.7	54.2	20.1
$HQIC_1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_3$	52.3	24.8	4.2	2.5	69.6	39.0	8.2	10.3	81.3	73.9	45.1	8.3
$HQIC_4$	34.6	22.5	6.5	7.6	65.9	42.2	17.7	15.8	79.9	73.3	48.3	10.9
BIC_1	0.0	0.0	0.0	0.0	0.2	0.1	0.3	0.3	47.7	28.1	3.4	0.4
BIC_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BIC_3	79.2	9.6	0.3	0.2	87.5	23.1	0.7	0.9	96.2	53.2	8.5	1.2
BIC_4	68.8	24.1	1.7	1.7	86.8	34.9	3.6	5.9	92.4	68.4	20.9	3.5
$CAIC_1$	12.0	0.0	0.0	0.0	58.1	8.1	0.0	0.2	74.5	33.3	1.5	0.3
$CAIC_2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.6	15.8	3.5	0.2
$CAIC_3$	84.9	5.9	0.2	0.1	92.8	17.4	0.0	0.4	97.8	42.2	3.7	0.7
$CAIC_4$	83.7	15.5	0.6	0.8	91.5	26.6	1.2	2.5	95.9	57.0	10.4	1.6
ML estimation (t = 4)												
AIC	47.0	26.7	16.7	13.8	51.6	43.3	30.2	20.7	55.9	57.9	53.7	31.6
$AICc_1$	76.5	29.1	9.2	0.9	66.6	48.2	19.2	7.3	62.6	66.6	51.9	24.7
$AICc_2$	54.4	27.2	15.8	10.0	55.4	44.9	29.7	18.3	58.1	60.1	53.8	30.2
$HQIC_1$	74.0	20.1	7.9	4.6	82.2	42.7	14.9	7.6	85.7	74.8	33.3	10.9
$HQIC_2$	59.6	25.3	14.4	8.8	73.3	45.6	22.3	10.5	79.8	73.7	41.5	16.3
BIC_1	86.4	9.2	0.4	0.3	95.5	19.4	0.9	0.6	97.2	55.2	8.3	1.8
BIC_2	78.8	18.6	2.7	1.1	88.0	34.5	7.3	2.2	92.7	69.7	20.9	9.1
$CAIC_1$	91.8	3.8	0.1	0.1	97.2	12.1	0.5	0.2	98.0	45.0	4.4	0.4
$CAIC_2$	86.6	11.1	1.3	0.7	94.2	23.6	3.3	1.4	96.5	60.5	12.7	2.3
REML estimation (t = 8)												
AIC_1	0.0	0.8	0.0	5.2	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.3
AIC_2	57.9	36.3	42.4	71.5	93.5	57.4	73.7	88.9	97.3	70.7	84.2	92.4
$AICc_1$	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$AICc_2$	0.0	0.5	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$AICc_3$	36.6	45.5	42.9	73.3	51.0	62.9	74.4	86.5	61.4	71.0	84.7	92.8
$AICc_4$	0.2	0.2	0.0	88.7	59.4	87.0	82.0	0.0	65.7	83.8	96.0	95.0
$HQIC_1$	0.0	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_2$	0.0	0.4	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_3$	76.5	76.8	27.0	43.6	81.5	88.8	71.0	94.1	88.4	93.2	94.	96.6
$HQIC_4$	53.0	58.7	43.7	75.0	70.6	80.3	78.8	92.8	82.0	88.9	92.0	95.7
BIC_1	0.0	0.9	0.0	1.9	0.2	0.0	0.0	16.0	0.0	0.0	1.8	70.6
BIC_2	0.0	0.6	0.0	3.8	0.2	0.0	0.0	0.6	0.0	0.0	0.0	0.2
BIC_3	89.4	74.4	1.8	7.0	94.1	97.6	24.5	29.4	98.0	99.3	84.8	99.0
BIC_4	76.8	76.5	26.8	43.6	85.9	91.9	63.6	93.3	93.6	97.0	94.6	98.5
$CAIC_1$	0.0	0.4	0.0	1.6	1.8	3.6	4.6	8.7	50.8	54.4	48.3	83.2

Table 6 (continued)

	$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
$CAIC_2$	0.0	0.0	0.0	2.8	0.2	0.0	0.0	5.3	0.0	0.0	0.0	29.9
$CAIC_3$	90.1	59.4	0.2	3.7	95.7	98.7	11.2	5.9	98.9	99.5	74.1	98.4
$CAIC_4$	85.8	79.1	8.5	9.0	90.6	95.0	44.9	74.2	96.4	98.6	92.8	99.0
ML estimation ($t = 8$)												
AIC	56.7	60.8	57.6	73.6	59.7	67.6	79.9	79.3	62.0	70.7	82.3	81.5
$AICc_1$	0.1	0.0	0.0	79.2	64.3	83.4	75.6	0.4	69.8	88.6	89.2	95.4
$AICc_2$	61.6	69.4	58.5	74.9	62.1	71.5	81.5	82.6	63.4	73.0	83.5	83.3
$HQIC_1$	84.9	86.5	27.0	49.0	87.8	92.7	74.4	90.0	89.1	92.4	93.6	93.5
$HQIC_2$	70.6	73.2	53.7	73.9	79.0	85.8	82.2	85.5	82.8	88.8	91.2	90.7
BIC_1	93.3	80.5	1.8	2.3	96.5	98.2	26.1	33.6	97.9	98.9	85.0	99.2
BIC_2	84.6	86.6	27.6	49.0	89.8	94.4	67.0	90.2	93.2	96.2	94.9	96.4
$CAIC_1$	92.2	66.1	0.1	1.2	98.2	98.6	12.2	7.3	98.6	99.5	74.0	98.8
$CAIC_2$	89.8	87.1	9.3	12.2	93.8	97.0	46.6	77.6	95.8	98.0	92.6	97.9

See the note from Table 4

Table 7 Percentages of correct identifications for non-normal data and reduced model ($t = 4, 8$)

	$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
REML estimation ($t = 4$)												
AIC_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AIC_2	2.2	8.5	4.5	5.2	10.7	20.5	20.3	19.9	12.5	29.9	37.5	23.4
$AICc_1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$AICc_2$	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0
$AICc_3$	4.7	8.5	5.9	5.1	11.6	22.0	21.2	17.2	13.1	30.1	38.7	22.5
$AICc_4$	17.8	30.9	3.0	0.3	30.3	37.8	21.1	15.5	37.2	38.8	44.0	22.3
$HQIC_1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$HQIC_3$	36.2	23.7	5.6	4.3	42.0	41.6	16.3	6.2	43.3	60.6	43.6	11.9
$HQIC_4$	19.3	14.2	7.1	6.4	30.9	36.2	23.5	9.8	37.9	52.7	46.2	16.0
BIC_1	0.0	0.0	0.0	0.0	0.7	1.2	0.7	0.1	25.7	26.1	11.4	1.1
BIC_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BIC_3	60.5	25.1	2.1	0.7	71.5	44.1	5.3	1.3	77.2	70.9	20.6	2.5
BIC_4	47.9	25.8	5.3	3.2	59.2	45.3	11.8	4.6	64.4	68.7	36.5	5.9
$CAIC_1$	8.3	4.7	0.3	0.0	43.0	22.8	2.2	0.5	58.5	47.4	11.1	0.8
$CAIC_2$	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.0	3.2	8.3	8.5	0.5
$CAIC_3$	71.3	21.9	0.8	0.1	83.3	40.7	3.9	0.8	83.8	65.8	14.2	1.1
$CAIC_4$	58.5	25.9	2.7	1.1	74.9	45.6	7.8	1.9	78.9	71.5	24.1	3.6
ML estimation ($t = 4$)												
AIC	17.2	21.3	11.3	19.3	17.0	35.4	25.4	22.6	20.3	34.9	39.8	29.1
$AICc_1$	44.1	34.2	3.4	0.4	27.8	50.8	22.2	10.2	29.0	40.6	46.6	23.0
$AICc_2$	21.6	27.5	10.5	13.5	19.5	38.0	25.7	20.3	23.0	35.1	42.3	27.3

Table 7 (continued)

	$n = 30$ ($n_1 = n_2 = 15$)				$n = 60$ ($n_1 = n_2 = 30$)				$n = 120$ ($n_1 = n_2 = 60$)			
	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4	CP_1	CP_2	CP_3	CP_4
$HQIC_1$	39.0	30.8	6.7	5.7	43.9	53.2	16.4	8.1	48.8	64.8	46.0	11.5
$HQIC_2$	25.2	27.6	10.0	13.0	32.2	49.6	21.1	12.8	41.5	56.7	49.4	16.7
BIC_1	63.6	25.0	2.5	0.5	72.0	49.4	4.9	1.5	78.0	70.2	22.1	2.9
BIC_2	44.5	30.6	5.5	3.5	53.7	52.5	11.5	4.7	67.8	71.7	35.5	6.5
$CAIC_1$	72.2	23.4	1.1	0.1	80.6	44.1	2.3	0.6	84.3	68.0	13.8	1.2
$CAIC_2$	59.8	26.5	3.2	1.6	68.9	51.6	6.2	2.9	76.5	71.5	24.9	3.7
REML estimation ($t = 8$)												
AIC_1	0.0	0.0	0.0	23.8	0.0	0.0	0.0	2.3	0.0	0.0	0.0	4.0
AIC_2	6.7	21.1	27.0	68.4	33.8	30.4	47.0	81.6	36.3	35.8	58.6	90.5
$AICc_1$	0.0	0.0	0.0	32.4	0.0	0.0	0.0	12.3	0.0	0.0	0.0	4.0
$AICc_2$	0.0	0.0	0.0	36.1	0.0	0.0	0.0	6.7	0.0	0.0	0.0	3.9
$AICc_3$	7.8	30.6	36.5	67.4	11.4	40.1	54.6	82.6	19.2	44.2	62.8	90.8
$AICc_4$	0.2	0.8	0.2	75.0	20.3	77.2	75.4	16.5	23.8	71.6	91.6	93.0
$HQIC_1$	0.0	0.0	0.0	19.6	0.0	0.0	0.0	11.7	0.0	0.0	0.0	3.7
$HQIC_2$	0.0	0.0	0.0	22.4	0.0	0.0	0.0	11.9	0.0	0.0	0.0	3.8
$HQIC_3$	39.6	62.9	30.3	53.8	46.7	78.5	64.5	86.6	49.0	75.5	90.1	95.0
$HQIC_4$	16.9	37.0	39.4	68.1	25.9	60.8	68.0	84.6	35.1	74.0	87.0	93.8
BIC_1	0.0	0.0	0.0	22.5	0.0	0.0	0.0	25.7	0.0	0.0	0.9	67.4
BIC_2	0.0	0.0	0.0	19.7	0.0	0.0	0.0	11.6	0.0	0.0	0.0	3.9
BIC_3	70.6	75.8	7.9	12.0	80.5	95.7	29.3	51.8	86.4	98.1	76.0	96.5
BIC_4	39.8	63.1	31.6	54.1	56.5	84.4	60.0	86.0	65.6	92.3	89.6	96.0
$CAIC_1$	0.7	0.4	0.1	9.3	0.8	5.8	4.2	25.1	37.9	47.8	38.0	79.9
$CAIC_2$	0.0	0.0	0.0	14.8	0.0	0.0	0.0	12.8	0.0	0.0	0.0	23.0
$CAIC_3$	78.7	73.3	4.7	9.4	87.0	96.6	16.7	37.6	91.6	99.0	65.6	85.0
$CAIC_4$	61.0	74.6	22.4	24.6	70.6	92.9	45.8	73.0	77.5	93.7	85.3	96.4
ML estimation ($t = 8$)												
AIC	10.7	30.9	31.5	65.2	8.7	35.5	51.4	73.9	12.4	36.9	54.8	79.1
$AICc_1$	0.1	1.1	0.0	86.3	27.2	61.2	69.0	27.0	38.6	83.6	87.3	94.0
$AICc_2$	16.6	40.4	43.9	68.4	10.5	41.3	62.2	79.6	12.5	44.2	60.2	81.2
$HQIC_1$	44.8	71.7	31.4	57.7	46.6	83.6	68.7	85.2	52.4	86.2	90.8	91.4
$HQIC_2$	22.6	47.5	39.0	67.1	29.8	63.7	71.8	81.0	33.9	74.0	87.0	88.9
BIC_1	81.3	80.8	6.6	11.5	83.4	96.4	32.2	46.6	87.6	98.2	79.1	96.4
BIC_2	44.6	72.9	31.4	57.2	54.2	89.1	62.9	85.8	66.1	93.7	89.8	94.3
$CAIC_1$	87.2	75.0	2.7	7.8	91.3	97.4	18.5	29.3	91.3	99.0	61.2	83.5
$CAIC_2$	65.9	80.6	18.5	31.3	72.1	95.0	49.0	77.1	80.2	96.8	85.4	95.6

See the note from Table 4

tion results pointed towards the use of m in the penalty terms of the AICc when the true covariance structures were TOEPH and UN and towards the use of n when the true covariance structures were AR and ARH, although overall differences were generally small in this case.

5) Finally, the five information criteria and their examined variations performed slightly better under REML than under ML estimation. The average differences between the two methods of estimation were approximately 8 percentage points for the efficient criteria and approximately 15 percentage points for the consistent criteria. It is also noteworthy that the ability of criteria to select the correct model was superior under $REML_1$ than under $REML_2$. The average differences between the two versions of REML estimation were approximately 8 percentage points for the efficient criteria and approximately 18 percentage points for the consistent criteria.

Performances associated with normal data and reduced model

The percentages of correct decisions when the data were normally distributed and the reduced model was adopted are found in Table 6. These results can be summarized as follows:

1) Contrary to what happened when the data were generated from a full model, it is very apparent that in this case $REML_1$-based information criteria are not appropriate for selecting the best model. A comparison of the results for both REML methods shows that the average differences were approximately 46 and 48 percentage points for the efficient and consistent criteria, respectively
2) The overall performance of the criteria under $REML_2$ estimation was roughly equivalent to the performance of the criteria under ML estimation. Specifically, the success rates averaged across covariance structures, number of subjects, and length of the repeated measures were 51.4 and 53.4% for the AIC, 46.8 and 50.7% for the AICc, 52.5 and 53.9% for the BIC, 48.9 and 49.8% for the CAIC, and 57.1 and 59.4% for the HQIC under $REML_2$ and ML, respectively. Results also show that the performance of the consistent criteria was better under simple covariance patterns than under complex covariance patterns, and vice versa for the efficient criteria.
3) Finally, results also show that the consistent criteria based on *n* performed better than the consistent criteria based on *m* when the true covariance structures were TOEPH and UN, and vice versa when the true covariance structures were AR and ARH. On the other hand, the simulation results pointed towards the use of *m* in the penalty terms of the AICc when the true covariance structures were TOEPH and UN and towards the use of *n* when the true covariance structures were AR and ARH.

Performances associated with non-normal data and reduced model

The percentages of correct decisions when the data were obtained from a moderately skewed distribution and the reduced model was adopted are found in Table 7. These results can be summarized as follows:

1) In general, the results obtained for the non-normal distributions were qualitatively similar to those found for the normal distributions. However, from a quantitative point of view the shape of the distribution substantially altered the performance of the efficient criteria and only slightly altered the performance of the consistent criteria, regardless of whether ML or REML estimation was used.
2) A comparison of the results for both normal and non-normal distributions shows that the average differences were approximately 17 and six percentage points for the efficient and consistent criteria, respectively. Thus, the consistent criteria performed better overall than their efficient counterparts. In fact, a careful examination of Table 7 shows that the success rates averaged across covariance structures, number of subjects, length of the repeated measures, and estimation methods (ML and $REML_2$), were 31.6% for the AIC, 34.8% for the AICc, 49.9% for the BIC, 49.3% for the CAIC, and 45.4% for the HQIC.
3) The performance of the five criteria was substantially better under $REML_2$ than under $REML_1$ estimation. As occurred for the reduced model and normally distributed data, the $REML_1$ function was not supported here. Nevertheless, it is very apparent that the pattern of correct identifications obtained under $REML_2$ closely resembled that obtained under ML. Specifically, the success rates averaged across covariance structures, total sample size, and length of the repeated measures were 30.8 and 31.6% for the AIC, 33.2 and 34.8% for the AICc, 49.2 and 49.9% for the BIC, 48.9 and 49.3% for the CAIC, and 44.8 and 45.4% for the HQIC under $REML_2$ and ML, respectively.
4) Finally, under this condition, it is also noteworthy that the consistent criteria based on *n* performed better than the consistent criteria based on *m* when the true covariance structures were TOEPH and UN, and vice versa when the true covariance structures were AR and ARH. In contrast, our results pointed towards the use of *m* in the penalty terms of the AICc when the true covariance structures were TOEPH and UN and towards the use of *n* when the true covariance structures were AR and ARH.

Discussion and recommendations

The purpose of this study was to examine the relative merits of several information criteria in selecting the best regression model when both the mean structure and the covariance structure are unknown. Unbalanced designs due to missing data involving both a moderate and large number of repeated measurements and varying total sample sizes were investigated. Also, the study investigated the impact of using different methods of estimation for information criteria, the impact of different adjustments for calculating the information criteria, and the impact of different distribution shapes. Results of this investigation are consistent with those obtained in past studies. More-

over, they contribute new findings that may aid researchers in the model selection process.

Simulation results showed that none of the procedures examined behaved correctly under all the conditions. Fortunately, despite differences in performance among the five criteria and their examined variations, performance of all criteria substantially improved with increased sample size and length of repeated measures. Averaging across the covariance structures, type of models, methods of estimation, length of repeated measures, and distribution shape the overall success rates for sample size configurations of (n_1, n_2) = (15, 15), (n_1, n_2) = (30, 30), and (n_1, n_2) = (60, 60) were as follows: AIC, 29, 40, and 48%; AICc, 21, 36, and 52%; BIC, 25, 36, and 52%; CAIC, 22, 32, and 51%; and HQIC, 30, 42, and 55%. Although no single criterion can be uniformly recommended, the AIC and HQIC had the best overall performance among the procedures examined. HQIC should in fact perform better in this study because selection is from a finite set of models (Burnham & Anderson, 2002).

From a quantitative standpoint, this pattern of results indicates that the percentages of times that the model generating the data was chosen by the information criteria were lower than those obtained in Gurka's (2006) similar study. However, Gurka (2006), in addition to considering only a normal and large complete-data sample size case, allowed a choice between only six candidate models, and in our study 36 candidate models were fit for each generated dataset. Consequently, poor performance for selecting the true model is not surprising given that several of the models used in this study may provide good approximations.

On the other hand, from a qualitative standpoint, our results partially corroborated findings by Gurka (2006), who reported that the criteria performed better or equally well under REML estimation compared to ML estimation when choosing the proper mean and covariance structure simultaneously. In fact, Gurka (2006) suggested using $REML_1$ estimation for the consistent criteria in selecting the proper model, and using $REML_2$ estimation for the efficient criteria. Note, however, that our results indicated that while $REML_2$-based information criteria performed comparably to the ML-based information criteria, performance of criteria was worse under $REML_1$ than ML. Thus, Gurka's recommendation was not unequivocally supported here.

A possible explanation for this discrepancy is that in Gurka's study datasets were simulated using the uniform correlation structure commonly assumed in designs where the within-subject factor is randomly allocated to subjects. According to this model, variances are constant across time as are the correlations between any pair of measurements. This assumption is insufficient and often unrealistic for describing real-time series data in the health and behavioral sciences (Fitzmaurice et al., 2004; Molenberghs & Kenward, 2007).

Beyond this, we also found that overall the consistent criteria (BIC, CAIC, and HQIC) performed better than their efficient counterparts (AIC and AICC) when the covariance patterns used to generate the data were relatively simple (i.e., AR and ARH). In contrast, the efficient criteria performed better than their consistent counterparts when the covariance patterns used to generate the data were more complex (i.e., TOEPH and UN). The consistent criteria tended to select a simpler model than the true model, particularly for the BIC and CAIC when the number of repeated measures was smaller. The opposite was true for the efficient criteria. Because excessively parsimonious structures are rarely adequate in real problems, selecting a model with few parameters is a type of error more severe than selecting a model with too many parameters. Hence, at least for the AIC and BIC, this result is consistent with the findings of Keselman, Algina et al. (1998) and Gomez et al. (2005).

With regard to discrepancies in the formulas involving in the penalty term of the criteria, our simulation studies showed that, regardless of whether ML or $REML_2$ estimation had been used, the consistent criteria based on total number of subjects (n) were more effective when selecting the best model than the consistent criteria based on total number of observations (m), particularly for the BIC and CAIC criteria. These results corroborate and generalize those found in Gurka (2006) study, while confirming the recommendation to use n in the penalty term of the BIC and CAIC criteria (see, Carlin & Louis, 2001; Kass & Raftery, 1995). In contrast, the simulation results pointed towards the use of m in the penalty terms of the AICc, as used by *Proc Mixed* in SAS and *mixed* command in SPSS/PASW. As indicated above, sample size in SAS is equal to n whereas sample size in SPSS/PASW is equal to m under ML and REML, respectively, when computing the BIC and CAIC.

The simulation studies covered in this paper also revealed that the criteria exhibited a clear superiority in their ability to accurately select the correct model when data were obtained from normal distributions. However, none of the procedures considered performed well when data were obtained from moderately skewed distributions, regardless of whether ML or REML estimation was used. For this scenario, consistent criteria and, even more so, efficient criteria, picked the wrong model more frequently than the correct model. This finding serves to reinforce the importance of testing for evidence of skewness in the data (see Vallejo, Ato, & Fernández, 2010, for details).

To conclude, we would like to add four brief comments. First and foremost, it is very important to emphasize that the error rate for all criteria decreased as the sample size

increased. Therefore, 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. In order to reach an acceptable power to distinguish between competing models, the rule of thumb $n \geq 10t$ per group is suggested. Second, with small samples the standard selection criteria may be highly inefficient, particularly if some data are missing and/or the form of the matrix plays an important role in the estimation. In these cases, it may be more appropriate to select the best mean model assuming the unstructured pattern to model the within-subject errors in the analysis (Siddiqui, Hung, & O'Neill, 2009). Third, it should be noted that information criteria do not automatically select the best model from all possible candidate models. Thus, the joint modeling of mean and covariance structures can be impractical if the number of explanatory variables is large. Fourth, the results are of course limited to the conditions examined in our study, though we sense that they may be generalizable to a considerably wider range of conditions that could conceivably be obtained in behavioral science research. We conclude by noting that this study did not address the effects of heterogeneity of covariance across groups on the performance of the criteria in selecting the best regression model. Based on this research, it can be expected that between-groups heterogeneity will substantially affect the consistent criteria's performance. In future research, it would be informative to examine the performance of the linear model, using techniques (e.g., generalized linear mixed models) that allow for non-normal error term distributions and relax the requirement of constant variability. Currently, SAS *Proc Glimmix* allows users to fit statistical models to data when assumptions of normality and variance homogeneity are not necessarily satisfied.

Author Note This work was supported by a Grant PSI2008-03624 from the Spanish Ministry of Science and Innovation. We gratefully thank the anonymous reviewers and the associate editor for their helpful comments and very constructive suggestions.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, AC-19*, 716–723. doi:10.1109/TAC.1974.1100705

Algina, J., & Keselman, H. J. (2004). A comparison of methods for longitudinal analysis with missing data. *Journal of Modern Applied Statistical Methods, 3*, 13–26.

Azari, R., Li, L., & Tsai, C. L. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis, 50*, 3053–3066. doi:10.1016/j.csda.2005.05.009

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370. doi:10.1007/BF02294361

Bozdogan, H. (2000). Akaike's information criterion and recent developments information complexity. *Journal of Mathematical Psychology, 44*, 62–91. doi:10.1006/jmps.1999.1277

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approach*. New York, NY: Springer.

Carlin, B. P., & Louis, T. A. (2001). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). London: Chapman & Hall / CRC Press.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. New York, NY: Cambridge University Press.

Cooper, D. M. K., & Thompson, R. (1977). A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika, 64*, 625–628. doi:10.1093/biomet/64.3.625

DeSouza, C. M., Legedza, A. T., & Sankoh, A. J. (2009). An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics, 6*, 1055–1073. doi:10.1080/10543400903242795

Feng, R., Zhou, G., Zhang, M., & Zhang, H. (2009). Analysis of twin data using SAS. *Biometrics, 65*, 584–589. doi:10.1111/j.1541-0420.2008.01098.x

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

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521–532. doi:10.1007/BF02293811

Gomez, V. E., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation, 34*, 377–392. doi:10.1081/SAC-200055719

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *American Statistician, 60*, 19–26. doi:10.1198/000313006X90396

Gurka, M. J., & Edwards, L. J. (2008). Mixed models. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Handbook of statistics, vol 27, epidemiological and medical statistics* (pp. 253–280). New York, NY: Elsevier.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B, 41*, 190–195.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika, 61*, 383–385. doi:10.1093/biomet/61.2.383

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307. doi:10.1093/biomet/76.2.297

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated–measures models with structured covariance matrices. *Biometrics, 42*, 805–820.

Jiang, J., & Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhya, 65*, 23–42.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics-Simulation and Computation, 27*, 591–604. doi:10.1080/03610919808813497

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B.,Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R.(1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA,

and ANCOVA analyses. *Review of Educational Research, 68*, 350–386. doi:10.3102/00346543068003350

Kitagawa, G., & Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Annals of the Institute of Statistical Mathematics, 62*, 209–234. doi:10.1007/s10463-009-0237-1

Kleinman, K., & Horton, N. J. (2010). *SAS and R: Data management, statistical analysis, and graphics*. London: Chapman & Hall/CRC.

Kowalchuk, R. K., Lix, L. M., & Keselman, H. J. (1996). *The analysis of repeated measures designs*. Paper presented at the annual meeting of the Psychometric Society, 1996, Banff, Alberta, Canada

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

Lee, H., & Ghosh, S. K. (2009). Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation, 79*, 93–106. doi:10.1080/00949650701611143

Liang, H., Wu, H., & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika, 95*, 773–778. doi:10.1093/biomet/asn023

Lindstrom, M. J., & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*, 1014–1022.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS System for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793–1819. doi:10.1002/1097-0258(20000715

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalize linear, and mixed models* (2nd ed.). Hoboken, NJ: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166. doi:10.1037/0033-2909.105.1.156

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. New York, NY: Wiley.

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling, 1*, 235–269. doi:10.1177/1471082X0100100402

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*, 545–554. doi:10.1093/biomet/58.3.545

Ripley, B. E. (1987). *Stochastic simulation*. New York, NY: Wiley.

SAS Institute Inc. (2008). *SAS/STAT ® 9.2 user's guide.* Cary, NC: Author.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. doi:10.1037/1082-989X.7.2.147

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

Shang, J., & Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis, 52*, 2004–2021. doi:10.1016/j.csda.2007.06.019

Siddiqui, O., Hung, H. M. J., & O'Neill, R. (2009). MMRM vs. LOCF: A comprehensive comparison based on simulation study and 25 NDA datasets. *Journal of Biopharmaceutical Statistics, 19*, 227–246. doi:10.1080/10543400802609797

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Sugiura, N. (1978). Further analysis of de data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods, 7*, 13–26. doi:10.1080/03610927808827599

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92*, 351–370. doi:10.1093/biomet/92.2.351

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika, 48*, 465–471. doi:10.1007/BF02293687

Vallejo, G., Ato, M., & Fernández, M. P. (2010). A robust approach for analyzing unbalanced factorial designs with fixed levels. *Behavior Research Methods, 42*, 607–617. doi:10.3758/BRM.42.2.607

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*, 10–21. doi:10.1027/1016-9040.12.1.10

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Wang, J., & Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics - Simulation and Computation, 38*, 788–801. doi:10.1080/03610910802645362

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. London: Chapman & Hall/CRC.

7.1.3. Objetivo 3

Los datos caracterizados por una estructura agrupada jerárquicamente, en la cual las unidades de observación de un determinado nivel están anidadas en un nivel más alto, son muy comunes en el campo de la investigación educativa. En este sentido, cuando se trabaja con estos datos hay que tener suma cautela en su análisis, dado que las consecuencias de no prestar atención a la estructura jerárquica de los datos conlleva problemas gravísimos (Vallejo et al., 2008) de unidades de análisis y de sesgos de agregación. Estas y otras complicaciones no pueden ser resueltas utilizando técnicas estadísticas basadas en el clásico modelo lineal general. Para solventarlo se requieren técnicas analíticas más sofisticadas como los Modelos Multinivel erigidos en el tiempo como la estrategia metodológica más adecuada para estos datos. Así, estos modelos reconocen la estructura anidada de los datos y permiten estimar las variaciones que se producen en los distintos estratos producidos por el agrupamiento, tanto en los estudios de carácter longitudinal (como se ha podido apreciar en los artículos anteriormente presentados) como en los de carácter transversal.

Para realizar un análisis de los datos jerárquicos adecuado, el investigador se ve en el deber de elegir un conjunto de modelos candidatos, una técnica de modelización estadística apropiada y una herramienta óptima para encontrar el mejor modelo que proporcione la aproximación lo más cercana posible al verdadero modelo desconocido de entre las alternativas que compiten. Estudios recientes (Gurka, 2006; Vallejo et al., 2010, 2011b) han evaluado ampliamente la selección de los modelos mediante criterios de verosimilitud en el contexto longitudinal (tanto con modelos de medidas repetidas anidados como no anidados). Sin embargo, el uso de los criterios de selección en la modelización multinivel continúa siendo un tema de debate en curso.

Teniendo esto presente, se parte ahora de un estudio transversal para llevar a cabo el **Tercer Objetivo** de esta Tesis. Así, la investigación que a continuación se expone fue diseñada para encontrar la mejor estrategia de selección del modelo correcto multinivel entre varias alternativas posibles. Primeramente se examina esta cuestión mediante un estudio de simulación Monte Carlo, tanto desde un punto de vista más clásico como desde un punto de vista bayesiano. Para los propósitos de comparación son utilizados los Criterios de Información AIC, AICC, BIC, CAIC, HQIC y son introducidos dos más, el DIC (Criterio de Información de la Desvianza) y el cAIC (AIC Condicional). En segundo lugar, y mediante un conjunto de datos previamente

publicados en un estudio empírico, se analiza el comportamiento de los criterios para explorar la posibilidad de generalización de los resultados.

Paper III

International Journal of Clinical and Health Psychology
ISSN 1697-2600

Asociación Española de Psicología Conductual (AEPC)

Director: Juan Carlos Sierra
Directores asociados: Stephen N. Haynes, Michael W. Eysenck y Gualberto Buela-Casal

Juan Carlos Sierra, director de **International Journal of Clinical and Health Psychology**, informa QUE:

El artículo

"**Performance evaluation of recent information criteria for selecting multilevel models in behavioral and social sciences**" (Guillermo Vallejo, Ellián Tuero-Herrero, José Carlos Núñez y Pedro Rosário) [Ref. 2013037).

ha sido aceptado para su publicación en el volumen 14 (número 1, enero 2014) de dicha revista.

Firmado en Granada, a 17 de junio de 2013.

Juan Carlos Sierra
Director *Int J Clin Health Psychol*

International Journal of Clinical and Health Psychology está incluida en Journal Citation Index-Social Sciences, Social Sciences Citation Index, Current Contents/Social Behavioral Sciences, Journal Citation Reports, PsycINFO, Scopus, National Library of Medicine, IN-RECS (Psicología), ISOC, PSICODOC, Red AlyC y Latindex.

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

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

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.

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

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. 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.

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

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:

$$BA_{ij} = b_{0j} + b_{1j} LA_{ij} + b_2 PD_{ij} + b_3 CA_{ij} + e_{ij},$$

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_{ij}, have a normal distribution with mean zero and variance σ_e^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

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

which illustrates that the *BA* may be viewed as a function of the overall intercept (γ_{00}), the main effect of teacher's $TA(\gamma_{01})$, the main effect of teacher's $TE(\gamma_{02})$, the main effect of student's $LA(\gamma_{10})$, the main effect of student's $PD(\gamma_{20})$, the main effect of student's $CA(\gamma_{30})$, and cross-level interactions involving TE with $LA(\gamma_{12})$ and TA with $LA(\gamma_{11})$, plus a random error: $u_{0j} + u_{1j} LA_{ij} + e_{ij}$. The variable e_{ij} 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 $LA_{ij} \times TA_j$ from the model;

(M_3) by dropping $LA_{ij} \times TE_j$ from the model; (M_4) by dropping u_{1j} 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 $T\mathrm{A}_j$ and TE_j; (M_9) by dropping PD_{ij} and including an interaction between LA_{ij} and CA_{ij}; (M_{10}) by dropping $LA_{ij} \times TA_j$ 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.

Table 1 Formulas for commonly used information criteria.

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

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.

Conditional Akaike's Information Criteria (*c-AIC*)

The conditional AIC is similar in form to the marginal AIC; however, these focuses have different likelihood functions and a different number of parameters. The c-AIC in "smaller-is-better" form is defined as

$$\text{cAIC} = \hat{d} + 2s_{\text{c}},$$

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 (DIC)

The DIC is a generalization of AIC (see 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:

$$\text{DIC} = \text{D}(\bar{\boldsymbol{\theta}}) + 2p_{\text{D}},$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \mathbf{u}', \sigma)'$, $\text{D}(\bar{\boldsymbol{\theta}}) = -2\log L(\mathbf{y} \mid \bar{\boldsymbol{\theta}})$ is the deviance of the model evaluated at the means of the posterior distributions of the parameters, and $p_D = \overline{\text{D}(\boldsymbol{\theta})} - \text{D}(\bar{\boldsymbol{\theta}})$ is the effective number of parameters. SAS Version 9.3 (SAS Institute, 2011) PROC MCMC calculates DIC taking $\overline{\text{D}(\boldsymbol{\theta})}$ to be the posterior mean of $-2\log L(\mathbf{y} \mid \boldsymbol{\theta})$, and evaluating $\text{D}(\bar{\boldsymbol{\theta}})$ 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 ($N = m \times n$) 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:

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.

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

	AIC *(SPSS/SAS)*	*c-AIC* *(SAS+R)*	*AICC* *(SPSS/SAS)*	*BIC_N* *(SPSS)*	*BIC_m* *(SAS)*	*$CAIC_N$* *(SPSS)*	*$CAIC_m$* *(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. PM = Parameter magnitude; ICC = Intraclass correlation.

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.

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

	AIC	*cAIC*	*AICC*	BIC_N	BIC_m	$CAIC_N$	$CAIC_m$	*HQC*	*DIC*
$PM = 0.5 / ICC = .1 / \rho_{u_{01}} = .2$									
	26/32	55/53	26/32	03/05	11/15	02/03	07/10	19/24	20
$PM = 1.0 / ICC = .1 / \rho_{u_{01}} = .2$									
	36/35	57/58	35/35	06/06	20/18	04/04	13/13	28/29	30
$PM = 0.5 / ICC = .3 / \rho_{u_{01}} = .2$									
	25/36	61/57	24/36	03/07	09/19	03/05	06/14	17/28	22
$PM = 1.0 / ICC = .3 / \rho_{u_{01}} = .2$									
	39/42	71/73	38/41	08/08	21/21	05/06	15/17	32/33	29
$PM = 0.5 / ICC = .1 / \rho_{u_{01}} = .4$									
	53/63	67/63	53/63	14/27	33/47	10/21	25/40	45/57	42
$PM = 1.0 / ICC = .1 / \rho_{u_{01}} = .4$									
	70/70	73/71	69/70	31/29	54/51	25/23	45/42	64/63	62
$PM = 0.5 / ICC = .3 / \rho_{u_{01}} = .4$									
	49/70	77/70	48/70	10/36	27/57	07/31	19/49	39/66	46

$PM = 1.0 / ICC = .3 / \rho_{u_{01}} = .4$									
	75/80	87/86	74/79	36/43	59/64	29/36	50/57	69/74	63

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

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 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.

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

	Criterion								
Model	*AIC*	*c-AIC*	*AICC*	BIC_N	BIC_m	$CAIC_N$	$CAIC_m$	*HQC*	*DIC*
M_1	**5009.7/**	**5090.5/**	**5009.8/**	5069.9/	**5032.2/**	5080.9/	5043.2/	**5018.4/**	
	5002.4	**5103.6**	**5002.4**	**5018.9**	**5008.6**	**5021.9**	**5011.6**	**5004.8**	4976.0
M_2	5014.9/	5099.3/	5015.0/	5069.7/	5035.3/	5079.7/	5045.3/	5022.8/	
	5010.3	5112.9	5010.3	5026.7	5016.4	5029.7	5019.4	5012.7	4980.1
M_3	5012.8/	5092.1/	5012.9/	5067.5/	5033.2/	5077.5/	5043.2/	5020.7/	
	5007.9	5104.5	5008.0	5024.4	5014.1	5027.4	5017.1	5010.3	4977.6
M_4	5011.9/	5098.4/	5012.0/	**5066.6/**	5032.3/	**5076.6/**	**5042.3/**	5019.8/	
	5018.2	5113.3	5018.2	5029.1	5022.2	5031.1	5024.2	5019.8	4988.8
M_5	5011.3/	5092.0/	5011.4/	5077.0/	5035.8/	5089.0/	5047.8/	5020.8/	
	5007.2	5110.4	5007.2	5023.6	5013.3	5026.6	5016.3	5009.6	4979.0
M_6	5011.7/	5092.3/	5011.9/	5077.4/	5036.2/	5089.4/	5048.2/	5021.2/	
	5003.0	5106.5	5003.0	5019.4	5009.1	522.4	5012.1	5005.4	4979.1
M_7	5010.6/	5101.2/	5010.8/	5076.3/	5035.1/	5088.3/	5047.1/	5020.1/	
	5002.7	5114.8	5002.7	5024.6	5010.9	5028.6	5014.9	5005.9	**4975.8**

M_8	5010.4/	5092.5/	5010.5/	5070.6/	5032.8/	5081.6/	5043.8/	5019.1/	
	5006.6	5107.7	5006.6	5023.0	5012.7	5026.0	5015.7	5009.0	4977.0
M_9	5011.5/	5092.2/	5011.7/	5077.2/	5036.0/	5089.2/	5048.0/	5021.0/	
	5006.5	5109.5	5006.5	5022.9	5012.6	5025.9	5015.6	5008.9	4978.0
M_{10}	5012.8/	5108.5/	5012.9/	5073.0/	5035.2/	5084.0/	5046.2/	5021.5/	
	5007.4	5122.2	5007.5	5029.3	5015.6	5033.3	5019.6	5010.6	4976.8

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

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 (see 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.

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_7	
Fixed-effects	*Estimate*(*SE*)	*Pr*>|*t*|	*Estimate*(*SE*)	*Pr*>|*t*|	*Estimate*(*SE*)	*Pr*>|*t*|
γ_{00}(Intercept)	10.553(.477)	<.0001	10.519(.551)	<.0001	10.568(.567)	.0001
γ_1 (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_3 (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}(LA×TA)	-1.605(.590)	.0067			-1.671(.640)	.0117
γ_{12}(LA×TE)	1.256(.553)	.0234			1.256(.600)	.0368
Random-Effects	*Estimate*(*SE*)	*Pr*>*Z*	*Estimate*(*SE*)	*Pr*>*Z*	*Estimate*(*SE*)	*Pr*>*Z*
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_7	
Parameter [a]	*Mean*	*Posterior interval*	*Mean*	*Posterior interval*	*Mean*	*Posterior interval*
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]

a. Based on assuming uninformative priors.

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 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 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: 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.

7.1.4. Objetivo 4

El comienzo de esta Tesis arrancaba con una exposición de la investigación de la Eficacia Escolar, sus etapas, los múltiples estudios a los que dio lugar y, uno de los aspectos más importantes, las técnicas de análisis de los datos de estas investigaciones. En estas últimas nos hemos detenido un poco más dada su relevancia para llegar a conclusiones más precisas, válidas y fiables. Así las cosas, ha quedado claro que durante mucho tiempo los investigadores han pasado por alto las agrupaciones naturales reflejadas en los datos. La mayor parte de los datos registrados eran analizados mediante técnicas estadísticas basadas en el convencional modelo lineal general. Sin embargo, estas técnicas fueron mejoradas y superadas a partir de los años ochenta por los Modelos Lineales Multinivel (Goldstein, 1995). Debido entonces a la importancia de estos modelos para el análisis de los datos en el campo educativo, se realizaron 3 estudios mediante esta metodología con datos obtenidos mediante simulación. Sin embargo nos parecía apropiado realizar un cuarto estudio de carácter más aplicado, esto es, con datos reales para, a la postre, analizarlos desde la perspectiva multinivel. Y así lo hicimos.

Hasta la fecha, los datos aportados por diversas investigaciones no son concluyentes respecto al papel de las variables del estudiante y del contexto revisadas sobre el rendimiento académico de estudiantes preuniversitarios. Además, no se dispone de información sobre la relevancia de cada una de las variables tomadas en la determinación del rendimiento cuando se consideran conjuntamente, ni tampoco hay estudios que analicen estas variables considerando los resultados a nivel del estudiante ni al nivel de la clase. Así las cosas, la investigación que cierra esta Tesis es una investigación educativa realizada mediante un estudio ex pos-facto descriptivo analizado con modelos multinivel. Por ello, el **Cuarto Objetivo** de esta Tesis consiste en analizar el grado de asociación del rendimiento académico de los estudiantes en una asignatura concreta, Biología, con ciertas variables del estudiante (enfoques de aprendizaje, conocimientos previos, tiempos de estudio, grado de asistencia a clase y realización de deberes escolares); variables del profesor (años de experiencia, enfoques de enseñanza, género de los profesores); tamaño de la clase y nivel de estudios de los padres.

Paper IV

Revista de Psicodidáctica, 2014, 19(1), XX-XX
www.ehu.es/revista-psicodidactica

ISSN:1136-1034 eISSN: 2254-4372
UPV-EHU
DOI: 10.1387/RevPsicodidact.7127

Variables del estudiante, del profesor y del contexto en la predicción del rendimiento académico en Biología: Análisis desde una perspectiva multinivel

José C. Núñez*, Guillermo Vallejo*, Pedro Rosário**, Ellián Tuero*, y Antonio Valle***

*Departamento de Psicología, Universidad de Oviedo, España; **Escola de Psicologia, Universidade do Minho, Portugal; *** Departamento de Psicología Evolutiva y de la Educación, Universidad de A Coruña, España.

Resumen

En el presente estudio se analiza la contribución de variables del alumno y variables del contexto en la predicción del rendimiento académico en Bachillerato. Se han obtenido información de 988 estudiantes, de último curso de Bachillerato y de sus 57 profesores de Biología. Los datos fueron analizados desde una perspectiva multinivel. Los resultados indican que, de la variabilidad observada en el rendimiento en Biología, el 85.6% se debe a las variables de nivel de estudiante mientras que el 14.4% restante corresponde a las variables de nivel de clase. A nivel de estudiante, el rendimiento en Biología se encontró asociado con el enfoque de aprendizaje, con los conocimientos previos, con el absentismo escolar y con el nivel educativo de los padres. A nivel de clase, el rendimiento únicamente estuvo asociado con el enfoque de enseñanza del profesor, y no directamente, sino a través del enfoque de estudio del alumno.

Palabras clave: Enfoques de enseñanza, enfoques de aprendizaje, rendimiento en Biología, análisis multinivel.

Abstract

The current investigation analyzed how student variables and context variables predicted high school students' academic achievement. The participants were 988 twelfth graders and their corresponding 57 Biology teachers. Data were analyzed making use of the multilevel method. Results indicate that 85.6% of the variation observed in the biology achievement was explained by variables at student level, while the remaining 14.4% was explained by variables at class level. At student level, biology achievement was associated with approaches to learning, prior knowledge, class absence and parents education level. At class level, the academic achievement was only associated with teachers' approaches to teaching not directly, but through students' approaches to learning.

Keywords: Approaches to teaching, approaches to learning, Biology achievement, multilevel analysis.

Agradecimientos: Este trabajo ha sido realizado con financiación del Ministerio de Ciencia e Innovación de España (Proyectos: EDU2010-16231 y PSI-2011-23395/PSIC).

Correspondencia: José Carlos Núñez, Departamento de Psicología de la Universidad de Oviedo. Plaza de Feijoo, s/n. 33003 Oviedo. España. E-mail: jcarlosn@uniovi.es

Introducción

En línea con los resultados aportados en los informes PISA de 2003 y de 2006, en 2009 el alumnado de Portugal y España volvió a presentar resultados en Ciencias (493 y 488 respectivamente) por debajo de la media de la OCDE (501), lo que sugiere la necesidad de investigar qué puede explicar estos resultados. Analizando el impacto de las macro-estructuras sociales y centrando el debate en las cuestiones del proceso de enseñanza y aprendizaje, el mismo informe de la OCDE (2010) afirma que las variables económicas del país (en concreto, el producto interior bruto) solamente explica un 6% de las diferencias de rendimiento encontradas en los distintos sistemas educativos. Este resultado constituye un reto para que investigar las variables que explican el 94% de varianza que resta por explicar en el rendimiento académico del alumnado de Bachillerato. En la presente investigación se pretende aumentar la comprensión sobre qué condiciones se encuentran determinando el rendimiento académico en el Bachillerato. Se intentará responder a este reto analizando la contribución de algunas de las variables del alumnado teóricamente más relevantes (p. e., los enfoques de aprendizaje, el rendimiento previo, el tiempo de estudio, la asistencia a clase, la realización de deberes escolares), así como también algunas variables del contexto (p. e., los enfoques de enseñanza, el género del profesor, la experiencia docente, el número de alumnos por clase, el nivel educativo de los padres). En este estudio, dado que los datos están organizados en una estructura jerárquica (el alumnado está anidado en clases con su respectivo profesor), se utiliza una estrategia de análisis multinivel que posibilita examinar tanto los efectos intra-clase como inter-clases.

Variables del alumnado y rendimiento académico

Los enfoques de aprendizaje

Marton y Säljö (1976), hace ya tres décadas, describieron dos formas diferentes que el alumnado tenía de enfocar el trabajo de un texto académico. Este estudio constituyó el inicio de una importante línea de investigación centrada en el estudio de lo que se denominó *enfoques de aprendizaje* del alumnado (Entwistle, 2009). Estos autores han identificado un nivel de procesamiento profundo y uno superficial de acuerdo con el enfoque de aprendizaje que el alumno utilizaba para acercarse a la tarea. El alumnado que utiliza preferencialmente un enfoque superficial está movido por un objetivo que es extrínseco a la tarea de aprendizaje; su implicación en la tarea es baja y su esfuerzo es ajustado a la mínima exigencia. Por contra, el alumnado que utiliza preferencialmente un enfoque profundo está motivado por la intención de maximizar la comprensión y construcción de significados al relacionar la tarea con sus conocimientos previos (Entwistle, 2009; Rosário et al., 2010; Rosário, Núñez, Valle, Paiva, y Polydoro, 2013).

El conocimiento previo

El alumnado organiza los conocimientos jerárquicamente, lo cual le permite comprender las nuevas experiencias. Por este motivo, lagunas graves en los conocimientos previos en un dominio pueden comprometer seriamente la adquisición de nuevos conocimientos (Alexander, Kulikowich, y Schulze, 1994; Miñano y Castejón, 2011). En consecuencia, el nivel de conocimiento previo parece ser una variable de interés a incluir en este estudio.

El tiempo de estudio

El tiempo de estudio, en general, es considerado un buen predictor del rendimiento escolar (Plant, Ericsson, Hill, y Asberg, 2005). No obstante, para que esto sea así, el tiempo de estudio y la correspondiente implicación del alumno tienen que ser constantemente ajustados en función tanto de los objetivos del alumnado como de la naturaleza de las tareas demandadas (p.e., el grado de dificultad, la utilidad percibida), y de las variables del contexto (p.e., el nivel de ruido, la temperatura). Quizás, por esta razón los datos de la literatura no apoyan inequívocamente la relación directa entre el tiempo de estudio y el rendimiento escolar (Gortner-Lahmers y Zulauf, 2000; Núñez, Rosário, Vallejo, y González-Pienda, 2013).

Los deberes escolares

Pese a la larga historia de investigación sobre el papel de los deberes escolares, aún falta por determinar claramente la fuerza de la relación entre la prescripción de éstos y el rendimiento académico (Dettmers, Trautwein, y Ludtke, 2009; Rosário et al., 2009; Trautwein y Köller, 2003). Mientras que en algunos estudios se encontró una relación positiva (p. e., Cooper, Robinson, y Pattal, 2006; Paschal et al., 1984), en otros las conclusiones son menos optimistas, indicando que esta relación es muy débil y que está mediada por variables personales, escolares y familiares (Ronning, 2011).

La asistencia a clase

Finalmente, en este estudio también se consideró importante incluir la variable absentismo escolar, pues ha despertado gran interés entre los investigadores (Jonanssen, 2011; McIntyre-Bhatty, 2008) por su asociación con el bajo rendimiento del alumnado (Reid, 2006). Estudiar esta variable en conexión con otras variables personales o del contexto de aprendizaje, por ejemplo las incluidas en esta investigación, podrá aportar algunas pistas para mejorar el proceso de enseñanza y de aprendizaje.

Variables del contexto y rendimiento académico

Los enfoques de enseñanza

Prosser y Trigwell (p. e., Prosser, Trigwell, y Taylor, 1994) desarrollaron una línea de investigación sobre cómo los profesores enseñaban en el contexto de la educación superior. Considerando los resultados derivados de sus investigaciones, se identificaron dos formas diferentes de afrontar el proceso instruccional, a las que se denominó *enfoques de enseñanza:* el enfoque orientado a la transmisión de información, centrado en el profesor (Information Transmission/Teacher-Focused (ITTF) approach), y el enfoque orientado al cambio conceptual, centrado en el alumno (Conceptual Change/Student-Focused (CCSF) approach). Mientras que los profesores que adoptan preferencialmente un enfoque ITTF centran su actividad en la transmisión de información relacionada con los contenidos de aprendizaje y las cuestiones técnicas relativas al proceso de enseñanza, los que utilizan en su proceso de enseñanza preferentemente un enfoque CCSF están comprometidos con promover la implicación del alumnado en un proceso activo de construcción de significados. En este sentido, estos profesores tienen en consideración los conocimientos previos del alumnado y

desarrollan estrategias de enseñanza tendentes a ayudar a la construcción del conocimiento (Ramsden, Prosser, Trigwell, y Martin, 2007). La investigación sobre los enfoques de enseñanza fue orientada hacia el análisis de su relación tanto con variables del contexto, p. e., el tamaño de la clase (Lopes y Santos, 2013; Rosário et al., 2013; Singer, 1996; Stes, Gijbels, y Van Petegem, 2008), como con variables personales del profesor: la experiencia docente de los profesores (Prosser, Ramsden, Trigwell, y Martin, 2003; Rosário et al., 2013), o el género del docente (Nevgi, Postareff, y Lindblom-Ylänne, 2004; Rosário et al., 2013).

Enfoques de enseñanza y tamaño de clase

Los resultados de la investigación en el contexto universitario no son concluyentes respecto a la relevancia del número de alumnos en clase para la adopción de un enfoque de enseñanza determinado. Por ejemplo, mientras que en el estudio de Singer (1996) se obtiene que a medida que aumenta el número del alumnado en clase los profesores están más predispuestos a orientar su enseñanza mediante un enfoque ITTF, en el trabajo de Stes et al. (2008) no se encontró relación entre el enfoque CCSF y el número de alumnos por clase. Globalmente, los resultados aportan sobre todo controversia, dado que hay estudios que apuntan efectos favorables asociados con la reducción del número del alumnado por clase (Pong y Pallas, 2001; Rosário et al., 2013; Rosário, Núñez, Valle, González-Pienda, y Lourenço, en prensa), pero también existen otros que permiten concluir lo contrario (p. e., Greenwald, Hedges, y Laine 1996; Konstantopoulos, 2008; Milesi y Gamoran, 2006). En conjunto, estos resultados sugieren la importancia de estudiar la relación entre el papel del profesor en clase (p. e., enfoque de enseñanza) y el tamaño de la clase.

Género y enfoques de enseñanza

Respecto a la relación entre las variables personales del profesor y la predilección por uno u otro enfoque de enseñanza, Lacey, Saleh y Gorman (1998) encontraron relación entre el género y el enfoque de enseñanza y, al igual que en el estudio de Nevgi et al. (2004), los hombres puntuaron más alto en el enfoque de enseñanza ITTF, mientras que las mujeres lo hacían en el CCSF.

Enfoques de enseñanza y años de experiencia

Stes et al. (2008) analizaron la relación entre la experiencia docente y el enfoque CCSF, hipotetizando que a mayor experiencia docente mayor sería la probabilidad de utilizar un enfoque CCSF. Los datos aportados por este estudio no confirmaron esta hipótesis, aunque los autores de la investigación sugieren tomar estos resultados con cierta cautela pues el grupo de profesores era pequeño (50 de una universidad belga). Sin embargo, Rosário, Núñez, Ferrando, Paiva, Lourenço, Cerezo y Valle (en prensa) obtuvieron evidencia de que a más años de experiencia más uso de un enfoque de enseñanza orientado a la construcción del conocimiento (CCSF).

Nivel educativo de los padres y rendimiento académico

Según los datos aportados por un buen número de estudios empíricos, el nivel educativo de los padres es un importante predictor del comportamiento del alumnado

en clase y de su rendimiento (Davis-Kean, 2005; Dearing, McCartney, y Taylor, 2001; Duncan y Brooks-Gunn, 1997; Dubow, Boxer, y Huesmann, 2009). Por ejemplo, Duncan y Brooks-Gunn (1997) concluyeron que el nivel educacional de las madres estaba conectado significativamente con el rendimiento intelectual de los niños incluso después de controlados algunos indicadores socioeconómicos como el rendimiento económico de la familia. Davis-Kean (2005) encontró relaciones positivas entre el nivel educacional de los padres y sus expectativas en relación con el éxito de sus hijos, sugiriendo que los padres con niveles educativos superiores implican a sus hijos activamente para que desarrollen expectativas personales ambiciosas.

Objetivos del presente estudio

Tal como queda claro en la revisión previa, los datos aportados por los estudios realizados hasta la fecha no son concluyentes respecto al papel de las variables del alumno y del contexto revisadas sobre el rendimiento académico de estudiantes preuniversitarios (y menos en el área específica de Biología). Además, no se dispone de información sobre la relevancia de cada una de las variables tomadas en la determinación del rendimiento cuando se consideran conjuntamente, ni tampoco hay estudios que analicen estas variables considerando los resultados al nivel del sujeto y al nivel de la clase. Por ello, el objetivo de la presente investigación consistió en analizar el grado de asociación del rendimiento académico de los estudiantes en Biología con ciertas variables del alumno (enfoques de aprendizaje, conocimientos previos, tiempo de estudio, grado de asistencia a clase, realización de deberes escolares), variables del profesor (enfoques de enseñanza, género de los profesores, años de experiencia), tamaño de la clase y el nivel de estudios de los padres.

Dado que sobre muchas de las variables consideradas en este estudio los datos aportados por la investigación pasada son poco concluyentes, y tomando en consideración que los resultados aportados por los trabajos revisados no han sido analizados desde una perspectiva multinivel, se plantea el presente estudio desde una perspectiva exploratoria. No obstante, la propia estrategia de análisis de los datos conlleva la búsqueda de respuestas a las siguientes cuestiones:

a) ¿Las variables del nivel de clase, examinadas en este estudio, condicionan significativamente el logro de los estudiantes en Biología? Si la respuesta a esta pregunta fuera afirmativa, entonces ¿qué variables de nivel de clase son relevantes en dicha determinación? En este nivel, en primer lugar, se espera que el enfoque de enseñanza sea una variable relevante, de modo que el rendimiento del alumnado será mayor en la medida en que los profesores desplieguen usualmente una instrucción centrada en el estudiante (en la construcción de significados), y será menor cuando su enfoque de enseñanza se encuentre centrado principalmente en la transmisión de información. En segundo lugar, en relación al resto de variables del nivel de clase, se espera que el tamaño de clase se encuentre relacionado negativamente con el rendimiento, mientras que la experiencia docente debería mostrar una asociación positiva con el rendimiento en Biología.
b) ¿Las variables de nivel individual analizadas explican significativamente el rendimiento del alumnado en Biología? Al igual que en caso anterior, si la variabilidad explicada a nivel individual fuera significativa, interesa conocer qué capacidad predictiva tiene cada una de estas variables. Tomando en consideración

los estudios previos, por una parte, se espera que cuanto más el alumno utilice un enfoque profundo de aprendizaje (centrado en la comprensión y adquisición de competencia) mayor será el rendimiento en Biología y, a la inversa, cuanto más utilice un aprendizaje superficial (interés por la adquisición de información y cumplir con criterios de logro externos) menor será el rendimiento académico en Biología. Por otra parte, aunque los resultados de la investigación pasada no son concluyentes, también se espera que el tiempo de estudio, la asistencia a clase, el nivel de conocimientos previos de Biología, la realización de deberes escolares y el nivel educativo de los padres muestren una asociación positiva con el rendimiento en esta área académica.

c) ¿Existe interacción entre el enfoque de aprendizaje (nivel de estudiante) y el enfoque de enseñanza (nivel de clase)? En concreto, ¿el enfoque de enseñanza de los profesores modera la relación entre el enfoque de aprendizaje de los estudiantes y el rendimiento en Biología?

Método

Participantes

En el estudio han participado 10 Institutos del norte de Portugal, los cuales fueron elegidos al azar de entre un total de 45 posibles. De estos institutos, han participado 57 profesores de Biología y sus correspondientes 988 estudiantes de tercero de Bachillerato. Los estudiantes presentaron las autorizaciones de sus padres para participar, y los profesores enviaron un correo electrónico al investigador principal comunicando su voluntad de participar en la investigación. De los 988 alumnos, 384 (38.9%) son varones y 604 (61.1%) mujeres, oscilando sus edades desde los 16 a los 19 años (M = 17.2; DT= .69). De los 57 profesores de Biología que participaron en la investigación, 11 (19.3%) son varones y 46 (80.7%) mujeres, oscilando su edad entre los 26 y los 61 años (M = 46.9, DT = 9.2). Su experiencia docente estuvo comprendida entre los 2 y los 36 años (M = 23.5; DT= 9.6).

Instrumentos de medida

Variables del estudiante

- *Enfoques de aprendizaje*. Los datos relativos a los enfoques de aprendizaje fueron obtenidos a través del cuestionario IEA (Inventario de Enfoques de Aprendizaje) (Rosário et al., 2007). El IEA está constituido por 12 ítems, que se contestan utilizando una escala tipo Likert de 5 puntos, entre 1 (completamente en desacuerdo) y 5 (completamente de acuerdo). Los análisis factoriales confirmatorios realizados mostraron una estructura factorial del IEA de dos factores (Rosário, Núñez, Ferrando et al., en prensa): enfoques superficial y enfoque profundo, con un buen ajuste del modelo, χ^2 (49) = 116.64, $p < .001$, χ^2/df = 2.38, GFI = .98, AGFI=.98, CFI = .99, TLI = .98, RMSEA = .03 (CI: .02 – .03). Los índices de fiabilidad (α de Cronbach) fueron muy satisfactorios: enfoque profundo (α = .91) y enfoque superficial (α = .90).
- *Asistencia a clase.* Esta variable fue evaluada al final de curso computando el número total de ausencias o faltas a la clase de Biología. Esta información fue recogida a finales de curso en la secretaría de los colegios participantes (M = 3.18; DT = 4.16).

- *Tiempo de estudio*. El tiempo fue evaluado diariamente durante una semana con una cuestión abierta, preguntando sobre el número de horas que el alumnado dedicaba a su estudio personal. Todos respondieron rellenando un diario que fue devuelto a los investigadores al final de la semana en un sobre cerrado. La media obtenida fue de 7.47 horas semanales de estudio (*DT* =5.52).
- *Conocimiento previo en Biología.* Esta variable fue evaluada a través de la nota obtenida por el alumnado en los dos cursos de Bachillerato. En Portugal, las notas oscilan entre 0 y 20 puntos, siendo el 10 la nota de corte para el aprobado. El alumnado fue distribuido de la siguiente forma: 1 para las notas entre 10 y 13 (n = 686; 45.6%), 2 para las notas entre 14 y 16 (n = 352; 23.4%) y 3 para las notas entre 17 y 20 (n = 466; 31.0%).
- *Deberes escolares*. Al final de curso, los profesores asignaron un 1 a todo el alumnado que había completado menos del 80% de los deberes asignados (41.4%), y un 2 cuando se había completado más del 80% (58.6%).

Variables de clase

- *Enfoques de enseñanza*. Los datos relativos a los enfoques de enseñanza fueron obtenidos a través del cuestionario IEE (Inventario de Enfoques de Enseñanza). Basado en el marco teórico asociado al modelo de Prosser y Trigwell (1999) y Ramsden et al. (2007), este instrumento está integrado por 12 ítems que aportan información sobre los dos enfoques de enseñanza (ITTF y CCSF). Como cada enfoque está constituido por una motivación y una estrategia, la escala también ofrece datos de las dos dimensiones de cada uno de los dos enfoques. Se contesta utilizando una escala tipo Likert de 5 puntos, entre 1 (completamente en desacuerdo) y 5 (completamente de acuerdo). Mediante análisis factorial confirmatorio se contrastó la estructura teórica de cuatro factores de primer orden (motivaciones y estrategias) y dos de segundo orden (enfoques). Los resultados mostraron un buen ajuste del modelo, $\chi^2(49) = 101.92$, $p < .001$, $\chi^2/df = 2.08$, GFI = .97, AGFI = .95, CFI = .98, RMSEA = .04 (.03 - .05), obteniendo evidencia, por tanto, de la validez de constructo del inventario (Rosário et al., 2010; Rosário, Núñez, Ferrando et al., en prensa). En cuanto a la fiabilidad, ambos factores mostraron niveles apropiados ($\alpha_{ITTF} = .92$ y $\alpha_{CCSF} = .94$).
- *Experiencia docente*. Los datos relativos a la experiencia docente fueron obtenidos en las secretarías de los institutos. La media obtenida fue de 22.81 años (*DT*=9.84).
- *Número de alumnos por clase*. La información relativa a la variable tamaño de la clase (número de alumnos por clase) fue obtenida en las secretarías de los institutos participantes.
- *Nivel educativo de los padres.* Esta variable fue categorizada del siguiente modo: 1 (enseñanza primaria), 2 (ESO), 3 (bachillerato), 4 (licenciatura) y 5 (Pos graduado). Esta información fue obtenida en las secretarías de los institutos participantes.

Rendimiento académico

Para cursar una licenciatura del área de Ciencias (p. e., Química, Medicina, Biología) el alumnado portugués deben realizar un examen nacional de Biología. Para preparar al alumnado para este examen, el Ministerio de Educación organiza tres

pruebas, una en cada trimestre. En la presente investigación fue calculada la media obtenida en las tres pruebas de Biología y tomada como medida del rendimiento académico en esta asignatura.

Procedimiento

El alumnado y los profesores fueron informados de los objetivos de esta investigación. La información fue recogida en el segundo semestre del curso (desde los meses de enero a abril) después de obtener la autorización de los directores de los institutos. Se indicó que para contestar a los inventarios tuvieran en cuenta la asignatura de Biología.

Análisis de datos

La naturaleza jerárquica de los datos aconseja analizarlos mediante un modelo jerárquico de dos niveles. El proceso de modelado estadístico será llevado a cabo en cuatro etapas. Inicialmente se formulará un modelo ANOVA de efectos aleatorios o modelo incondicional, el cual permite conocer la cantidad de varianza que pudiera explicarse a nivel individual (nivel 1) y a nivel de clase (nivel 2). Además, servirá como referente para evaluar la bondad de ajuste de modelos condicionales más complejos. Una vez realizado este primer paso, se ajustará el modelo correspondiente al nivel 2 (class-level) con el fin de conocer en qué medida las variables del contexto instruccional explican el rendimiento de los estudiantes. Seguidamente, se ajustará el modelo correspondiente a las variables del alumno (individual-level), con el fin de observar el grado en qué las variables del alumno predicen el rendimiento académico en biología. Finalmente, se procederá al estudio de la interacción entre ambos modelos (individual y class-level), al objeto de estimar el grado de interacción existente entre variables del nivel instruccional y variables del nivel de individuo.

En todos los análisis realizados, la variable dependiente fue la calificación obtenida al finalizar el curso predicha por un conjunto de variables explicativas registradas tanto en el nivel del estudiante (nivel 1) como en el nivel de clase (nivel 2). Las variables medidas en el nivel 1 fueron las siguientes: (a) *enfoques de aprendizaje*, medidos con la escala IEA y dicotomizada por encima de un punto de corte en función de la puntuación obtenida en esta escala. En concreto, si la puntuación promedio obtenida en las subescalas asociadas con el enfoque superficial (motivación y estrategia) > 9, entonces enfoque de aprendizaje = 0; mientras que si la puntuación promedio obtenida en las subescalas asociadas con el enfoque profundo (motivación y estrategia) > 9, entonces enfoque de aprendizaje = 1; (b) el *rendimiento previo*; (c) el *grado de realización de las tareas asignadas por los profesores*: inferior al 80% = 0, superior al 80% = 1; (d) el *género* del estudiante: varones = 0, mujeres = 1; (e) las *horas dedicadas al estudio* de la asignatura a lo largo de la semana: mínimo = 0, máximo = 25; (f) las *faltas de asistencia a clase* durante el curso escolar: mínimo = 0, máximo = 20; (g) el *nivel educativo de los padres*: primaria = 1,..., doctorado = 5.

Por lo que respecta a las variables explicativas registradas en el nivel 2, cabe destacar: (a) el *enfoque de enseñanza* de los profesores, medido mediante la escala IEE y dicotomizada por encima de un punto de corte en función de la puntuación obtenida en las subescalas de esta escala. En concreto, si la puntuación promedio obtenida en las subescalas asociadas con la docencia centrada en la transmisión de información (intención y estrategia) > 9, entonces enfoque de enseñanza = 0; mientras que si la

puntuación promedio obtenida en las subescalas asociadas con la docencia centrada en la construcción del conocimiento (intención y estrategia) > 9, entonces enfoque de enseñanza = 1; (b) el *género* de los profesores: varones = 0, mujeres = 1; (c) los *años de experiencia docente*: mínimo = 1, máximo = 36; (d) el *número de estudiantes por clase*: mínimo = 8, máximo = 33.

Resultados

Estadística descriptiva

En la Tabla 1 se ofrece la estadística descriptiva correspondiente a las variables de nivel 1 y de nivel 2 usadas en la presente investigación.

Tabla 1

Estadísticos Descriptivos de las Variables a Nivel de Estudiante y a Nivel de Clase

	M	*DT*	Mínimo	Máximo
Variables Nivel 1 (estudiante)				
Enfoque de aprendizaje	.64	.48	.00	1.00
Conocimientos previos	1.85	.85	1.00	3.00
Deberes escolares	.61	.49	.00	1.00
Género alumno	.61	.48	.00	1.00
Tiempo de estudio	7.79	5.77	.00	25.00
Absentismo escolar	3.03	4.19	.00	20.00
Nivel educativo de los padres	2.68	1.22	1.00	5.00
Variables Nivel 2 (clase)				
Enfoque de enseñanza	.77	.42	.00	1.00
Género profesor	.80	.40	.00	1.00
Nivel de experiencia docente	23.12	9.99	2.00	36.00
Número de alumnos por clase	20.28	4.77	8.00	33.00

Nota: Nivel 1 ($N = 988$); Nivel 2 ($N = 57$).

Análisis multinivel

Modelo incondicional de medias

Se comienza el análisis de los datos, ajustando el modelo nulo o incondicional de medias que sigue:

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

donde Y_{ij} es el rendimiento observado para el *i*-ésimo estudiante anidado en la *j*-ésima clase, γ_{00} es el rendimiento promedio global de los estudiantes, u_{0j} denota la variabilidad que existe entre los profesores en términos del rendimiento promedio de los estudiantes y e_{ij} denota la variabilidad que existe en el rendimiento de los estudiantes anidados en *j*-ésima clase. Se asume que los términos aleatorios del modelo son *NID* (normal e independientemente distribuidos) con media cero y varianza constante; o sea, $u_{0j} \sim NID(0, \tau_{00})$ y $e_{ij} \sim NID(0, \sigma_e^2)$. Repárese que se ha asumido que las clases

estudiadas representan una muestra aleatoria de una determinada población, lo que hace que las inferencias no sean exclusivas para la muestra de clases estudiadas.

Con este modelo incondicional se formula que el rendimiento se puede explicar mediante una parte fija, la cual contiene un valor global que es igual para todas las clases y para todos los estudiantes, más una parte aleatoria que indica la variabilidad asociada con los diferentes niveles implicados en el análisis, a saber: nivel del estudiante (nivel 1) y nivel del profesor o clase (nivel 2). Este modelo preliminar sirve como referente para comparar la bondad de ajuste de sucesivos modelos condicionales a los datos. En nuestro caso, se trata de verificar si los componentes de varianza asociados con el rendimiento de los estudiantes dentro de las clases y con el rendimiento promedio de los estudiantes entre las clases difieren significativamente de cero, pues si no fuera así no tendría sentido analizar los datos a dos niveles.

En la Tabla 2 se muestran los resultados obtenidos tras ajustar el modelo referido a los datos de la presente investigación. Como se puede observar, se constata que la estimación del rendimiento promedio en esta muestra de clases (13.02) difiere de cero (p <.0001). Sin embargo, el resultado más destacable es la existencia de diferencias estadísticamente significativas en el rendimiento de los estudiantes dentro de las clases (u_{0j} = 1.61; p < .0001), así como en su rendimiento promedio a través de las mismas (e_{ij} = 9.84; p < .0001). En el 95% de los casos cabe esperar que la magnitud de la variación entre las clases, en cuanto al rendimiento promedio se refiere, se encuentre dentro del intervalo (10.45, 15.56). Esto indica un rango moderado de variabilidad en los niveles de rendimiento promedio entre las clases en esta muestra de datos. A su vez, de la variabilidad observada en el rendimiento académico (1.62 + 9.84 = 11.46), es principalmente debida a las variables de nivel 1: un 85.9% se debe a las variables de nivel de estudiante y el 14.4% restante es debida a las variables de nivel de clase (unas clases generan más rendimiento que otras).

El grado de dependencia entre las observaciones de los estudiantes dentro de una misma clase, aproximadamente 0.141 en nuestro caso, impide el cumplimiento de la hipótesis de independencia, asumida por el modelo de regresión clásico, y aconseja el análisis de los datos a dos niveles (individuo y clase).

Tabla 2

Resumen de los Resultados Obtenidos con el Modelo Incondicional de Medias

Solución para los efectos fijos					
Efecto	Estimador	Error estándar	GL	Valor t	Pr > |t|
Intercepto	13.0233	.1974	56	65.98	< .0001
Estimadores parámetros de covarianza					
Par Cov	Efecto	Estimador	SE	Valor Z	Pr > Z
u_{0j}	Clases	1.6100	.4156	3.90	< .0001
e_{ij}	Residual	9.8398	.4560	21.58	< .0001
Estadísticos de ajuste					
Descripción	Valor				
Desvianza	5138.6				
Criterio AIC	5144.2				
Criterio BIC	5150.7				

Nota: SE = error estándar; GL = grados de libertad; Desvianza = menos dos veces el logaritmo de la función de máxima verosimilitud; AIC = Criterio de Información de Akaike; BIC = Criterio de Información Bayesiano.

Modelos con predictores a nivel de clase

El modelo incondicional de medias no contempla las características de los estudiantes ni de las clases; únicamente proporciona una base sobre la cual poder comparar modelos más complejos. Sin embargo, el rendimiento podría ser explicado por las características de los estudiantes que conforman las clases, por las características de cada clase, así como por el efecto conjunto de ambas. Por consiguiente, una vez que se ha puesto de relieve que el rendimiento promedio es más elevado en unas clases que en otras, se requiere comprender por qué el desempeño académico obtenido en unas clases es mayor que el obtenido en otras. Para dar cuenta de esto, se llevó a cabo un nuevo análisis incorporando las variables explicativas registradas en el nivel de clase (nivel 2), a saber, el enfoque de enseñanza, el género del profesor, el número de estudiantes por clase y los años de experiencia docente, prestando especial atención a la variable enfoque de enseñanza.

Específicamente, se formula a nivel-2 el modelo condicional que sigue:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(enfoques\_enseñanza)_j + \gamma_{02}(género)_j + \gamma_{03}(tamaño\_clase)_j + \gamma_{04}(\exp eriencia\_docente)_j + u_{0j} + e_{ij}.$$

donde Y_{ij} denota el rendimiento observado para el *i*-ésimo estudiante anidado en la *j*-ésima clase, γ_{00} representa el desempeño promedio de los estudiantes instruidos por docentes de experiencia media en grupos de tamaño medio, γ_{01} indica si el rendimiento de los estudiantes instruidos con métodos centrados principalmente en el docente (enfoque de enseñanza centrado en la trasmisión de información) difiere de los instruidos con métodos principalmente centrados en el discente (enfoque de enseñanza centrado en la construcción del conocimiento por parte del alumno), controlando los efectos de las variables género del profesor, número de alumnos por clase y experiencia docente; γ_{02} indica si el rendimiento de los estudiantes instruidos por mujeres difiere de los instruidos por varones, controlando los efectos de las variables enfoque de enseñanza, número de alumnos por clase y experiencia docente; γ_{03} denota el cambio en el rendimiento promedio de los estudiantes por cada unidad de aumento en el tamaño de los grupos de clase, controlando los efectos de las variables enfoque de enseñanza, género y experiencia docente; γ_{04} denota el cambio en el rendimiento promedio de los estudiantes como consecuencia del incremento de la experiencia del profesor, controlando los efectos de las variables enfoque de enseñanza, género y número de alumnos por clase. Finalmente, u_{0j} representa la variación en el rendimiento promedio entre las clases, mientras que e_{ij} representa la variación dentro de las mismas.

Los resultados de ajustar sendos modelos condicionales de intercepto aleatorio con predictores de nivel 2 aparecen recogidos en la Tabla 3. Por un lado, de acuerdo con el primero de los dos modelos ajustados (Modelo A), no hay evidencia de que exista un cambio estadísticamente significativo en el rendimiento promedio de los estudiantes en función del método de instrucción empleado (enfoque de enseñanza), del género de los profesores, del número de estudiantes por profesor y de los años de experiencia docente. Repárese que el panorama cambia ligeramente ajustando un modelo condicional más parco (Modelo B de la Tabla 2), pues la diferencia entre los valores del intercepto de

cada modelo es pequeña (13.16 - 12.31= 0.85). Aunque con el modelo reducido (Modelo B) se aprecia una relación marginalmente no significativa entre la forma de instruir de los profesores (enfoque de enseñanza) y el rendimiento promedio de los estudiantes (γ_{01} = .907; p = .055). Por otra parte, examinando la varianza correspondiente al nivel 2, tampoco se aprecia que ésta se reduzca significativamente al incorporar la variable forma de instruir (enfoque de enseñanza) en el nivel de la clase; específicamente, mientras la varianza incondicional valía 1.61 la varianza condicional vale 1.46. Esto indica que alrededor de un 10% de la variabilidad observada en el rendimiento promedio es explicado por el enfoque de enseñanza. También se puede observar que el coeficiente de correlación intra-clase condicional o residual sólo se redujo en dos centésimas, tras controlar el efecto de la variable forma de instruir de los profesores: antes .14 y ahora .12 (1.46/11.31= .12).

Tabla 3

Resumen de los Resultados Obtenidos con el Modelo Condicional de Intercepto Aleatorio con Múltiples Predictores de Nivel 2

	Modelo A			Modelo B		
Efectos fijos						
Efecto	Estimador (SE)	GL	Pr > |*t*|	Estimador (SE)	GL	Pr > |*t*|
Intercepto	13.162 (1.02)	52	<.0001	12.31(0.41)	55	<.0001
Enfoque de enseñanza	.724(0.47)	52	.131	.91(0.46)	55	.054
Género de profesores	-.037(0.49)	52	.940			
Nº alumnos por clase	-.052(0.04)	52	.200			
Experiencia docente	.015(0.02)	52	.459			
Efectos aleatorios						
Par Cov	Estimador (SE)	*Z*	Pr > *Z*	Estimador (SE)	*Z*	Pr > *Z*
u_{0k}	1.38(0.37)	3.68	<.0001	1.46(0.39)	3.75	<.0001
e_{ij}	9.85(0.46)	21.56	<.0001	9.84(0.45)	21.57	<.0001
Estadísticos de ajuste						
Descripción	Valor			Valor		
Desvianza	5132.9			5134.9		
Criterio AIC	5146.9			5142.9		
Criterio BIC	5162.2			5151.1		

Nota. SE = error estándar; GL = grados de libertad; Desvianza = menos dos veces el logaritmo de la función de máxima verosimilitud; AIC = Criterio de Información de Akaike; BIC = Criterio de Información Bayesiano.

Si bien es cierto que el Modelo B no permite concluir, estadísticamente hablando, que la forma de instruir de los profesores afecte al rendimiento de los estudiantes, dicha variable no será extraída del análisis por resultar marginalmente no significativa (p = .055) y ser central en la presente investigación Además, conviene tener presente que el Modelo B, con los criterios de información más pequeños, AIC y BIC en nuestro caso, es el modelo que logra un mejor ajuste a los datos. A la misma conclusión hubiésemos llegado de haber utilizado el AIC basado en la verosimilitud condicional, en

lugar de la verosimilitud marginal, y el DIC basado en la inferencia bayesiana (véase Vallejo, Tuero, Núñez, y Rosário, en prensa).

Modelos con predictores a nivel de estudiante

El modelo ajustado previamente sólo contempla el efecto de las variables de composición y contexto de las clases pero no considera las características de los estudiantes, por lo que se desconocen las razones que llevan a que existan diferencias en el rendimiento de los estudiantes, ni tampoco hay evidencias de que la variabilidad observada entre las clases no sea más que un artefacto debido al perfil diferente de los estudiantes que son instruidos por los profesores en cada clase. Para responder a esta cuestión se realiza un nuevo análisis con siete variables de nivel estudiante, a saber, *rendimiento previo*, *realización de deberes escolares*, *género*, *enfoque de aprendizaje*, *nivel educativo de los padres*, *horas dedicadas al estudio* y *absentismo escolar*; estas dos últimas variables centradas con respecto a la media de su grupo. Inicialmente, se realizó un testeo para comprobar la variación aleatoria de las pendientes una tras otra y se observó que todas se mantenían constantes, a excepción de la correspondiente al factor enfoques de aprendizaje que variaba a lo largo de las clases.

El modelo de coeficientes aleatorios resultante puede ser escrito como sigue:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10}(tiempo\_estudio)_{ij} + \gamma_{20}(conocimiento\_previo)_{ij} + \gamma_{30}(deberes)_{ij} + \\ & \gamma_{40}(absentismo)_{ij} + \gamma_{50}(género)_{ij} + \gamma_{60}(educación\_padres)_{ij} + \\ & \gamma_{70}(enfoque\_aprendizaje)_{ij} + u_{0j} + u_{1j}(enfoque\_aprendizaje)_{ij} + e_{ij} \end{aligned}$$

donde Y_{ij} denota el rendimiento observado para el *i*-ésimo estudiante anidado en la *j*-ésima clase, γ_{00} representa el desempeño promedio de los estudiantes, γ_{10} denota el cambio en el rendimiento promedio de los estudiantes por cada unidad de aumento en las horas de estudio, controlando los efectos de las variables restantes; γ_{20} indica la relación existente entre el conocimiento previo y el rendimiento, controlando los efectos de las variables restantes; γ_{30} indica la relación existente entre la realización de los deberes y el rendimiento, controlando los efectos de las variables restantes; γ_{40} denota el cambio en el rendimiento promedio de los estudiantes por cada unidad de aumento en el absentismo escolar, controlando los efectos de las variables restantes; γ_{50} indica la relación existente entre el género de los estudiantes y el rendimiento de los mismos, controlando los efectos de las variables restantes; γ_{60} indica la relación existente entre el nivel formativo de los padres y el rendimiento de los hijos, controlando los efectos de las variables restantes; γ_{70} indica cómo afecta el método de estudio (enfoque de aprendizaje) al rendimiento, controlando los efectos de las variables restantes. Finalmente, u_{1j} indica si la relación entre los enfoques de aprendizaje y el rendimiento promedio varían a través de las clases.

En la Tabla 4 se muestran los resultados más importantes que se obtuvieron tras ajustar sendos modelos de coeficientes aleatorios. De acuerdo con el primero de los dos modelos ajustados, Modelo A, no hay evidencia de que existan cambios en el rendimiento promedio en función de las horas dedicadas al estudio de la asignatura a lo

largo de la semana (p = .074). Consideramos de interés mencionar que si existía una relación estadísticamente significativa entre las variables horas de estudio y rendimiento cuando no se controlaba el efecto de la variable enfoques de aprendizaje usadas por los estudiantes; no obstante, como se observa en la Tabla 4, dicha relación resultó marginalmente no significativa cuando se controló el efecto esta última variable. Además, no existían diferencias estadísticamente significativas de género en el rendimiento de los estudiantes (p = .389). Obsérvese que tampoco se pudo rechazar la hipótesis nula de falta de asociación entre la variable grado de realización de los deberes asignados por los profesores y la variable rendimiento (p = .431).

Por último, los resultados reportados en la Tabla 4 para el Modelo A, también ponen de relieve que la relación entre la forma de estudiar y el rendimiento promedio dentro de las clases variaba significativamente a lo largo de las mismas (u_{1j}= .948, p = .015). Sin embargo, no existía ninguna evidencia de que la forma de estudiar dependiese del rendimiento promedio de la clase, pues la covarianza pendiente e intercepto a través de las clases no fue estadísticamente significativa (p = .227).

Tabla 4

Resumen de los Resultados Obtenidos con los Modelos de Interceptos y Pendientes Aleatorias con Múltiples Predictores de Nivel 1

	Modelo A			Modelo B		
Efectos fijos						
Efecto	Estimador (SE)	GL	Pr > |*t*|	Estimador (SE)	GL	Pr > |*t*|
Intercepto	9.846(.477)	56	<.0001	9.779(.446)	56	<.0001
Tiempo de estudio	.029(.020)	924	.0736			
Conocimientos previos	.692(.182)	924	.0001	.745(.179)	927	<.0001
Deberes escolares	.904(1.147)	924	.4311			
Absentismo escolar	-.105(.024)	924	<.0001	-.109(.024)	927	<.0001
Género alumnos	-.985(1.143)	924	.3891			
Nivel de estudios padres	.356(.086)	924	.0001	.372(.086)	927	<.0001
Enfoques de aprendizaje	1.746(.297)	924	<.0001	1.821(.258)	927	<.0001
Efectos aleatorios						
Par Cov	Estimador (SE)	*Z*	Pr > *Z*	Estimador (SE)	*Z*	Pr > *Z*
u_{0j}	1.128(.506)	2.23	.0130	.677(.288)	2.35	0.0094
u_{1j}	1.833(.857)	2.14	.0162	.984(.459)	2.17	0.0151
u_{01}	-.720(.561)	-1.28	.2268			
e_{ij}	8.218(.447)	21.05	<.0001	8.343(.394)	21.19	<.0001
Estadísticos de ajuste						
Descripción	Valor			Valor		
Desvianza	4975.5			4980.1		
Criterio AIC	4999.5			4996.1		
Criterio BIC	5024.0			5012.4		

Nota. SE = error estándar; GL = grados de libertad; Desvianza = menos dos veces el logaritmo de la función de máxima verosimilitud; AIC = Criterio de Información de Akaike; BIC = Criterio de Información Bayesiano.

De lo dicho se colige la conveniencia de ajustar un modelo más simple, por ejemplo, uno en el cual aún se permita al intercepto y a la pendiente variar a través de las clases, pero se elimine las variables explicativas que no resultaron significativas en el paso anterior; es decir, que es factible que un modelo más parco, Modelo B, ofrezca un ajuste razonable de los datos. Lo anteriormente dicho se puede comprobar fácilmente examinando los estadísticos de ajuste mostrados en la Tabla 4, recuérdese que buscamos modelos con los valores más pequeños de los criterios AIC y BIC. Dado que el Modelo A no explica mejor los datos que el Modelo B y éste es más parsimonioso, seleccionaremos el segundo modelo.

Lo primero que cabe destacar tras ajustar el Modelo B es que, en promedio, existe una relación estadísticamente significativa dentro de las clases entre los enfoques de aprendizaje de los estudiantes y su rendimiento académico (γ_{70}= 1.821; p < .0001). Más en concreto, tomando en cuenta el signo de la asociación, los resultados obtenidos indican que los estudiantes que suelen emplear un enfoque profundo en su estudio alcanzan logros significativamente mayores que los estudiantes que suelen emplear un enfoque superficial. También se constató que el conocimiento previo predice positiva y significativamente el rendimiento académico presente (γ_{20}= .745; p < .0001). Además, se encontró que tanto el absentismo escolar como nivel educativo de los padres afectaban significativamente al rendimiento de los participantes (γ_{40}= -.109, p < .0001 y γ_{60}= .372, p < .0001, respectivamente); sin embargo, mientras que el primero reducía el rendimiento el segundo lo incrementaba. Resaltar, por último, que la variabilidad residual dentro de las clases y a través de los mismas aún permanece significativa (e_{ij} = 8.343, p < .0001*;* u_{0j} = .667, p = .009). Por este motivo, es muy importante seguir investigando otras posibles causas no tenidas en cuenta en este análisis y que pueden explicar, al menos en parte, tales variabilidades. Repárese, no obstante, que en este caso no sólo se redujo la varianza dentro de las clases desde 9.848 hasta 8.343, sino que también se redujo varianza entre las clases desde 1.385 hasta .677.

Modelos con predictores de niveles 1 y 2

Por ello, una vez ajustado por separado un modelo para las variables registradas en el nivel de los estudiantes (nivel 1) y otro para las variables registradas en el nivel de las clases (nivel 2), consideraremos un modelo que contenga variables de ambos niveles. Dicho modelo nos permitirá detectar la posible existencia interacciones cruzadas o transversales entre los mismos.

Combinando el modelo ajustado en el nivel estudiante y el modelo ajustado en el nivel clase se obtiene la ecuación:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01}(enfoque\_enseñanza)_j + \gamma_{10}(conocimiento\_previo)_{ij} + \gamma_{20}(absentismo)_{ij} + \\ & \gamma_{30}(estudios\_padres)_{ij} + \gamma_{40}(enfoque\_aprendizaje)_{ij} + \gamma_{11}(enfoque\_enseñanza)_{ij} \times \\ & (enfoque\_aprendizaje)_{ij} + u_{0j} + u_{1j}(enfoque\_aprendizaje)_{ij} + e_{ij} \end{aligned}$$

la cual pone de manifiesto que el rendimiento puede ser visto como una función de los efectos fijos más los aleatorios. Los efectos fijos serían: media general (γ_{00}), efecto principal del enfoque de enseñanza (γ_{01}), efecto principal rendimiento previo (γ_{10}),

efecto principal absentismo escolar (γ_{20}), efecto principal nivel educativo de los padres (γ_{30}), efecto principal enfoque de aprendizaje (γ_{40}), e interacción cruzada entre los enfoques de enseñanza y los enfoques de aprendizaje (γ_{11}). Los efectos aleatorios representan la variabilidad que existe entre las clases (u_{0j}), entre los enfoques de aprendizaje a través de las clases (u_{1j}) y dentro de las clases (e_{ij}). Finalmente, dado que todas las cuestiones que motivan este análisis ya han sido especificadas, excepto la referida a la interacción cruzada, señalar que estimamos γ_{11} para examinar si el enfoque de enseñanza centrado en el profesor (consistente en transmitir información) difiere del enfoque de enseñanza centrado en los estudiantes (que consiste en facilitar la construcción del conocimiento por parte del alumno) en términos de la fuerza de asociación entre los enfoques de enseñanza y el rendimiento académico del alumnado.

En la Tabla 5 aparecen los resultados más importantes que se obtuvieron tras ajustar el modelo que incluye predictores de nivel 1 y de nivel 2. En concreto, se constata que el enfoque de enseñanza empleado por los profesores no resultó estadísticamente significativo como efecto principal (γ_{01} = .673, p = .125), aunque sí lo hizo como efecto secundario, a través de su interacción con los enfoques de aprendizaje del alumnado (γ_{11} = -1.403, p = .018). No obstante, el alumnado instruido por profesores que utilizan preferentemente un enfoque de enseñanza centrado en el discente obtienen un rendimiento promedio ligeramente superior (10.10) al de aquellos otros estudiantes instruidos por profesores que utilizan preferentemente un enfoque de enseñanza centrado en el docente (9.42). Con respecto a la interacción, conviene resaltar que los profesores que utilizaban habitualmente un enfoque de enseñanza centrado en el discente diferían de los profesores que utilizaban habitualmente un enfoque de enseñanza centrado en el docente en términos de la fuerza de asociación entre los enfoques de aprendizaje de los estudiantes y el rendimiento obtenido en biología. En otras palabras, debido al efecto moderador ejercido por la variable enfoque de enseñanza de los profesores, las diferencias de desempeño entre los estudiantes que utilizaban preferentemente un enfoque profundo y los que utilizaban un enfoque superficial era superior bajo aquellos profesores que utilizaban preferentemente un enfoque de enseñanza centrado en el docente que bajo los que utilizaban en su docencia un enfoque preferentemente centrado en el discente.

Finalmente, conviene advertir que los componentes de varianza de intercepto y pendiente aún se mantienen estadísticamente significativas (p = .011 y p = .022, respectivamente), lo que indica una variación significativa a través de las clases en ambos coeficientes. Nótese también, que la adición de la variable enfoque de enseñanza y su interacción cruzada con el enfoque de aprendizaje redujo ligeramente la varianza residual del intercepto (≈ 3%) y varianza residual de la pendiente para los enfoques de aprendizaje (≈ 13%), en comparación con la estimada para el modelo de coeficientes aleatorios del la sección anterior. No obstante, el rechazo de la hipótesis nula estaría indicando que todavía queda una variación significativa del rendimiento promedio entre las clases por ser explicada. Es previsible que la inclusión de variables adicionales a nivel de la clase redujese aún más la varianza correspondiente a las clases. Por consiguiente, existen adicionales características de los estudiantes y de los profesores no tenidas en cuenta en este análisis que podrían explicar la variación reseñada.

Tabla 5

Resumen de los Resultados Obtenidos con el Modelo Combinado de Interceptos Aleatorios

Efectos fijos

Efecto	Estimador	(SE)	GL	Valor *t*	Pr > |*t*|
Intercepto	9.424	(.516)	55	18.28	<.0001
Enfoque de enseñanza	.673	(.432)	55	1.56	.1248
Conocimientos previos	.694	(.178)	926	3.90	<.0001
Absentismo escolar	-.108	(.024)	926	-4.49	<.0001
Nivel estudios de padres	.374	(.086)	926	4.36	<.0001
Enfoque de aprendizaje	2.884	(.526)	926	5.49	<.0001
Enfoque de enseñanza × Enfoque de aprendizaje	-1.403	(.594)	926	-2.37	.0184

Efectos aleatorios

Par Cov	Efecto	Estimador	SE	Valor *Z*	Pr > *Z*
u_{0j}	Clases	.658	.267	2.30	.0108
u_{1j}	Enfoque de aprendizaje	.867	.431	2.01	.0222
e_{ij}	Residual	8.319	.392	21.25	<.0001

Estadísticos de ajuste

Descripción	Valor
Desvianza	4974.3
Criterio AIC	4994.3
Criterio BIC	5014.8

Nota. SE = error estándar; GL = grados de libertad; Desvianza = menos dos veces el logaritmo de la función de máxima verosimilitud; AIC = Criterio de Información de Akaike; BIC = Criterio de Información Bayesiano.

Discusión

El objetivo de la presente investigación consistió en analizar en qué medida el rendimiento académico en Biología, de los estudiantes de último curso de bachillerato, es predicho por ciertas variables del alumno (enfoques de aprendizaje, conocimientos previos, tiempo de estudio, grado de asistencia a clase, realización de deberes escolares), variables del profesor (enfoques de enseñanza, género, experiencia docente) y variables del contexto (tamaño de la clase, nivel de estudios de los padres). Dado que los datos tenían una estructura jerárquica (alumnos dentro de clases), fueron analizados a partir de una estrategia multinivel. Mediante este tipo de análisis, el presente estudio no sólo permitió conocer la relevancia de las variables a nivel de estudiante y a nivel de clase en su predicción del rendimiento en Biología sino también estudiar la interacción entre variables de ambos niveles, aspecto escasamente estudiado en la investigación pasada, pero de especial importancia teórica y aplicada.

A nivel general, se obtuvo que, mientras que las hipótesis formuladas a nivel de clase resultaron principalmente no confirmadas, las hipótesis a nivel de estudiante fueron en gran medida confirmadas. Así, se constató que la mayor parte de la variabilidad en el rendimiento en Biología estaba asociada con las variables tomadas a nivel de estudiante (el 85.6%), mientras que las variables tomadas a nivel de clase sólo aportaron un 14.4% de la misma. Sin embargo, de especial relevancia resultaron los

datos correspondientes a la interacción entre cómo enseñan los profesores, cómo aprenden los estudiantes y el rendimiento académico obtenido. Seguidamente, se discuten los hallazgos más relevantes.

Análisis a nivel de estudiante

En relación con las variables analizadas a nivel de estudiante (nivel 1), resultaron ser buenos predictores del rendimiento en Biología el conocimiento previo de la materia, el nivel de absentismo escolar, el nivel de estudios de los padres y el enfoque de aprendizaje, siendo esta variable la más relevante en esta ecuación. Ni el tiempo de estudio, ni la cantidad de deberes realizados ni el género de los estudiantes mostraron efectos principales significativos.

En cuanto a los efectos encontrados significativos, como era de esperar, se obtuvo que a mayor nivel de conocimientos previos mayor rendimiento en biología. Asimismo, también se observó que a mayor absentismo menor rendimiento académico (Reid, 2006). Al igual que en algunos trabajos previos, en este estudio también se halló que a mayor nivel de estudios de los padres mayor es el rendimiento en Biología de los hijos (Davis-Kean, 2005; Dubow et al., 2009). Finalmente, en esta investigación se aporta evidencia clara de que cuanto más se utilice un enfoque profundo para el estudio mayor será el rendimiento y cuanto más superficial sea el enfoque de aprendizaje utilizado en el proceso de aprendizaje menor es el rendimiento en Biología. Aunque algunos trabajos habían aportado dudas sobre esta relación (Entwistle, 1991; Rosário, Núñez et al., 2010; Struyven et al., 2006), los datos de este trabajo apoyan claramente que los beneficios provienen del uso de un enfoque profundo, el cual implica una motivación intrínseca, o con orientación a la tarea, y el uso de estrategias cognitivas y metacognitivas necesarias para la comprensión y elaboración de la información.

Por lo que respecta a las variables no significativas en la explicación del rendimiento en Biología (tiempo de estudio, género del estudiante y cantidad de deberes realizados de los prescritos por los profesores), el tiempo dedicado al estudio de esta asignatura merece un comentario especial. En concreto, si bien no resultó relevante cuando se incluyeron en la ecuación todas las variables del estudiante, su efecto principal es? ¿significativo si se elimina de la ecuación las variables que resultaron significativas (conocimientos previos, absentismo escolar, nivel educativo de los padres y enfoques de estudio). Al contrario de lo que pudiera parecer, el tiempo de estudio sí es una variable importante, pero que al contemplar otras variables como los enfoques de aprendizaje, el efecto de aquella variable se vehicula a través de esta última (de hecho el estudio con un aprendizaje profundo conlleva mayor cantidad de tiempo que el estudio utilizando un enfoque superficial). En cuanto a las otras dos variables, nuestros datos indican que hacer más o menos deberes de los prescritos no explica una cantidad significativa de variabilidad en el rendimiento. ¿Cómo explicar estos datos? Por una parte, habría que considerar que el error de estimación fue alto (1.137), quizás debido a la dicotomización de la variable (lo que también ocurre con el error de estimación del género, 1.134). Por otra parte, es posible que, al igual que en el caso del tiempo de estudio, el efecto de la cantidad de deberes realizados también podría estar subsumido por la utilización de un determinado enfoque de aprendizaje (es posible que el trabajo escolar realizado con un enfoque profundo conlleve la realización de un mayor número de deberes escolares y de más tiempo de estudio, comparado con el trabajo realizado utilizando un enfoque superficial). Por ello, tal como se comentó para el tiempo de

estudio, la cantidad de deberes puede ser una variable más importante de lo que pudiera derivarse de los resultados del análisis cuando están todas las variables presentes. Futuras investigaciones deberían analizar en profundidad esta hipótesis (midiéndola como una variable continua), a la vez que considerarla como una variable de nivel de clase.

Análisis a nivel de clase

Ninguna de las variables incluidas en la ecuación a nivel de clase mostró efectos principales significativos. Únicamente, los enfoques de enseñanza mostraron un leve efecto principal sobre el rendimiento en biología a este nivel de análisis ($p < .1$), si bien este limitado efecto se disipó una vez que la forma de enseñar de los profesores se puso en relación (interacción) con la forma de estudiar del alumnado.

Interacción entre enfoques de enseñanza y enfoques de aprendizaje

Este estudio aporta información relevante y novedosa respecto de la interacción entre los enfoques de aprendizaje del alumno (nivel 1) y los enfoques de enseñanza de los profesores (nivel 2). Como ya se indicó, los resultados a nivel de estudiante indicaron que cuanto más el alumnado estudia con un enfoque profundo mayor es su rendimiento y, viceversa, cuanto más utilizan un enfoque superficial menor es el rendimiento. Cuando se tuvo en cuenta los dos niveles de análisis, se confirmó que esta diferencia en el rendimiento era mayor en alumnado instruido por profesores con un enfoque de enseñanza principalmente centrado en transmitir información, que en el alumnado cuyos profesores usaban un enfoque de enseñanza orientado preferentemente a ayudar al alumno a construir significados (desarrollo de procesos de comprensión y elaboración de la información). ¿Por qué puede ocurrir esto? Es posible que cuando un profesor plantea su estrategia de enseñanza centrada en la organización y transmisión de la información, el aprendizaje y el rendimiento del alumnado estén determinados de modo significativo por sus características personales aquí consideradas y por otras no consideradas como por ejemplo los objetivos, mientras que si el profesor promueve contextos instruccionales en los que se solicita al estudiante una implicación activa y significativa para la construcción personal del conocimiento, el alumnado tiende a no utilizar tanto un enfoque superficial, pues no sería útil en ese contexto de enseñanza.

Limitaciones del estudio

La presente investigación ha supuesto un gran esfuerzo por reunir datos suficientes, de alumnado, padres y profesores, como para realizar análisis de desde una perspectiva multinivel. Sin embargo, existen algunos aspectos que podrían modular la interpretación de los resultados obtenidos. En primer lugar, el hecho de que la información sobre la forma de aprender y la de enseñar fue obtenida mediante instrumentos tipo auto-informe, por lo que dicha información tiene que ver con lo que alumnado y profesores creen hacer en sus respectivas tareas. Aunque muy común en la investigación en el campo de la educación, esto no deja de ser una limitación pues los resultados deben ser tomados por lo que profesores y estudiantes creen que hacen y no lo que en realidad ocurre. En segundo lugar, las conclusiones derivadas del estudio podrían no ser completamente transferibles a otras disciplinas académicas, o a otras edades del alumnado (Stes et al., 2008). Por ello, sería de interés que en futuras

investigaciones se buscasen respuestas a los muchos interrogantes que todavía persisten en este campo científico.

Referencias

Alexander, P. A., Kulikowich, J. M., y Schulze, S. K. (1994).The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences, 6,* 379-397. http://dx.doi.org/10.1016/1041-6080(94)90001-9.

Cooper, H., Robinson, J. C., y Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research, 76,* 1-62. http://dx.doi.org/10.3102/00346543076001001.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*, 294-304. http://dx.doi.org/10.1037/0893-3200.19.2.294.

Dearing E., McCartney K., y Taylor, B.A. (2001). Change in family income matters more for children with less. *Child Development*, *72*, 1779-1793. http://dx.doi.org/10.1111/1467-8624.00378.

Dettmers, S., Trautwein, U., y Ludtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement, 20,* 375-405. http://dx.doi.org/10.1080/09243450902904601.

Dubow, E., Boxer, P., y Huesmann, L. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill Palmer Quarterly*, *55*, 224-249. http://dx.doi.org/10.1353/mpq.0.0030.

Duncan, G. J., y Brooks-Gunn, J. (1997). *Consequences of growing up poor*. New York: Russell Sage Foundation.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education, 22*, 201-204. http://dx.doi.org/10.1007/BF00132287.

Entwistle, N. J. (2009). *Teaching for understanding at University: Deep approaches and distinctive ways of thinking*. Basingstoke, UK: Palgrave Macmillan.

Gortner-Lahmers, A., y Zulauf, C. R. (2000). Factors associated with academic time use and academic performance of college students: A recursive approach. *Journal of College Student Development, 41*, 544-556.

Greenwald, R., Hedges, L. V., y Laine, R. D. (1996). The effects of school resources on student achievement. *Review of Educational Research*, *66*, 361-396. http://dx.doi.org/10.2307/1170528.

Jonanssen, C. (2011). The dynamics of absence behaviour: Interrelations between absence from class and absence in class. *Educational Research*, *53*, 17-32.

Konstantopoulos, S. (2008). Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *Elementary School Journal*, *108*, 275-291. http://dx.doi.org/10.1086/528972.

Lacey, C. H., Saleh, A., y Gorman, R. (1998, October). *Teaching nine to five: A study of the teaching styles of male and female professors*. Paper presented at the annual meeting of the Women in Educational Leadership Conference, Lincoln, NE.

Lopes, J., y Santos, M. (2013).Teachers' beliefs, teachers' goals and teachers' classroom management: A study with primary teachers. *Revista de Psicodidáctica, 18*, 5-24. doi: 10.1387/RevPsicodidact.4615.

Marton, F., y Säljö, R. (1976). On qualitative differences in learning: I – Outcome and process. *British Journal of Educational Psychology, 46*, 4-11. http://dx.doi.org/10.1111/j.2044-8279.1976.tb02980.x.

McIntyre-Bhatty, K. (2008). Truancy and coercive consent: Is there an alternative? *Educational Review*, *60*, 375-390. http://dx.doi.org/10.1080/00131910802393407.

Milesi, C., y Gamoran, A. (2006). Effects of class size and instruction on kindergarten achievement. *Educational Evaluation and Policy Analysis*, *28*, 287-313. http://dx.doi.org/10.3102/01623737028004287.

Miñano, P., y Castejón, J. L. (2011). Cognitive and motivational variables in the academic achievement in language and mathematics subjects: A structural model. *Revista de Psicodidáctica, 16*, 203-230.

Nevgi, A., Postareff, L., y Lindblom-Ylänne , S. (2004, June). *The effect of discipline on motivational and self-efficacy beliefs and on approaches to teaching of Finnish and English university teachers*. A paper presented at the EARLI SIG Higher Education Conference.

Núñez, J. C., Rosário, P., Vallejo, G., y 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. http://dx.doi.org/10.1016/j.cedpsych.2012.10.002.

OECD (2010). *PISA 2009 Results: What students know and can do: Student performance in reading, mathematics and science* (Volume I). Retrieved at http://dx.doi.org/10.1787/9789264091450-en.

Paschal, R. A., Weinstein, T., y Walberg, H. J. (1984). The effects of homework on learning: A quantitative synthesis. *Journal of Educational Research*, *78*, 97-104.

Plant, E. A., Ericsson, K. A., Hill, L., y Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology, 30,* 96-116. http://dx.doi.org/10.1016/j.cedpsych.2004.06.001.

Pong, S., y Pallas, A. (2001). Class size and eighth-grade math achievement in the united states and abroad. *Educational Evaluation and Policy Analysis*, *23*, 251-273. http://dx.doi.org/10.3102/01623737023003251.

Prosser, M., y Trigwell, K. (1999). *Understanding learning and teaching. The experience in higher education*. Buckingham, UK: Open University Press.

Prosser, M., Ramsden, P., Trigwell, K., y Martin, E. (2003). Dissonance in experience of teaching and its relation to the quality of student learning. *Studies in Higher Education*, *28*, 37-48. http://dx.doi.org/10.1080/03075070309299.

Prosser, M., Trigwell, K., y Taylor, P. (1994). A phenomenographic study of academics' conceptions of science learning and teaching. *Learning and Instruction, 4*, 217-231. http://dx.doi.org/10.1016/0959-4752(94)90024-8.

Ramsden, P., Prosser, M., Trigwell, K., y Martin, E. (2007). University teachers' experiences of academic leadership and their approaches to teaching. *Learning and Instruction, 17,* 140-155. http://dx.doi.org/10.1016/j.learninstruc.2007.01.004

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

Reid, K. (2006). An evaluation of the views of secondary staff towards school attendance issues. *Oxford Review of Education, 32*, 303-324. http://dx.doi.org/10.1080/03054980600775557.

Ronning, M. (2011). Who benefits from homework assignments? *Economics of Education Review, 30,* 55-64.

Rosário, P., Mourão, R., Baldaque, M., Nunes, T., Núñez, J. C., González-Pienda, J., y Valle, A. (2009). Tareas para casa, autorregulación del aprendizaje y rendimiento en matemáticas. *Revista de Psicodidáctica, 14*, 179-192.

Rosário, P., Mourão, R., Núñez, J. C., González-Pienda, J. A., Solano, P., y Valle, A. (2007). Evaluating the efficacy of a program to enhance college students' self-regulation learning processes and learning strategies. *Psicothema, 19*, 353-358.

Rosário, P., Núñez, J. C., González-Pienda, J. A., Valle, A., Trigo, L. y Guimarães, C. (2010). Enhancing self-regulation and approaches to learning in first-year college students: A narrative-based program assessed in the Iberian Peninsula. *European Journal of Psychology of Education*, *25*, 411-428. http://dx.doi.org/10.1007/s10212-010-0020-y.

Rosário, P., Núñez, J. C., Valle, A., Paiva, O., y Polydoro, S. (2013). Approaches to teaching in high school when considering contextual variables and teacher variables. *Revista de Psicodidáctica, 18,* 25-45. doi:10.1387/RevPsicodidact.6215.

Rosário, P., Núñez, J. C., Valle, A., González-Pienda, J. A., y Lourenço, A. (en prensa). Grade level, study time, and grade retention and their effects on motivation, self-regulated learning strategies, and mathematics achievement: a structural equation model. *European Journal of Psychology of Education.* doi:10.1007/s10212-012-0167-9.

Rosário, P., Núñez, J. A., Ferrando, J. P., Paiva, O., Lourenço, A., Cerezo, R., y Valle, A. (2013). The relationship between approaches to teaching and approaches to studying: A two-level structural equation model for biology achievement in high school. *Metacognition and Learning, 8,* 47-77. doi: 10.1007/s11409-013-9095-6.

Singer, E. (1996). Espoused teaching paradigms of college faculty. *Research in Higher Education*, *37*, 659-679. http://dx.doi.org/10.1007/BF01792951.

Stes, A., Gijbels, D., y Van Petegem, P. (2008). Student-focused approaches to teaching in relation to context and teacher characteristics. *Higher Education*, *55*, 255-267. http://dx.doi.org/10.1007/s10734-007-9053-9.

Struyven, K., Dochy, F., Janssens, S., y Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and Instruction, 16*, 279-294. http://dx.doi.org/10.1016/j.learninstruc.2006.07.001.

Trautwein, U., y Köller, O. (2003). The relationship between homework and achievement—still much of a mystery. *Educational Psychology Review*, *15*, 116-145. http://dx.doi.org/10.1023/A:1023460414243.

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

José Carlos Núñez, Catedrático de Dificultades de Aprendizaje de la Universidad de Oviedo (España) y Director del Departamento de Psicología. Sus dos principales líneas de investigación son: a) dimensiones psicológicas y educativas del aprendizaje auto-regulado en contextos educativos; b) dificultades del aprendizaje escolar y TDAH. Actualmente, es responsable de un proyecto del Plan Nacional de Investigación (EDU2010-16231).

Guillermo Vallejo, Catedrático de Metodología de las Ciencias del Comportamiento de la Universidad de Oviedo (España). Imparte docencia sobre Diseños de Investigación en la Facultad de Psicología de dicha universidad. Es Investigador Principal de un proyecto del Plan Nacional de Investigación (PSI-2011-23395/PSIC).

Pedro Rosário, Profesor Titular de Psicología de la Educación de la Universidad de Minho (Portugal). Sus dos principales líneas de investigación son: a) dimensiones psicológicas y educativas del aprendizaje auto-regulado; b) procesos de autorregulación en ambientes de aprendizaje tecnológicos y en pizarras electrónicas. Tiene numerosas publicaciones en su país y en el extranjero en cualquiera de las dos líneas de investigación. Es el investigador principal del grupo GUIA (www.guia-psiedu.com).

Ellián Tuero, Becaria de FPI, en fase de contratada, en el Departamento de Psicología de la Universidad de Oviedo (España). Imparte docencia en la Facultad de Psicología y en la de Formación del Profesorado y Educación. A nivel de investigación, se ha especializado en el análisis de datos en diseños de medidas repetidas y en estructuras que requieren estrategias de análisis multinivel.

Antonio Valle, Catedrático de Psicología de la Educación y Director del Departamento de Psicología Evolutiva y de la Educación en la Universidad de A Coruña (España). La motivación académica, las estrategias de estudio y el aprendizaje auto-regulado son sus principales tópicos de investigación. Actualmente, es responsable de un Proyecto de Investigación de la Xunta de Galicia (10PXIB106293PR).

Fecha de recepción: 22-01-13 Fecha de revisión: 17-03-13 Fecha de aceptación: 15-05-13

Revista de Psicodidáctica, 2014, 19(1), XX-XX
www.ehu.es/revista-psicodidactica

ISSN:1136-1034 eISSN: 2254-4372
UPV-EHU
DOI: 10.1387/RevPsicodidact.7127

Student, Teacher, and School Context Variables Predicting Academic Achievement in Biology: Analysis from a Multilevel Perspective

José C. Núñez*, Guillermo Vallejo*, Pedro Rosário**, Ellián Tuero*, and Antonio Valle***

*University of Oviedo, Spain, **University of Minho, Portugal, ***University of A Coruña, Spain.

Abstract

The current investigation analyzed how student variables and context variables predicted high school students' academic achievement. The participants were 988 twelfth graders and their corresponding 57 Biology teachers. Data were analyzed using the multilevel method. Results indicate that 85.6% of the variation observed in Biology achievement was explained by variables at the individual level, while the remaining 14.4% was explained by variables at the class level. At the individual level, Biology achievement was associated with approaches to learning, prior knowledge, class absence, and parents' education level. At the class level, academic achievement was only associated with teachers' approaches to teaching, not directly, but through students' approaches to learning.

Keywords: Approaches to teaching, approaches to learning, Biology achievement, multilevel analysis.

Resumen

En el presente estudio se analiza la contribución de variables del alumno y variables del contexto en la predicción del rendimiento académico en Bachillerato. Se han obtenido información de 988 estudiantes, de último curso de Bachillerato y de sus 57 profesores de Biología. Los datos fueron analizados desde una perspectiva multinivel. Los resultados indican que, de la variabilidad observada en el rendimiento en Biología, el 85.6% se debe a las variables de nivel de estudiante mientras que el 14.4% restante corresponde a las variables de nivel de clase. A nivel de estudiante, el rendimiento en Biología se encontró asociado con el enfoque de aprendizaje, con los conocimientos previos, con el absentismo escolar y con el nivel educativo de los padres. A nivel de clase, el rendimiento únicamente estuvo asociado con el enfoque de enseñanza del profesor, y no directamente, sino a través del enfoque de estudio del alumno.

Palabras clave: Enfoques de enseñanza; enfoques de aprendizaje; rendimiento en Biología; análisis multinivel.

Acknowledgements: This work was carried out with funding from the Ministry of Science and Innovation of Spain (Projects: EDU2010-16231 and PSI-2011-23395/PSIC).

Correspondence: José Carlos Núñez, Departament of Psychology; University of Oviedo. Plaza de Feijoo, s/n. 33003 Oviedo. Spain. E-mail: jcarlosn@uniovi.es

Introduction

Consistent with the 2003 and 2006 PISA (OCDE, 2010) reports, in 2009, the students of Portugal and Spain once again obtained results in the Sciences (493 and 488, respectively) that were lower than the OCDE mean (501), suggesting the need to understand these results. After analyzing the impact of social macro-structures and focusing on the teaching-learning process, this report states that economic variables in each country (specifically, the gross domestic product) only explain 6% of the differences in achievement found in the diverse educational systems. This result represents a challenge to investigate the variables that explain the remaining 94% of variance in academic achievement of high school students. The present investigation seeks to deepen our understanding of the conditions that determine academic achievement in high school. We will attempt to respond to this challenge by analyzing the contribution of various theoretically relevant student variables (e.g., approaches to learning, prior achievement, study time, class absence, homework), as well as contextual variables (e.g., approaches to teaching, teacher gender, teacher experience, class size, parents' educational level). As the data are organized in a hierarchical structure (students are nested in classes with their respective teacher), we used a multilevel analysis strategy in this study, which allowed examination of intraclass and interclass effects.

Student variables and academic achievement

Approaches to learning

Three decades ago, Marton and Säljö (1976) described two different approaches employed by students to deal with academic texts. This study initiated an important line of research focused on what was referred to as students' *approaches to learning* (Entwistle, 2009). The authors identified a deep level and a surface level of processing, depending on the approach to learning used by the student to deal with the task. Students who prefer a surface approach are motivated by a goal that is extrinsic to the learning-task; their task involvement is low, and they expend the minimal effort required to complete the task. In contrast, students who prefer the deep approach are motivated by the goal of maximizing their comprehension and constructing meaning by relating the task to their prior knowledge (Entwistle, 2009; Rosário et al., 2010; Rosário, Núñez, Valle, Paiva, & Polydoro, 2013).

Prior knowledge

Students organize knowledge hierarchically, allowing them to understand new experiences. Severe gaps in prior knowledge of a domain can therefore seriously compromise the acquisition of new knowledge (Alexander, Kulikowich, & Schulze, 1994; Miñano & Castejón, 2011). As a consequence, the level of prior knowledge seems to be a relevant variable to include in this study.

Study time

In general, study time is considered a good predictor of school achievement (Plant, Ericsson, Hill, & Asberg, 2005). Nevertheless, for this to occur, study time and

students' corresponding engagement must constantly be adjusted, depending on the students' goals, the nature of the required tasks (e.g., degree of difficulty, perceived utility), and contextual variables (e.g., level of noise, temperature). This may help to explain why the data in the literature do not unequivocally support a direct relationship between study time and school achievement (Gortner-Lahmers & Zulauf, 2000; Núñez, Rosário, Vallejo, & González-Pienda, 2013).

Homework

Despite the long history of research on the role of homework, the strength of the relationship between homework assignment and academic achievement remains inconclusive (Dettmers, Trautwein, & Ludtke, 2009; Rosário et al., 2009; Trautwein & Köller, 2003). Whereas in some studies, a positive relationship was found (e.g., Cooper, Robinson, & Pattal, 2006; Paschal, Weinstein, & Walberg, 1984), others reach less optimistic conclusions, indicating that this relationship is very weak and is mediated by personal, school, and family variables (Ronning, 2011).

Class absence

Finally, in this study, we also considered it important to include the variable class absence, as it has aroused much interest in researchers (Jonanssen, 2011; McIntyre-Bhatty, 2008) due to its association with students' low achievement (Reid, 2006). The study of this variable in connection with other personal or contextual learning variables, for example, those included in this investigation, could provide some clues on how to improve the teaching-learning process.

Contextual variables and academic achievement

Approaches to teaching

Prosser and Trigwell (e.g., Prosser, Trigwell, & Taylor, 1994) developed a line of research about how teachers teach within the context of higher education. Considering the results derived from their investigations, two different ways of coping with the instructional process (*approaches to teaching*) were identified*:* the Information Transmission/Teacher-Focused (ITTF) approach and the Conceptual Change/Student-Focused (CCSF) approach. Whereas teachers who preferentially adopt an ITTF approach focus their activity on the transmission of information related to the learning contents and on technical issues related to the teaching process, teachers who preferentially use a CCSF approach to teaching are committed to promoting students' engagement in an active process of construction of meaning. Accordingly, those teachers who tend to use the CCSF approach take students' prior knowledge into account and develop teaching strategies to promote the construction of knowledge (Ramsden, Prosser, Trigwell, & Martin, 2007). Research on approaches to teaching was oriented towards the analysis of their relation to contextual variables, for example, class size (Lopes & Santos, 2013; Rosário, Nuñez, Valle, et al., 2013; Singer, 1996; Stes, Gijbels, & Van Petegem, 2008), and to teachers' personal variables, including teacher experience (Prosser, Ramsden, Trigwell, & Martin, 2003; Rosário, Núñez, Valle, et al., 2013) and teacher gender (Nevgi, Postareff & Lindblom-Ylänne, 2004; Rosário, Núñez, Ferrando, et al., 2013).

Approaches to teaching and class size

The results of research on the relevance of class size to teachers' adoption of a certain approach to teaching in the university context are inconclusive. For example, whereas a study by Singer (1996) found that as class size increases, teachers are more apt to adopt an ITTF approach to teaching, Stes et al. (2008) found no relationship between the CCSF approach and class size. Globally, the results are controversial, as some studies indicate favorable effects associated with the reduction of class size (Pong & Pallas, 2001; Rosário, Núñez Valle, et al., 2013; Rosário, Núñez, Valle, González-Pienda, & Lourenço, in press) but other studies reach the opposite conclusion (Greenwald, Hedges, & Laine 1996; Konstantopoulos, 2008; Milesi & Gamoran, 2006). As a whole, these results suggest the importance of studying the relationship between the teachers' role in class (e.g., approach to teaching) and class size.

Gender and approaches to teaching

With regard to the relationship between personal teacher variables and preference for a certain approach to teaching, Lacey, Saleh, and Gorman (1998) found a relationship between gender and approach to teaching, and, as in the study by Nevgi et al. (2004), male teachers were more likely to use the ITTF approach to teaching, whereas females were more likely to use the CCSF approach.

Approaches to teaching and teacher experience

Stes et al. (2008) analyzed the relationship between teacher experience and the CCSF approach, hypothesizing that greater teacher experience would be related to a higher probability of using the CCSF approach. The data provided by this study did not confirm this hypothesis, although the authors advised interpreting these results cautiously because the number of subjects was small (50 teachers from a Belgian university). However, Rosário, Núñez, Ferrando, et al. (2013) obtained evidence that more years of experience were related to greater use of teaching oriented to the construction of knowledge (CCSF).

Parents' educational level and academic achievement

According to the data provided by a large number of empirical studies, parents' educational level is an important predictor of students' behavior in class and of academic achievement (Davis-Kean, 2005; Dearing, McCartney, & Taylor, 2001; Duncan & Brooks-Gunn, 1997; Dubow, Boxer, & Huesmann, 2009). For example, Duncan and Brooks-Gunn concluded that mothers' educational level was significantly related to their children's intellectual achievement even after controlling for some socioeconomic indicators such as the family's economic status. Davis-Kean found a positive relationship between parents' educational level and their expectations for their children, suggesting that parents with higher educational levels actively involve their children in the development of ambitious personal expectations.

Goals of the present study

As mentioned above, thus far, the data provided by the research into the role of the aforementioned student and contextual variables in pre-university students' academic achievement are inconclusive (and even more so in the specific area of Biology). Additionally, there is no relevant information available about the variables considered concurrently in the determination of achievement, nor are there any studies of these variables that consider the results at the individual and class level. Therefore, the goal of this investigation is to analyze the degree of association between Biology students' academic achievement and certain student variables (approaches to learning, prior knowledge, study time, degree of class absence, homework), teacher variables (approaches to teaching, teacher gender, teacher experience), class size, and parents' educational level.

As the data provided by past research into many of the variables considered herein are inconclusive and the results provided by the reviewed works have not been analyzed from a multilevel perspective, we propose this study from an exploratory perspective. The data analysis strategy seeks answers to the following questions:

a) Do the explanatory variables measured at the class level in this study affect students' achievement in Biology? If so, then which class level variables are relevant to this conditioning? First, we expect that the approach to teaching will be a relevant variable at the class level, such that students' achievement will be better when the teacher's preferred approach to teaching is student-focused (aimed at the construction of meaning), and students' achievement will be poorer when the teacher's approach to teaching is mainly focused on the transmission of information. Second, with regard to the remaining class level variables, we expect that class size will be negatively related with achievement, whereas teacher experience should be positively associated with achievement in Biology.
b) Do the explanatory variables measured at the individual level affect students' achievement in Biology? As before, if the variability explained at the individual level is significant, the predictive value of each of these variables should be determined. Taking prior studies into consideration, we expect that students' greater use of a deep approach to learning (focused on comprehension and acquisition of competence) will be related to higher achievement in Biology and vice versa: greater use of surface learning (interest in acquiring information and meeting criteria of external achievement) will be related to poorer academic achievement in Biology. Although the results of past research have been inconclusive, we also expect study time, class absence, level of prior knowledge of Biology, homework, and parents' educational level to be positively associated with achievement in this academic area.
c) Is there any interaction between the approach to learning (individual level) and the approach to teaching (class level)? Specifically, does the teacher's approach to teaching moderate the relationship between students' approach to learning and their achievement in Biology?

Method

Participants

Ten high schools situated in the north of Portugal, randomly selected from a total of 45 schools, participated in the study. From these high schools, 57 Biology teachers and their corresponding 988 students in the third year of high school participated. The students presented their parents' authorization to participate in the investigation, and the teachers agreed to participate via e-mail to the main investigator. Of the 988 students, 384 (38.9%) were male, and 604 (61.1%) were female, with ages ranging from 16 to 19 years (M = 17.2, SD = .69). Of the 57 Biology teachers who participated in the investigation, 11 (19.3%) were male, and 46 (80.7%) were female, with ages ranging from 26 to 61 years (M = 46.9, SD = 9.2). Their teaching experience ranged from 2 to 36 years (M = 23.5, SD = 9.6).

Measurement instruments

Student variables

- *Approaches to learning*. The data about approaches to learning were obtained through the Students' Approaches to Learning Inventory (SALI, High School; Rosário, et al., 2007; Rosário, Núñez, Ferrando, et al., 2013). The SALI is made up of 12 items, rated on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The confirmatory factor analyses carried out yielded a two-factor structure of the SALI (Rosário, Núñez, Ferrando et al., 2013): surface and deep approaches, with a good fit of the model, χ^2 (49) = 116.64, $p < .001$, χ^2/df = 2.38, GFI = .98, AGFI = .98, CFI = .99, TLI = .98, RMSEA = .03, CI [.02, .03]. The reliability indexes (Cronbach's alpha) were highly satisfactory: α = .91 for the deep approach, and α = .90 for the surface approach.
- *Class absence.* This variable was assessed at the end of the course by computing the total number of class absences. The information was obtained from the secretariat of the participant high schools (M = 3.18, SD = 4.16).
- *Study time*. Study time was assessed daily for one week with an open question, which asked the students about the number of hours they dedicated to their personal study. All of the students responded by filling in a diary that was returned to the investigators in a sealed envelope at the end of the week. The mean study time obtained was 7.47 weekly hours (SD = 5.52).
- *Prior knowledge of Biology.* This variable was assessed by means of the students' grades in high school. In Portugal, grades range between 0 and 20 points, and 10 is the cut-off point to pass. The student body was distributed as follows: a value of 1 was assigned for grades between 10 and 13 (n = 686; 45.6%); a value of 2 was assigned for grades between 14 and 16 (n = 352; 23.4%); and a value of 3 was assigned for grades between 17 and 20 (n = 466; 31.0%).
- *Homework.* At the end of the course, the teachers gave a 1 to all of the students who had completed at least 80% of the assigned homework (41.4%) and a 2 to those who completed more than 80% (58.6%).

Class variables

- *Approaches to teaching*. The data on approaches to teaching were obtained through the Teachers' Approaches to Teaching Inventory (TATI; Rosário, et al., 2007; Rosário, Núñez, Ferrando, et al., 2013). Based on the theoretical framework associated with the models of Prosser and Trigwell (1999) and Ramsden et al. (2007), this instrument has 12 items that provide information about the two approaches to teaching (ITTF and CCSF). As each approach is made up of one motivation and one strategy, the scale also provides data about the two dimensions of each of the approaches. The scale is rated on a 5-point Likert-type scale, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). By means of a confirmatory factor analysis, we contrasted the theoretical structure of four first-order factors (Motivations and Strategies) and two second-order factors (Approaches). The results showed a good fit of the model: $\chi^2(49) = 101.92$, $p < .001$, $\chi^2/df = 2.08$, GFI = .97, AGFI = .95, CFI = .98, RMSEA = .04, CI [(.03, .05], thereby providing evidence of the construct validity of the inventory (Rosário et al., 2010; Rosário, Núñez, Ferrando et al., 2013). With regard to reliability, both factors had adequate levels: $\alpha_{ITTF} = .92$ and $\alpha_{CCSF} = .94$.
- *Teaching experience*. The data on teaching experience were obtained from the secretariat of the institutes. The mean number of years of experience was 22.81 (*SD* = 9.84).
- *Class size*. The information on class size was obtained from the secretariat of the participant high schools.
- *Parents' educational level.* This variable was categorized as follows: 1 (Elementary school), 2 (Compulsory secondary school), 3 (High school), 4 (Licentiate degree), and 5 (Postgraduate). The information was obtained from the secretariat of the participant high schools.

Academic achievement

In order to study toward a Licentiate degree in the area of Sciences (e.g., Chemistry, Medicine, Biology), Portuguese students must take a national Biology exam. To prepare the students for this exam, the Ministry of Education organizes three tests, one each trimester. In the present investigation, the mean of all three Biology tests was calculated and used as the measurement of academic achievement in Biology.

Procedure

The students and teachers were informed of the goals of this investigation. The information was collected during the second semester of the Biology course (between January and April) after obtaining authorization from the directors of the high schools. Participants were instructed to complete the inventories with reference to the subject of Biology.

Data analysis

The hierarchical nature of the data encouraged analysis with a two-level hierarchical model. The statistical modeling process was carried out in four stages. Initially, a random effect ANOVA model, or unconditional model, was formulated,

which allows determination of the amount of variance explained at the individual level (Level 1) and the class level (Level 2). Additionally, it serves as referent against which to assess the goodness of fit of more complex conditional models. After performing this first step, the model corresponding to class level was fitted in order to determine the extent to which the contextual instructional variables explain students' achievement. Then, the model corresponding to the individual level was fitted in order to observe the extent to which student variables predict academic achievement in Biology. Finally, we studied the interaction of the two models (individual and class level) in order to estimate the degree of interaction among the variables at the class level and the variables at the individual level.

In all the analyses, the dependent variable was the students' grades obtained at the end of the course as predicted by a set of explanatory variables recorded at the individual level and at the class level. The following variables were measured at Level 1: (a) *approaches to learning*, measured with the SALI scale and dichotomized by a cut-off point as a function of the score obtained on this scale. Specifically, if the mean score obtained on the subscales associated with the surface approach (Motivation and Strategy) > 9, then the approach to learning = 0; whereas if the mean score obtained on the subscales associated with the deep approach (Motivation and Strategy) > 9,then the approach to learning = 1; (b) *prior achievement*; (c) the degree of completion of *homework* assigned by the teachers: less than 80% = 0, more than 80% = 1; (d) *student gender* : males = 0, females = 1; (e) *study time* (hours dedicated to study) of the subject over the week: minimum = 0, maximum = 25; (f) *class absences* during the school term: minimum = 0, maximum = 20; (g) *parents*' *educational level*: elementary = 1, ..., postgraduate = 5.

With regard to the explanatory variables recorded at Level 2, we note: (a) the teachers' *approach to teaching,* measured by means of the TATI scale and dichotomized at a cut-off point as a function of the score obtained on the subscales of this scale. Specifically, if the mean score obtained on the subscales associated with teaching focused on transmission of information (Intention and Strategy) > 9, then the approach to teaching = 0; whereas if the mean score on the subscales associated with teaching focused on the construction of knowledge (Intention and Strategy) > 9, then the approach to teaching = 1; (b) *teacher gender*: males = 0, females = 1; (c) *teacher experience*: minimum = 1, maximum =36; (d) *class size*: minimum = 8, maximum = 33.

Results

Descriptive statistics

Table 1 presents the descriptive statistics of the Level 1 and Level 2 variables used in this investigation.

Table 1

Descriptive Statistics of the Variables at the Individual and Class Level

	M	*SD*	Minimum	Maximum
Level 1 Variables (individual)				
Approach to learning	.64	.48	.00	1.00
Prior knowledge	1.85	.85	1.00	3.00
Homework	.61	.49	.00	1.00
Student gender	.61	.48	.00	1.00
Study time	7.79	5.77	.00	25.00
Class absence	3.03	4.19	.00	20.00
Parents' educational level	2.68	1.22	1.00	5.00
Level 2 Variables (class)				
Approach to teaching	.77	.42	.00	1.00
Teacher gender	.80	.40	.00	1.00
Level of teacher experience	23.12	9.99	2.00	36.00
Class size	20.28	4.77	8.00	33.00

Note: Level 1 (N = 988); Level 2 (N = 57).

Multilevel analysis

Unconditional means model

Data analysis began by fitting the following null or unconditional means model:

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

where Y_{ij} is the achievement observed for the *i*th student nested in the *j*th class, γ_{00} is the grand mean (the mean global achievement of the students), u_{0j} represents the variability between classes in terms of students' mean achievement, and e_{ij} represents the variability in the achievement of the students nested in the *j*th class. It is assumed that the random terms of the model are *NID* (normally and independently distributed) with a mean of zero and constant variance; that is to say, $u_{0j} \sim NID(0, \tau_{00})$ and $e_{ij} \sim NID(0, \sigma_e^2)$. Note that it is assumed that the classes studied represent a random sample of a certain population, so that the inferences are not exclusive to the sample of students studied.

With this unconditional model, achievement can be explained through a fixed component containing a global value that is the same for all classes and all students, plus a random component indicating the variability associated with the different levels of the analysis, that is: individual level and class level. This preliminary model served as a referent against which to compare the goodness of fit of successive conditional models. In our case, we verified whether the variance components—one representing the variation between class means (τ_{00}) and the other representing the variation among students within classes (σ_e^2)—were significantly different from zero; if this were not the case, there would be no point in analyzing the data at both levels.

Table 2 shows the results obtained after fitting the model to the data in the present investigation. It can be observed that the estimated mean achievement in this sample of

classes (13.02) is different from zero ($p < .0001$). However, the most notable result is the existence of statistically significant differences in average achievement levels of students among classes ($u_{0j} = 1.61$; $p < .0001$) and in their achievement levels within classes ($e_{ij} = 9.84$; $p < .0001$). In 95% of the cases, the magnitude of the variation among classes in mean achievement levels was expected to fall within the interval [10.45, 15.56]. This indicates a moderate range of variability in average achievement levels among classes in this sample of data. Further, the observed variability in academic achievement (1.62 + 9.84 = 11.46) is mainly due to the Level 1 variables: 85.9% of the variability is due to individual level variables, and the remaining 14.4% is due to class level variables (achievement is higher in some classes than in others).

The degree of dependence among the students' observations within the same class, approximately 0.141 in our case, contradicts the hypothesis of independence assumed by the classic regression model, arguing for data analysis at two levels (individual and class).

Table 2

Summary of the Results Obtained with the Unconditional Means Model

Solution for fixed effects					
Effect	Estimator	Standard error	*df*	*t*-value	Pr > |t|
Intercept	13.0233	.1974	56	65.98	< .0001
Estimators of covariance parameters					
Par Cov	Effect	Estimator	SE	Z-value	Pr > Z
u_{0j}	Classes	1.6100	.4156	3.90	< .0001
e_{ij}	Residual	9.8398	.4560	21.58	< .0001
Fit statistics					
Description	Value				
Deviance	5138.6				
AIC Criterion	5144.2				
BIC Criterion	5150.7				

Note: SE = standard error; df = degrees of freedom; Deviance = minus twice the logarithm of the maximum similarity function; AIC = Akaikes' information criterion; BIC = Bayesian information criterion.

Models with class level predictors

The unconditional means model does not consider either student or class characteristics; it merely provides a basis for comparison against more complex models. However, achievement could be explained by the characteristics of the students who make up the classes, the characteristics of each class, as well as the combined effect of both. We sought to understand why mean achievement is higher in some classes than in others. To explain this, we carried out a new analysis, incorporating the explanatory variables recorded at the class level, Level 2, (the approach to teaching, teacher gender, class size, and teacher experience), paying particular attention to teaching approach.

Specifically, at Level 2, the following conditional model was formulated:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(\text{teaching approach})_j + \gamma_{02}(\text{gender})_j + \gamma_{03}(\text{class size})_j + \gamma_{04}(\text{teacher experience})_j + u_{0j} + e_{ij}.$$

where Y_{ij} represents the achievement observed for the *i*th student nested in the *j*th class, γ_{00} represents the mean achievement of the students instructed by teachers of average experience in average sized groups, γ_{01} indicates whether the achievement of students instructed with methods mainly focused on the teacher (approach to teaching focused on information transmission, ITTF) differs from the achievement of students instructed with methods mainly focused on the student (approach to teaching focused on students' construction of knowledge, CCSF), while controlling for the effects of the variables teacher gender, class size, and teacher experience; γ_{02} indicates whether the achievement of students instructed by women differs from that of students instructed by men, while controlling for the effects of the variables approach to teaching, class size, and teacher experience; γ_{03} represents the change in students' mean achievement for each unit of increase in class size, controlling for the effects of the variables approach to teaching, teacher gender, and teacher experience; γ_{04} represents the change in students' mean achievement as a consequence of the increase in teacher experience, controlling for the effects of the variables approach to teaching, teacher gender, and class size. Finally, u_{0j} represents the variation in class means in academic achievement, and e_{ij} represents the within-class variation.

The results of fitting both conditional random intercept models with Level 2 predictors are shown in Table 3. According to the first of the two fitted models (Model A), there is no evidence of a statistically significant change in the students' mean achievement as a function of the instruction method employed (approach to teaching), teacher gender, class size, or teacher experience. Note that this changes slightly when fitting a more parsimonious conditional model (Model B in Table 2) because the difference between the intercept values of each model is small (13.16 - 12.31= 0.85). However, with the reduced model (Model B), a marginally nonsignificant relationship between the teachers' approach to teaching and the students' mean achievement is observed (γ_{01} = .907, p = .055). But examination of the variance corresponding to Level 2 shows that this relationship does not change significantly when incorporating the variable approach to teaching at class level; specifically, the unconditional variance was 1.61, and the conditional variance was 1.46. This indicates that approximately 10% of the variability observed in mean achievement is explained by the approach to teaching. We also observe that the conditional or residual intraclass correlation only decreases by two hundredths after controlling for the effect of the variable teachers' approach to teaching, dropping from .14 to .12 (1.46/11.31 = .12).

Table 3

Summary of the Results Obtained with the Random Intercept Conditional Model with multiple Level 2 Predictors

	Model A			Model B		
Fixed effects						
Effect	Estimator (SE)	*df*	Pr > |*t*|	Estimator (SE)	*df*	Pr > |*t*|
Intercept	13.162 (1.02)	52	<.0001	12.31(0.41)	55	<.0001
Approach to teaching	.724(0.47)	52	.131	.91(0.46)	55	.054
Teacher gender	-.037(0.49)	52	.940			
Class size	-.052(0.04)	52	.200			
Teacher experience	.015(0.02)	52	.459			
Random effects						
Par Cov	Estimator (SE)	*Z*	Pr > *Z*	Estimator (SE)	*Z*	Pr > *Z*
u_{0k}	1.38(0.37)	3.68	<.0001	1.46(0.39)	3.75	<.0001
e_{ij}	9.85(0.46)	21.56	<.0001	9.84(0.45)	21.57	<.0001
Fit statistics						
Description	Value			Value		
Deviance	5132.9			5134.9		
AIC Criterion	5146.9			5142.9		
BIC Criterion	5162.2			5151.1		

Note. SE = standard error; df = degrees of freedom; Deviance = minus twice the logarithm of the maximum similarity function; AIC = Akaikes' information criterion; BIC = Bayesian information criterion.

Although Model B does not allow us, in statistical terms, to conclude that the teachers' approach to teaching affects student achievement, this variable was not removed from the analysis because it was marginally nonsignificant (p = .055) and central to the present investigation. Moreover, it is noted that Model B, with smaller information criteria—AIC and BIC in our case—is the model that best fits the data. We would have reached the same conclusion if we had used the AIC based on conditional likelihood instead of the AIC based on marginal likelihood; DIC (Deviance Information Criterion) is routinely used for Bayesian model comparison (see Vallejo, Tuero, Núñez, & Rosário, in press).

Models with individual level predictors

The previously fitted model only considers the effect of the variables of class composition and context; it does not consider the students' characteristics. The reasons for the differences in students' achievement is therefore unknown, and there is no evidence that the between-classes variability observed is not an artifact due to the different profiles of the students who are instructed by the teachers in each class. To clarify to this issue, we performed a new analysis with seven individual-level variables: *prior achievement*, *doing homework*, *student gender*, *approach to learning*, *parents' educational level*, *study time,* and *class absence,* with the latter two variables centered around the group mean. Initially, we performed a test to verify the random variation of the slopes one by one, observing that they remained constant except for the slope corresponding to the factor approaches to learning, which varied across classes.

The resulting model of random coefficients can be expressed as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(\text{study time})_{ij} + \gamma_{20}(\text{prior knowledge})_{ij} + \gamma_{30}(\text{homework})_{ij} + \gamma_{40}(\text{absence})_{ij} + \gamma_{50}(\text{gender})_{ij} + \gamma_{60}(\text{parent education})_{ij} + \gamma_{70}(\text{learning approach})_{ij} + u_{0j} + u_{1j}(\text{learning approach})_{ij} + e_{ij}$$

where Y_{ij} represents the achievement observed for the *i*th student nested in the *j*th class, γ_{00} represents the students' mean achievement, γ_{10} represents the change in students' mean achievement for each unit of increase in hours of study time, controlling for the effects of the remaining variables; γ_{20} indicates the relationship between prior knowledge and achievement, controlling for the effects of the remaining variables; γ_{30} indicates the relationship between doing homework and achievement, controlling for the effects of the remaining variables; γ_{40} represents the change in students' mean achievement for each unit of increase in class absence, controlling for the effects of the remaining variables; γ_{50} indicates the relationship between students' gender and their achievement, controlling for the effects of the remaining variables; γ_{60} indicates the relationship between the parents' educational level and their children's achievement, controlling for the effects of the remaining variables; and γ_{70} indicates how the approach to learning affects achievement, controlling for the effects of the remaining variables. Finally, u_{1j} indicates whether the relationship between the approaches to learning and mean achievement varies across classes.

Table 4 shows the most relevant results obtained after fitting both random coefficient models. According to Model A, there is no evidence of changes in mean achievement as a function of hours of study time over the week (p = .074). It is interesting to note that there is a statistically significant relationship between the variables study time and achievement when not controlling for the effect of the variable approaches to learning used by the students; nevertheless, as observed in Table 4, this relationship is marginally nonsignificant when controlling for the effect of approaches to learning. Moreover, there were no statistically significant gender differences in the students' achievement (p = .389). We could not reject the null hypothesis of an absence of association between the variable degree of completing homework assigned by the teachers and the variable achievement (p = .431).

Finally, the results in Table 4 for Model A also show that the relationship between students' approach to learning and mean within-class achievement varied significantly across classes (u_{1j} = .948, p = .015). However, there is no evidence that the effects of approaches to learning on students' academic achievement differ depending upon the average level of academic achievement in the class. In our study, the covariance slope and intercept across classes were not statistically significant (p = .227).

Table 4

Summary of the Results Obtained with the Random Intercept and Slope Models and with Multiple Level 1 Predictors

	Model A			*Model B*		
Fixed effects						
Effect	Estimator (SE)	df	Pr > |*t*|	Estimator (SE)	df	Pr > |*t*|
Intercept	9.846(.477)	56	<.0001	9.779(.446)	56	<.0001
Study time	.029(.020)	924	.0736			
Prior knowledge	.692(.182)	924	.0001	.745(.179)	927	<.0001
Homework	.904(1.147)	924	.4311			
Class absence	-.105(.024)	924	<.0001	-.109(.024)	927	<.0001
Student gender	-.985(1.143)	924	.3891			
Parents' educational level	.356(.086)	924	.0001	.372(.086)	927	<.0001
Approaches to learning	1.746(.297)	924	<.0001	1.821(.258)	927	<.0001
Random effects						
Par Cov	Estimator (SE)	*Z*	Pr > *Z*	Estimator (SE)	*Z*	Pr > *Z*
u_{0j}	1.128(.506)	2.23	.0130	.677(.288)	2.35	0.0094
u_{1j}	1.833(.857)	2.14	.0162	.984(.459)	2.17	0.0151
u_{01}	-.720(.561)	-1.28	.2268			
e_{ij}	8.218(.447)	21.05	<.0001	8.343(.394)	21.19	<.0001
Fit statistics						
Description	Value			Value		
Deviance	4975.5			4980.1		
AIC Criterion	4999.5			4996.1		
BIC Criterion	5024.0			5012.4		

Note. SE = standard error; df = degrees of freedom; Deviance = minus twice the logarithm of the maximum similarity function; AIC = Akaikes' information criterion; BIC = Bayesian information criterion.

The above results indicate the appropriateness of fitting a simpler model, for example, one in which the intercept and slope are allowed to vary across classes, and eliminating the explanatory variables that were nonsignificant in the previous step; in other words, a simpler model, Model B, may provide a reasonable fit to the data. This statement can be easily verified by examining the fit statistics in Table 4; note that we are seeking models with the lowest values in the AIC and BIC criteria. As Model A does not explain the data better than Model B, and Model B is more parsimonious, we chose the latter model.

First, note that after fitting the Model B, there is on average a statistically significant relationship between students' approaches to learning and their academic achievement (γ_{70}= 1.821, p < .0001) within classes. More specifically, taking the direction of the association into account, the results indicate that the achievements of students who usually use a deep approach to learning were significantly higher than those of students who usually use a surface approach. We also verified that prior knowledge positively and significantly predicted present academic achievement (γ_{20}= .745, p < .0001). Further, we found that both class absence and parents' educational level significantly affected the participants' achievement (γ_{40}= -.109, p < .0001 and

γ_{60} = .372, p < .0001, respectively); however, whereas the former reduced achievement, the latter increased it. Finally, the variance components (within and between classes) remain significantly different from zero (e_{ij} = 8.343, p <.0001; u_{0j} = .667, p = .009). It is therefore very important to continue to investigate other causes not taken into account in this analysis that might—at least partially—explain these variabilities. Nevertheless, in this case, not only did the within-class variance decrease from 9.848 to 8.343, but the between-class variance also dropped from 1.385 to .677.

Models with individual and class level predictors

After separately fitting a model for individual level variables and another for class level variables, we will consider a model that includes variables at both levels. This model will allow us to detect the possible existence of crossed interactions between the levels.

Combining the models fitted at the student and at the class level, the following equation is obtained:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(\text{teaching approach})_j + \gamma_{10}(\text{prior knowledge})_{ij} + \gamma_{20}(\text{class absence})_{ij} + \gamma_{30}(\text{parent education})_{ij} + \gamma_{40}(\text{learning approach})_{ij} + \gamma_{11}(\text{teaching approach})_{ij} \times (\text{learning approach})_{ij} + u_{0j} + u_{1j}(\text{learning approach})_{ij} + e_{ij}$$

This reveals that achievement can be considered as a function of fixed effects plus random effects. The fixed effects are: general mean (γ_{00}), main effect of the approach to teaching (γ_{01}), main effect of prior achievement (γ_{10}), main effect of class absence (γ_{20}), main effect of parents' educational level (γ_{30}), main effect of approach to learning (γ_{40}), and crossed interaction between approaches to teaching and approaches to learning (γ_{11}). The random effects represent the between-class variability (u_{0j}) among the approaches to learning across the classes (u_{1j}) and within classes (e_{ij}). As all of the issues that motivated this analysis have been specified except for the one referring to crossed interaction, we estimated γ_{11} to examine whether the teacher-focused approach to teaching (information transmission) differs from the student-focused approach to teaching (facilitating students' construction of knowledge) in terms of the strength of the association between approaches to teaching and students' academic achievement.

Table 5 presents the most important results obtained after fitting the model that includes Level 1 and Level 2 predictors. Specifically, we verified that the teachers' approach to teaching had no statistically significant main effects (γ_{01} = .673, p = .125), although it had a secondary effect through its interaction with students' approaches to learning (γ_{11} = -1.403, p = .018). Nevertheless, the achievement of students instructed by teachers preferentially using a student-focused approach to teaching was slightly better (10.10) than that of students instructed by teachers preferentially using a teacher-focused approach to teaching (9.42). With regard to the interaction, the strength of the association between the students' approaches to learning and their achievement in Biology varied depending on whether the teachers regularly used a student-focused

approach or a teacher-focused approach to teaching. In other words, due to the moderating effect of the teachers' approach to teaching, the differences in the achievement of students who preferentially used a deep approach were higher than that of students who used a surface approach when the teachers preferentially used a student-focused approach instead of a teacher-focused approach.

Finally, the components of intercept and slope variance remained statistically significant (p = .011 and p = .022, respectively), indicating a significant between-class variation in both coefficients. The addition of the variable approach to teaching and its crossed interaction with approach to learning slightly reduced the residual variance of the intercept (≈ 3%) and the residual variance of the slope for approaches to learning (≈ 13%) in comparison with the estimated variance for the random coefficient model of the previous section. Nevertheless, rejection of the null hypothesis would indicate that there is additional variation in class mean achievement levels that is not explained by the variables included in the model. It is foreseeable that the inclusion of additional class-level variables would further reduce the variance corresponding to the classes. Therefore, additional student and teacher characteristics not taken into account in this analysis might explain this variation.

Table 5

Summary of the Results Obtained with the Random Intercept Combined Model

Fixed effects

Effect	Estimator	(SE)	*Df*	*t*-Value	Pr > |*t*|
Intercept	9.424	(.516)	55	18.28	<.0001
Approach to teaching	.673	(.432)	55	1.56	.1248
Prior knowledge	.694	(.178)	926	3.90	<.0001
Class absence	-.108	(.024)	926	-4.49	<.0001
Parents' educational level	.374	(.086)	926	4.36	<.0001
Approach to learning	2.884	(.526)	926	5.49	<.0001
Approach to teaching × Approach to learning	-1.403	(.594)	926	-2.37	.0184

Random effects

Par Cov	Effect	Estimator	SE	*Z-value*	Pr > *Z*
u_{0j}	Classes	.658	.267	2.30	.0108
u_{1j}	Approach to learning	.867	.431	2.01	.0222
e_{ij}	Residual	8.319	.392	21.25	<.0001

Fit statistics

Description	Value
Deviance	4974.3
AIC Criterion	4994.3
BIC Criterion	5014.8

Note. SE = standard error; df = degrees of freedom; Deviance = minus twice the logarithm of the maximum similarity function; AIC = Akaikes' information criterion; BIC = Bayesian information criterion.

Discussion

The goal of this investigation was to analyze the degree to which academic achievement in Biology of students in the last year of high school is predicted by certain student variables (i.e., approaches to learning, prior knowledge, study time, class absence, homework), teacher variables (i.e., approaches to teaching, gender, teacher experience), and contextual variables (i.e., class size, parents' educational level). As the data show a hierarchical structure (students nested within classes), a multilevel strategy was conducted. By means of this type of analysis, this study not only allowed us to determine the relevance of individual-level and class-level variables in the prediction of achievement in Biology, but also to study the interaction of the variables of both levels, an aspect that has received little attention in past research but is of great theoretical and applied importance.

In general, whereas the hypotheses formulated at the class level were mainly not confirmed, the hypotheses at the individual level were confirmed to a great extent. Thus, we confirmed that most of the variability in Biology achievement was associated with individual-level variables (85.6%), whereas the class-level variables only explained 14.4% of the variability. However, the data corresponding to the interaction between the teachers' way of teaching and students' way of learning and academic achievement were particularly relevant. Below, the most important findings are discussed.

Individual-level analysis

With regard to the variables analyzed at individual level (Level 1), prior knowledge of the subject, class absence, parents' educational level, and approach to learning were good predictors of achievement in Biology, and the last variable was the most relevant in this equation. Study time, the amount of homework done, and student gender had no significant main effects.

With regard to the significant effects, as expected, a higher level of prior knowledge was related to better achievement in Biology. Likewise, we observed that greater class absence was related to poorer academic achievement (Reid, 2006). In keeping with some previous research, in this study, we found that parents' higher educational level was associated with their children's higher Biology achievement (Davis-Kean, 2005; Dubow et al., 2009). Lastly, this investigation provides clear evidence that greater use of a deep approach to study leads to higher achievement and that use of a more superficial approach to learning is related to poorer achievement in Biology. Although some works have expressed doubts about this relationship (Entwistle, 1991; Rosário et al., 2010; Struyven, Dochy, Janssens, & Gielen, 2006), the data presented in the current study clearly show that the benefits come from the use of a deep approach, implying intrinsic or task-oriented motivation and the use of cognitive and metacognitive strategies to comprehend and elaborate the information.

With regard to the nonsignificant variables in the explanation of achievement in Biology (study time, student gender, and amount of homework done), the students study time on Biology deserves special mention. Specifically, although this variable was not relevant when all of the student variables were included in the equation, its main effect becomes significant if the variables that were significant (prior knowledge, class absence, parents' educational level, and approaches to study) are eliminated from the

equation. Thus, study time is an important variable, but when including other variables such as approaches to learning the effect of study time occurs through the latter (in fact, studying with a deep approach to learning involves more study time than studying with a surface approach). With regard to the other two variables, our data indicate that doing more or less homework does not explain a significant amount of the variability in achievement. How can we explain these data? On the one hand, error estimation was high (1.137), perhaps due to the dichotomization of the homework variable, (which also occurs with the error estimation of gender, 1.134). On the other hand, as in the case of study time, the effect of the amount of homework done could also be subsumed by the students' approach to learning (it is possible that doing homework with a deep approach could involve completing more homework tasks and spending more time compared to completing homework using a surface approach). Therefore, like study time, the amount of homework may play a more important role than the one suggested by the results of the analysis when all of the variables are present. Future research should analyze this hypothesis in depth (measuring homework as a continuous variable) while considering it as a class-level variable.

Class-level analysis

None of the variables included in the equation at class level showed significant main effects. Only the approaches to teaching showed a mild main effect on achievement in Biology at this level of analysis ($p < .10$), although this limited effect dissipated when the approach to teaching was related to the approach to studying.

Interaction between approaches to teaching and approaches to learning

This study provides relevant and novel information about the interaction between students' approaches to learning (Level 1) and teachers' approaches teaching (Level 2). As mentioned, the results at the individual level indicated that students who preferentially use a deep approach to studying perform better, and students who are more likely to use a surface approach show poorer achievement. When taking into account both levels of analysis, we confirmed that this difference in achievement was greater in the students instructed by teachers whose approach to teaching was mainly focused on transmitting information (ITTF) than in the students whose teachers used an approach to teaching preferentially oriented to helping the student to construct meaning (development of processes of comprehension and elaboration of the information, CCSF). What could be the reason for this finding? Students' learning and achievement may be significantly more determined by their personal characteristics considered herein (e.g., study time, homework completion) and by other characteristics not considered (e.g., students self-set goals, attitude towards learning) than by their teachers' ITTF approaches to teaching.

On the other hand, when a teacher promotes instructional contexts that demand students' active and significant involvement in the construction of knowledge (CCSF), students who are more likely to use a surface approach to learning will be encouraged to use a deeper approach to learning because a surface approach will not be useful in this teaching context.

Limitations of the study

The present investigation has involved a great effort to collect sufficient data from students, parents, and teachers in order to carry out the analysis from a multilevel perspective. However, there are some aspects of the study that could modulate the interpretation of the results obtained. First, the fact that information about approaches to learning and teaching was obtained by means of self-report instruments means that such information is based on what the students and teachers think they do in their respective tasks. Although the use of self-report measures is very common in research in the field of education, it is still a limitation because the results should be interpreted as what teachers and students think they do and not what really occurs. Second, the conclusions derived from this study may not be completely transferable to other academic disciplines or to students of other ages (Stes et al., 2008). It would therefore be interesting for future research to explore the many questions that persist in this area.

References

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences, 6,* 379-397. doi: 10.1016/1041-6080(94)90001-9.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research, 76,* 1-62. http://dx.doi.org/10.3102/00346543076001001.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*, 294-304. http://dx.doi.org/10.1037/0893-3200.19.2.294.

Dearing E., McCartney K., & Taylor, B.A. (2001). Change in family income matters more for children with less. *Child Development*, *72*, 1779-1793. http://dx.doi.org/10.1111/1467-8624.00378.

Dettmers, S., Trautwein, U., & Ludtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement, 20,* 375-405. http://dx.doi.org/10.1080/09243450902904601.

Dubow, E., Boxer, P., & Huesmann, L. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill Palmer Quarterly*, *55*, 224-249. http://dx.doi.org/10.1353/mpq.0.0030.

Duncan, G. J., & Brooks-Gunn, J. (1997). *Consequences of growing up poor*. New York: Russell Sage Foundation.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education, 22*, 201-204. http://dx.doi.org/10.1007/BF00132287.

Entwistle, N. J. (2009). *Teaching for understanding at University: Deep approaches and distinctive ways of thinking*. Basingstoke, UK: Palgrave Macmillan.

Gortner-Lahmers, A., & Zulauf, C. R. (2000). Factors associated with academic time use and academic performance of college students: A recursive approach. *Journal of College Student Development, 41*, 544-556.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effects of school resources on student achievement. *Review of Educational Research*, *66*, 361-396. http://dx.doi.org/10.2307/1170528.

Jonanssen, C. (2011). The dynamics of absence behaviour: Interrelations between absence from class and absence in class. *Educational Research*, *53*, 17-32.

Konstantopoulos, S. (2008). Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *Elementary School Journal*, *108*, 275-291. http://dx.doi.org/10.1086/528972.

Lacey, C. H., Saleh, A., & Gorman, R. (1998, October). *Teaching nine to five: A study of the teaching styles of male and female professors*. Paper presented at the annual meeting of the Women in Educational Leadership Conference, Lincoln, NE.

Lopes, J., & Santos, M. (2013).Teachers' beliefs, teachers' goals and teachers' classroom management: A study with primary teachers. *Revista de Psicodidáctica, 18*, 5-24. doi: 10.1387/RevPsicodidact.4615.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I – Outcome and process. *British Journal of Educational Psychology, 46*, 4-11. http://dx.doi.org/10.1111/j.2044-8279.1976.tb02980.x.

McIntyre-Bhatty, K. (2008). Truancy and coercive consent: Is there an alternative? *Educational Review*, *60*, 375-390. http://dx.doi.org/10.1080/00131910802393407.

Milesi, C., & Gamoran, A. (2006). Effects of class size and instruction on kindergarten achievement. *Educational Evaluation and Policy Analysis*, *28*, 287-313. http://dx.doi.org/10.3102/01623737028004287.

Miñano, P., & Castejón, J. L. (2011). Cognitive and motivational variables in the academic achievement in language and mathematics subjects: A structural model. *Revista de Psicodidáctica, 16*, 203-230.

Nevgi, A., Postareff, L., & Lindblom-Ylänne, S. (2004, June). *The effect of discipline on motivational and self-efficacy beliefs and on approaches to teaching of Finnish and English university teachers*. A paper presented at the EARLI SIG Higher Education Conference.

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. http://dx.doi.org/10.1016/j.cedpsych.2012.10.002.

OECD (2010). *PISA 2009 Results: What students know and can do: Student performance in reading, mathematics and science* (Volume I). Retrieved at http://dx.doi.org/10.1787/9789264091450-en.

Paschal, R. A., Weinstein, T., & Walberg, H. J. (1984). The effects of homework on learning: A quantitative synthesis. *Journal of Educational Research*, *78*, 97-104.

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology, 30,* 96-116. http://dx.doi.org/10.1016/j.cedpsych.2004.06.001.

Pong, S., & Pallas, A. (2001). Class size and eighth-grade math achievement in the United States and abroad. *Educational Evaluation and Policy Analysis*, *23*, 251-273. http://dx.doi.org/10.3102/01623737023003251.

Prosser, M., & Trigwell, K. (1999).*Understanding learning and teaching. The experience in higher education*. Buckingham, UK: Open University Press.

Prosser, M., Ramsden, P., Trigwell, K., & Martin, E. (2003). Dissonance in experience of teaching and its relation to the quality of student learning. *Studies in Higher Education*, *28*, 37-48. http://dx.doi.org/10.1080/03075070309299.

Prosser, M., Trigwell, K., & Taylor, P. (1994). A phenomenographic study of academics' conceptions of science learning and teaching. *Learning and Instruction, 4*, 217-231. http://dx.doi.org/10.1016/0959-4752(94)90024-8.

Ramsden, P., Prosser, M., Trigwell, K., & Martin, E. (2007). University teachers' experiences of academic leadership and their approaches to teaching. *Learning and Instruction, 17,* 140-155. http://dx.doi.org/10.1016/j.learninstruc.2007.01.004.

Reid, K. (2006). An evaluation of the views of secondary staff towards school attendance issues. *Oxford Review of Education, 32*, 303-324. http://dx.doi.org/10.1080/03054980600775557.

Ronning, M. (2011). Who benefits from homework assignments? *Economics of Education Review, 30,* 55-64.

Rosário, P., Mourão, R., Baldaque, M., Nunes, T., Núñez, J. C., González-Pienda, J., & Valle, A. (2009). Tareas para casa, autorregulación del aprendizaje y rendimiento en matemáticas [Homework, self-regulation of learning, and math achievement]. *Revista de Psicodidáctica, 14*, 179-192.

Rosário, P., Mourão, R., Núñez, J. C., González-Pienda, J. A., Solano, P., & Valle, A. (2007). Evaluating the efficacy of a program to enhance college students' self-regulation learning processes and learning strategies. *Psicothema, 19*, 353-358.

Rosário, P., Núñez, J. C., González-Pienda, J. A., Valle, A., Trigo, L. & Guimarães, C. (2010). Enhancing self-regulation and approaches to learning in first-year college students: A narrative-based program assessed in the Iberian Peninsula. *European Journal of Psychology of Education*, *25*, 411-428. http://dx.doi.org/10.1007/s10212-010-0020-y.

Rosário, P., Núñez, J. C., Valle, A., Paiva, O., & Polydoro, S. (2013). Approaches to teaching in High School when considering contextual variables and teacher variables. *Revista de Psicodidáctica, 18,* 25-45. doi: 10.1387/RevPsicodidact.6215.

Rosário, P., Núñez, J. C., Valle, A., González-Pienda, J. A., & Lourenço, A. (in press). Grade level, study time, and grade retention and their effects on motivation, self-regulated learning strategies, and mathematics achievement: A structural equation model. *European Journal of Psychology of Education.* doi: 10.1007/s10212-012-0167-9.

Rosário, P., Núñez, J. A., Ferrando, J. P., Paiva, O., Lourenço, A., Cerezo, R., & Valle, A. (2013). The relationship between approaches to teaching and approaches to studying: A two-level structural equation model for biology achievement in high school. *Metacognition and Learning, 8,* 47-77. doi: 10.1007/s11409-013-9095-6.

Singer, E. (1996). Espoused teaching paradigms of college faculty. *Research in Higher Education*, *37*, 659-679. http://dx.doi.org/10.1007/BF01792951.

Stes, A., Gijbels, D., & Van Petegem, P. (2008). Student-focused approaches to teaching in relation to context and teacher characteristics. *Higher Education*, *55*, 255-267. http://dx.doi.org/10.1007/s10734-007-9053-9.

Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning

environment. *Learning and Instruction, 16*, 279-294. http://dx.doi.org/10.1016/j.learninstruc.2006.07.001.

Trautwein, U., & Köller, O. (2003). The relationship between homework and achievement—still much of a mystery. *Educational Psychology Review*, *15*, 116-145. http://dx.doi.org/10.1023/A:1023460414243.

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*.

José Carlos Núñez is full Professor of Learning Difficulties at the University of Oviedo (Spain) and Director of the Department of Psychology. His main lines of research are: a) psychological and educational dimensions of self-regulated learning in educational settings and b) academic learning difficulties and ADHD. He is currently in charge of a project of the National Research Plan (EDU2010-16231).

Guillermo Vallejo is Full Professor of Methodology of Behavioral Sciences at the University of Oviedo (Spain) where he teaches Research Designs as part of the Faculty of Psychology. He is the Principal Investigator of a project of the National Research Plan (PSI-2011-23395/PSIC).

Pedro Rosário is Associate Professor of Educational Psychology at the University of Minho (Portugal). His main lines of research are: (a) psychological and educational dimensions of self-regulated learning, (b) processes of self-regulation in technological settings, and (c) interactive blackboard learning environments. He has numerous publications in his country and abroad in both lines of research. He is the Principal Investigator of the GUIA group (www.guia-psiedu.com).

Ellián Tuero is Scholarship holder for training research personnel (FPI) in the Department of Psychology of the University of Oviedo (Spain). She teaches at the Faculty of Psychology and the Faculty of Training Teachers and Education. At the research level, she specializes in data analysis of repeated measures designs and in structures that require strategies of multilevel analysis.

Antonio Valle is Full Professor of Educational Psychology and Director of the Department of Developmental and Educational Psychology of the University of A Coruña (Spain). His main research topics are academic motivation, study strategies, and self-regulated learning. He is currently in charge of a Research Project of the Xunta of Galicia (10PXIB106293PR).

Received date: 22-01-13 Review date: 17-03-13 Accepted date: 15-05-13

Evaluación de la Investigación

Todo investigador debe
pensar después cómo ha realizado su estudio

DISCUSIÓN GENERAL

Las exigencias del desarrollo social actual justifican la realización de investigaciones complejas en las que intervienen diversos métodos y técnicas de investigación con las que se obtienen múltiples datos de variables que pertenecen a diversos niveles de jerarquía y que es necesario analizar concienzudamente. Es ahí donde los Modelos Multinivel, lejos de revelarse como una moda, son el método inexcusable de análisis para lograr resultados integrales en la contemporánea realidad educativa.

Los Modelos Multinivel tienen una gran complejidad en lo que respecta al marco teórico que los evidencia y los sustenta, y también suponen un instrumento complejo en cuanto a su utilización para analizar los datos. En esta parte, el investigador no sólo debe tener claros sus objetivos y sus hipótesis sino también un buen dominio de la técnica para no errar en el análisis.

La meta que persigue todo investigador educativo es encontrar el modelo que explique la máxima variabilidad de sus datos con el mínimo número de parámetros. A esto se le denomina encontrar el modelo "*óptimo*" para sus datos. Así las cosas, y dado que no tener una teoría bien definida que explique nuestras hipótesis es más frecuente de lo que sería deseable, junto con que para una misma evidencia muestral es posible que existan múltiples modelos candidatos (García, 1996; Vallejo et al., 2010, 2011b); conocer qué estrategia es la más adecuada para seleccionar el modelo *óptimo* multinivel se torna prioritario, dado que cuando se elige un modelo inapropiado la precisión y la exactitud de la estimación de los parámetros empeora conduciendo a conclusiones erróneas sobre los resultados.

Para evaluar qué modelo explica mejor unos datos determinados la literatura estadística especializada nos ofrece diversos criterios de ajuste. Entre los más comúnmente utilizados destacan el criterio de ajuste condicional LRT y los Criterios de Información, siendo estos últimos los que más aceptación están teniendo por su versatilidad para comparar entre modelos anidados y no anidados y por estar implementados en la mayor parte de los programas estadísticos que ajustan modelos

mixtos, incluyendo el módulo *ProcMixed* en SAS, la función *lme* en S-PLUS/R o los comandos *Mixed y xtmixed* en SPSS y STATA. Otros criterios de selección de modelos, tales como los basados en la validación cruzada, los criterios predictivos (el coeficiente de determinación ajustado, el coeficiente de correlación de concordancia ponderado, la suma de cuadrados residual de predicción, etc.) o las técnicas gráficas carecen del vigor de los criterios anteriores, sobre todo, estas últimas. Es muy probable que la naturaleza subjetiva de las técnicas gráficas influya negativamente en su escasa aplicación en el ámbito de la selección de modelos.

El objetivo central de esta Tesis se situó en este contexto y consistió en evaluar el desempeño de una amplia variedad de estrategias de bondad de ajuste para elegir el modelo *óptimo* (el modelo más simple con la menor discrepancia entre los valores observados y los valores esperados) bajo diferentes condiciones de estudio. Específicamente, llevamos a cabo 4 investigaciones donde pusimos a prueba el rendimiento de diversos Criterios de Información que, en su mayor parte, están implementados en el módulo PROC MIXED y PROC GLIMMIX del programa SAS, en diseños tanto de corte longitudinal como de corte transversal.

En el primer y segundo trabajo, los estudios estuvieron enmarcados en el ámbito longitudinal. El primero versó sobre modelos anidados pertenecientes a diseños generados con datos balanceados (datos completos). Sin embargo, no se puede obviar el hecho de que en muchas ocasiones en las que se toman distintas medidas a lo largo del tiempo pueden provocarse abandonos, con los consiguientes datos faltantes (datos incompletos), en este contexto se encuadró la segunda investigación. En ambas, mediante simulación Monte Carlo comparamos globalmente el rendimiento de los criterios de ajuste para seleccionar el mejor modelo y el impacto del uso de los diferentes estimadores de los Criterios de Información.

En general, encontramos que ninguna de las herramientas exploradas eligió, de manera consistente, el modelo correcto. No obstante, el rendimiento de estos procedimientos de bondad de ajuste mejoró notablemente cuando aumentamos el número de medidas repetidas y el tamaño de la muestra.

Constatamos también que los Criterios de Información Eficientes (AIC y AICC) funcionaban mejor en la selección del modelo cuando los patrones de covarianza generados fueron complejos y peor cuando fueron simples. Al contrario obtuvimos para sus homólogos Consistentes (BIC, CAIC y HQIC), que rindieron mejor cuando los

patrones de covarianza fueron simples y peor cuando fueron complejos. El mismo comportamiento que obtuvimos para los criterios AIC y BIC fue encontrado en estudios anteriores por distintos investigadores (Gómez, Schaalje, & Felligham, 2005; Keselman, Algina, Kowalchuck, & Wolfinger, 1998).

Nuestros datos también pusieron de manifiesto que, independientemente del procedimiento de estimación utilizado, ML o REML, los Criterios de Información Consistentes basados en el número de individuos (*n*, nivel 2) fueron más efectivos en la selección del mejor modelo que cuando estos mismos Criterios Consistentes estaban basados en el número total de observaciones (*m*, nivel 1), específicamente para el BIC y el CAIC. Estos resultados generalizan los encontrados por Gurka (2006), lo que sugiere la utilización de *n* en el término de penalización del BIC y del CAIC (posturas afines se hallan en Carlin & Louis, 2001; Kass & Raftery, 1995). Además, este hallazgo sirve de soporte empírico para el uso de la estrategia del ProcMixed del SAS (2008), de calcular los Criterios Consistentes usando el número de participantes en el nivel 2 del modelo jerárquico, en contraposición a la seguida en el comando Mixed del SPSS (2008).

De modo particular, en nuestro primer estudio los resultados indicaron que de todas las herramientas examinadas para seleccionar el mejor modelo de un conjunto de modelos anidados (siendo conocido el verdadero proceso generador de los datos VPGD) tuvo un mejor desempeño el criterio de ajuste condicional LRT basado en el estimador de máxima verosimilitud completa MV. Pese a que el desempeño de los Criterios de Información fue inferior que el rendimiento del LRT, el rendimiento del Criterio AIC fue superior al de todos los demás Criterios de Información.

En líneas generales, los Criterios de Información (AIC, AICC, CAIC, BIC, CAIC y HQIC) seleccionaron el VPGD más veces bajo estimación REML que bajo estimación ML, coincidiendo con el estudio de Gurka (2006). Nuestros resultados también mostraron que el rendimiento de los Criterios de Información mejoró cuando el estimador REML incluyó el término constante (REML1), especialmente cuando el número de medidas repetidas tomado fue moderado, en concordancia con los resultados encontrados por Wang y Schaalje en su estudio de 2009.

En nuestro segundo estudio, cuando la selección del modelo se centró en modelos no anidados (de nuevo conocido el VPGD) tuvieron un rendimiento superior tanto el Criterio AIC como el Criterio HQIC. De hecho, el rendimiento de este último

era previsible dadas otras investigaciones con estudios de modelos finitos (Burham & Anderson, 2002).

Desde un punto de vista cuantitativo nuestros resultados indicaron que el porcentaje de veces que el modelo similar al VPGD fue elegido por los Criterios de Información fue inferior al encontrado en el estudio de Gurka (2006). Esto se justifica por el propio escenario examinado en dicho trabajo. Gurka (2006) realizó una investigación más parca en la que los datos considerados eran completos y la comparación de los modelos era restringida (6 modelos candidatos para cada conjunto de datos generados). En nuestro caso, el escenario fue muchísimo más complejo, generándose datos faltantes y comparándose 36 modelos candidatos para cada conjunto de datos.

Los resultados de simulación tratados en este segundo estudio también pusieron de manifiesto que los Criterios de Información rindieron mejor cuando los datos fueron generados a partir de distribuciones normales. De hecho, ninguno de los procedimientos considerados tuvo un buen desempeño cuando los datos se obtuvieron bajo distribuciones segadas. Este hallazgo refuerza la importancia de usar pruebas diseñadas para comprobar si se satisfacen los supuestos distribucionales subyacentes al modelo de análisis (Vallejo, Ato, & Fernández, 2010; Sterba & Pek, 2012).

La selección de modelos mediante Criterios de Información en el contexto transversal también fue objeto de estudio en nuestra tercera investigación. Así, tratamos de encontrar la estrategia más óptima de selección del mejor modelo multinivel entre distintas alternativas candidatas. Para ello se tuvieron en cuenta el tamaño de grupo, el número de grupos, el valor de los parámetros y la correlación intra-clase. Todas estas cuestiones fueron examinadas mediante un estudio de simulación Monte Carlo, tanto desde un punto de vista más Clásico como desde un punto de vista Bayesiano. Con fines de comparación, además de evaluar el rendimiento de los Criterios de Información utilizados en los dos primeros estudios (AIC, AICC, BIC, CAIC y HQIC), también fue examinado el desempeño del Criterio de Información de la Desvianza (DIC) bayesiano sugerido por Spiegelhalter, Best, Carlin y Van der Linde (2002) y el AIC condicional (cAIC) recomendado por Vaida y Blanchard (2005).

En general, los resultados fueron coincidentes con las investigaciones previas y ninguno de los criterios de selección se comportó correctamente en todas las condiciones examinadas. Si bien es cierto que en la mayoría de las veces el rendimiento de los

Criterios de Información fue mejor bajo estimación REML que bajo ML, confirmándose de nuevo lo encontrado por Gurka (2006).

En la selección del mejor modelo multinivel, en relación al tamaño de muestra, nuestros resultados revelaron que un gran número de grupos (NG) es más importante que un gran tamaño de grupos (TG). Este hallazgo sugiere que para distinguir entre modelos multinivel que están compitiendo la regla debe ser: NG ≥ 50 y N/NG ≥ 20, siendo N = NG × TG).

En cuanto a la importancia de la correlación intra-clase, los datos indicaron que esta variable afectó al rendimiento de los Criterios de Información, pero la magnitud de esa influencia fue menor en comparación con los valores de los parámetros y la correlación de los efectos aleatorios. Esta última sí que resultó tener un peso relevante. Así, los Criterios de Información Eficientes tuvieron un rendimiento superior cuando los efectos aleatorios fueron correlacionados, mientras que los Criterios Consistentes fueron más ventajosos cuando los efectos aleatorios no fueron correlacionados. Resultados similares fueron encontrados por Vallejo et al. (2010, 2011b), donde los criterios tipo AIC rindieron mejor que los criterios tipo BIC cuando los patrones de covarianza (efectos aleatorios) para generar los datos fueron complejos y peor cuando fueron simples y viceversa.

Respecto a las discrepancias en las fórmulas que involucran en el término de penalización los criterios, al menos para el BIC y CAIC, en la corrección se sugiere utilizar m en vez de N. Como se ha indicado anteriormente, el tamaño de la muestra en el SAS al calcular dichos criterios es igual a m, mientras que en el SPSS es igual a N tanto bajo estimación ML como REML.

Una vez estudiados en profundidad los criterios de selección de modelos y teniendo en consideración que los resultados obtenidos en nuestras investigaciones mediante datos generados por simulación concordaron con los encontrados en otros estudios, estimamos oportuno poner a prueba una aplicación práctica. De esta forma, estudiando datos reales (de alumnos agrupados en clases) constataríamos la importancia de utilizar la técnica multinivel para analizar los datos que se presentan jerárquicamente en el ámbito educativo. Así las cosas, analizamos en qué medida el rendimiento académico en Biología, de los estudiantes de último curso de Bachillerato, fue predicho por variables del alumno, del profesor y del contexto.

En general, nuestros resultados apuntaron que la mayor parte de la variabilidad en el rendimiento de los estudiantes, en la asignatura estudiada, estuvo asociada con variables del nivel estudiante en el 85,6 % y con variables de clase en el 14,4 %. A nivel estudiante (nivel 1) resultaron importantes: los conocimientos previos, el enfoque de aprendizaje, el absentismo escolar y el nivel educativo de los padres. A nivel de clase, el rendimiento solamente estuvo asociado directamente con el enfoque de enseñanza del docente, y no directamente, sino a través del enfoque de estudio del estudiante.

Con este último trabajo pretendimos arrojar un poco más de luz a la investigación educativa aportando un estudio donde se pusiera de manifiesto la importancia de distintas variables tomadas en la determinación del rendimiento cuando se consideran conjuntamente los resultados a nivel estudiante y a nivel de clase. Además, una de las contribuciones más importantes de este estudio fue la aplicación del conocimiento que ofrecen los Modelos Multinivel a la evaluación de los estudios en el Campo de la Educación. Así, aportamos conocimiento de esta técnica de análisis en dos sentidos. De una parte, conceptualizamos la técnica describiéndola de manera rigurosa y detallada. Y de otra, explicamos el proceso de modelado estadístico multinivel de forma precisa. Ambas partes se tornaron como una contribución pedagógica indispensable para las exigencias actuales en la investigación educativa.

Finalmente hay que tener presente que nuestras aportaciones han sido relevantes para el estudio del comportamiento de los estimadores en muestras tanto infinitas como finitas para variables dependientes continuas. Sin embargo, nuestros resultados no pueden ser extendidos a otras situaciones. Las generalizaciones no están exentas de problemas y por ejemplo, la utilización de Modelos Multinivel con variables discretas ha sido menos investigada y requiere especial atención a los supuestos teóricos, por lo que su estudio en profundidad queda aún pendiente. Otra posible línea de investigación relevante puede ser similar a la seguida aquí, con la salvedad ahora de que el objetivo principal sería comparar el desempeño de los Criterios de Información para seleccionar el modelo más próximo al VPGD. En las 3 investigaciones realizadas mediante métodos de simulación Monte Carlo se evaluó el rendimiento de los criterios de bondad de ajuste para seleccionar el modelo más óptimo pero siempre asumiendo conocido el VPGD, sin embargo en la práctica, cuando se trabaja con datos reales, es más probable que el VPGD sea desconocido, motivo por el cual sería altamente recomendable un estudio así. De igual modo, la escasez de investigaciones con un número amplio de unidades del segundo

nivel y un mayor número de niveles en los estudios sería muy reveladora, otra cosa distinta es que sea viable. En este sentido hay que efectuar una consideración a tener en cuenta. Los Modelos Multinivel requieren, como ha quedado patente en nuestros estudios, del uso de muestras grandes, de las que dependen en gran medida las buenas propiedades de los estimadores. Cuando se realizan investigaciones en el campo educativo a nivel macro, como por ejemplo los estudios PISA que analizan el rendimiento de los estudiantes a partir de unos exámenes mundiales, no parece existir inconveniente alguno en conseguir una muestra grande de participantes, por lo que este inconveniente quedaría solventado. Sí que podrían aparecer otros problemas como los de la validez de esos datos, pero eso no nos compete aquí. Ahora bien, la mayoría de las investigaciones que se realizan sobre la Eficacia Escolar promovidas desde las Universidades, Colegios, Instituciones... son menos costosas, menos pretenciosas pero más habituales, y en estos estudios el investigador juega con la ventaja del control de los datos pero se enfrenta a la dificultad de conseguir un tamaño de muestra grande junto con un número de grupos también grande. Por tanto, como esto es muy complicado de resolver, la solución pasaría por continuar investigando en los criterios de bondad de ajuste de los Modelos Multinivel con el fin de lograr modelos más acomodados a los datos sin necesidad de muestras tan grandes.

CONCLUSIONS

The main conclusions of this Thesis can be summarized as follow:

1. In order to provide results that contribute to furthering our knowledge of the variables affecting the improvement of School Effectiveness, the use of Multilevel Models is required.

2. Each hierarchy level is modeled by Multilevel Models, which combine and analyze the relations among levels, estimate the variability of the coefficients among the groups and the magnitude of the variances that operate at the diverse levels, correct the underestimation of the standard errors and provide more precise estimations.

3. Executing an adequate multilevel analysis requires the selection of the best model. However, as there are multiple candidate models for the same sample evidence, useful tools must be employed to develop a good fit and a better selection of the model in reference to the observed data.

4. None of the tools examined consistently chooses the correct model. Nevertheless, the performance of these fit procedures improves when the number of repeated measures and the sample size increase.

5. The Efficient Information Criteria (AIC and AICC) perform better when the covariance patterns are complex, and worse when they are simple. The Consistent Information Criteria (BIC, CAIC and HQIC) perform better when the covariance patterns are simple, and worse when they are complex.

6. The Efficient Information Criteria are greatly affected by the lack of data normality.

7. The Consistent Information Criteria based on the number of individuals (n, level 2) are more effective than when they are based on the total number of observations (m, level 1).

8. The Intra-Class Correlation affects the performance of the Information Criteria, but this influence is low in comparison with the values of the parameters and with the correlation of the random effects.

9. The DIC Deviance Criterion performs worst than the rest of the fit criteria assessed, so it is recommended for conducting inferences after selecting the model.

10. With regard to simple size when selecting the best model, a large number of groups (NG) is more important than a large size of groups (SG), so that $NG \geq 50$ and $N/NG \geq 20$, with $N = NG \times SG$ is suggested.

REFERENCIAS

Aitkin, M., & Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society, Ser A, 149*, 1-43.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Alker, H. R. (1969). A Tipology of Ecological Fallacies. En M. Dogan & S. Rokkan (Eds.), *Quantitative Ecological Analysis in the School Science.* Cambridge Mass: MIT Press.

Amador, M., & López-González, E. (2007). Una Aproximación Bibliométrica a los Modelos Multinivel. *Revista Electrónica de Investigación y Evaluación Educativa, 13*(1), 67-82. Consultado el 09/11/2012, en http://www.uv.es/RELIEVE/v13n1/RELIEVEv13n1_3.htm.

Andreu, J. (2011). El Análisis Multinivel: Una Revisión Actualizada en el Ámbito Sociológico. *Metodología de Encuestas, 13*, 161-176.

Ato, M., Losilla, J. M., Navarro, J. B., Palmer, A. L., & Rodrigo, M. F. (2000). *Modelo Lineal Generalizado*. Terrassa: CBS.

Ato, M., & Vallejo, G. (2007). *Diseños Experimentales en Psicología.* Madrid: Pirámide.

Báez, B. (1994). El Movimiento de las Escuelas Eficaces: Implicaciones para la Innovación Educativa. *Revista Iberoamérica de Educación, 4*, 93-116.

Bickel, R. (2007). *Multilevel Analysis for Applied Research: It´s just Regression*. New York: Guilford Press.

Blanco, A., González, C., & Ordóñez, X. (2009). Achievement Measurements Correlation Patterns in Longitudinal Assessments: A Multilevel Approach Simulation Study. *Revista de Educación, 348,* 195-215.

Bolívar, A. (1999). *Cómo Mejorar los Centros Educativos.* Barcelona: Síntesis.

Borich, G. (2009). *Effective Teaching Methods.* Upper Saddle River, NJ: Merril Pub Co.

Bozdogan, H. (1987). Model Selection Akaikes´s Information Criteria (AIC): the General Theory and its Analytic Extensions. *Psychometrika, 52*(3), 345-370.

Braun, H. I. (2005*). Using Student Progress to Evauate Teachers: A Primer on Value-Added Models.* Princeton, NJ: Educational Testing Service.

Bryk, A. S. & Raundenbush, S. W. (1987). Applications of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin, 101,* 147-158.

Brown, A. (2009). *Teaching Strategies: A Guide to Effective Instruction*. Boston MA: Houghton Mifflin Company.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and Multilmodel Inference. A parcatical Information-Theoric Approach.* New York, NY: Springer.

Carlin, B. P., & Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.). London: chapman & Hall / CRC Press.

Claeskens, G., & Hjort, N. L. (2008). *Model Selection and Model Averaging*. New York: Cambridge University Press.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data. *Statistics in Medicine, 16,* 2349-2380.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartlanad, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Cornejo, R., & Redondo, J. M. (2007). Variables and Factors Associated to the Scholastic Learning. A Discussion from Actual Investigation. *Estudios Pedagógicos, 33*(2), 155-175. Consultado el 02/01/2013 en http://www.scielo.cl/pdf/estped/v33n2/art09.pdf

Creemers, B. (1997). La Base de Conocimientos sobre Eficacia Escolar. En D. Reynolds (Coord.), *Las Escuelas Eficaces. Claves para Mejorar la Enseñanza* (pp. 51-70). Madrid: Aula XXI/Santillana.

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.

De la Cruz, F. (2008). Modelos Multinivel. *Revista Peruana de Epidemiología*, *12*(3), 1-8. Consultado el 14/12/2012 en http://rpe.epiredperu.net/rpe_ediciones/v12_n03_2008/AR1.pdf.

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.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from a Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society, Ser B, 39*, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American statistical Association, 76,* 341-353.

DeSouza, C. M., Legedza, A. T., & Sankoh, A. J. (2009). An Overview of practical Approaches for Handling Missing Data in Clinical Trials. *Journal of Biopharmaceutical Statistics, 6,* 1055-1073. doi: 10.1080/10543400903242795.

Diggle, P. G., & Kenward, M. G. (1994). Informative Dropout in Longitudinal Data Analysis (with Discussion). *Applied Statistics in Medicine, 43,* 49-93.

Draper, D. (1995). Inference and Hierarchical Modeling in the Social Sciences. *Journal of Educational and Behavioral Statistics, 20*(2), 115-147.

Edmons, R. (1979). Effective Schools for the Urban Poor. *Educational Leadership, 37*(1), 15-24.

Elston, R. C., & Grizzele, J. E. (1962). Estimation of the Time Response Curves and Their Confidence Bands. *Biometrics, 18*(2), 148-159.

Fai, A. H. T. & Cornelius, P. C. (1996). Approximate F-tests of Multiple Degree of freedom Hypotheses in Generalized Least squares Analyses of Unbalanced Split-Plot Experiments. *Journal of statistical Computation and Simulation, 54,* 363-378.

Fermín, W., Galindo, P., Martín, J. (2007). Multilevel Strategy for Detecting Behaviours of Dropout Cases in the Analysis of Longitudinal Data. *Saber, Universidad de Oriente, Venezuela, 19*(1), 65-73.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of Misspecifying the First-Level Error Structure in Two-Levels Models Change. *Multivariate Behavioral Research, 37,* 379-403.

Fernández, M. J., & González, A. (1997). Desarrollo y Situación Actual de los Estudios de Eficacia Escolar. *RELIEVE, 3*(1). Consultado el 06/01/2013 en http://www.uv.es/relieve/v3n1_3.htm.

Fernández, P., Livacic-Rojas., P., & Vallejo, G. (2007). Cómo Elegir la Mejor Prueba Estadística para Analizar Un Diseño de Medidas Repetidas. *International Journal of Clinical and Health Psychology, 7*(1), 153-175.

Filp, J. (Ed.). (1984). *La Educación Preescolar mirada desde la Escuela.* Santiago de Chile: CIDE.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis,* (2nd ed.). New York: John Wiley & Sons.

Galán, J., Jiménez, E., & Cervantes, C. (2003). Restricted Maximum Likelihood Estimation of Variance Components of Multiple Traits Under Designs I and II North Carolina. *Revista Fitotecnia Mexicana, 26*(1), 53-66.

García, C. (1996). Estabilidad de Algunos Criterios de Selección de Modelos. *Qüestiió, 20*(2), 147-166.

Gelman A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Geweke, J., & Meese, R. (1981). Estimating Regression Models of Finite but Unknown Order. *International Review, 22,* 55-70.

Goldstein, H. (1980). Fifteen Thousand Hours: A Review of the Statistical Procedures. *Journal of Child Psychology and Psychiatry, 21,* 364-366.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* New York: Oxford University Press.

Goldstein, H. (1989). Models for Multilevel Response Variables With An Application To Growth Curves. In R. D. Bock (Ed.), *Multilevel Analysis of Educational Data.* San Diego, CA: Academic Press.

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

Gómez, S., Torres, V., García, Y., & Navarro, J. A. (2012). Procedimientos Estadísticos más Utilizados en el Análisis de Medidas Repetidas en el Tiempo en el Sector Agropecuario. *Revista Cubana de Ciencia Agrícola, 46*(1), 1-7.

Gómez, V. E., Schaalje, G. B., Fellingham, G. W. (2005). Performance of the kenward-Roger Method when the Covariance Structure is Selected Using AIC and BIC. *Communications in Statistics – Simulation and Computation, 34,* 377-392. doi:10.1081/SAC-200055719.

Good, T.L., Wiley, C. R. H., & Florez, I.R. (2009). Effective Teaching: An Emerging Synthesis. In L. J. Saha & A. G. Dworking (Eds.), *International Handbook of Research on Teachers and Teaching* (pp. 803-815). New York: Springer.

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S., & Jesson, D. (1999). *Improving Schools: Performance and Potential.* Buckingham: Open University Press.

Gurka, M. J. (2006). Selecting the Best Linear Mixed Model Under REML. *The American Statistician, 60,* 19-26.

Gurka, M. J., & Edwards, L. J. (2008). Mixed Models. In C. R. Rao, J. P. Miller & D. C. Rao (Eds.), *Handbook of Statistics 27: Epidemiology and Medical Statistics* (pp. 253-280). New York: Elsevier.

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.

Hannan, E. J., & Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Ser B, 41,* 190-195.

Heck, R. H. & Thomas, S. L. (2000). *An Introduction to Multilevel Modeling Techniques.* Mahwah, NJ: Lawrence Erlbaum Associates.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction Under A Selection Model. *Biometrics, 31*(2), 423-447.

Hertzog, C., Lindenberger, U., Ghisletta, P., & Von Oertzen, T. (2008). Evaluating the Power of Latent Growth Curve Models to Detect Individual Differences in Change. *Structural Equation Modeling, 15*, 541–563.

Hill, P. W., & Rowe, K. J. (1996). Multilevel Modeling in School Effectiveness Research. *School Effectiveness and School Improvement, 17*(1), 1-34.

Hogan, J. W., Jason, R. Korkontzelou, C. (2004). Tutorial in Bioestatistics. Handling Drop-Out in Longitudinal Studies. *Statistics in Medicine, 23,* 1455-1497.

Hopkins, D. (1995). Towards Effective School Improvement. *School Effectiveness and School Improvement, 6*(3), 265-274.

Hox, J. J. (1995). *Applied Multilevel Analysis.* Amsterdam: TT-Publikaties.

Hox, J. J. (1998). Multilevel Modelling: When and Why. In I. Balderjahn & M. Schader (Eds.), *Classification, Data Analysis and Data Highways* (pp. 147-154). New York: Springer Verlag.

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

Hox, J. J., & Kreft, I. G. G. (1994). Multilevel Analysis Methods. *Sociological Methods and Research, 22*(3), 238-299.

Hox J. J., & Roberts, J. K. (2011). *Handbook Advanced Multilevel Analysis*. New York: Routledge.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika, 76,* 297-307.

Jencks, C. S., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., & Michelson, S. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America.* Nueva York: Basic Books.

Kass, R. E., & Reftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90,* 773-795.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53*(3), 983-997.

Kenward, M. G., & Roger, J. H. (2009). An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood. *Computational Statistics & Data Analysis, 53*, 2583-2595.

Keselman, H. J., Algina, J., Kowalchuck, R. K., & Wolfinger, R. D. (1998). A Comparison of Two Approaches for Selecting Covariance Structures in the Analysis of repeated Measurements. *Communications in Statistics - Simulation and Computation, 27,* 591-604. doi: 10.1080/03610919808813497.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). The Analysis of Repeated Measurements: A Comparison of the Mixed Model Satterthwaite F Test and A Nonpooled Adjusted Degree of Freedom Multivariate Test. *Communications in Statistics-Theory and Methods, 28*(12), 2976-2999.

Kim, J. S. (2009). Multilevel Analysis: An Overview and some Contemporary Issues. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 337-361). London: Sage.

Koehler, A. B., & Murphree, B. S. (1988). A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order. *Applied Statistics, 37,* 187-195.

Kowalchuck, R. K., Keselman, H. J., Algina, J. Y., & Wolfinger, R. D. (2004). The Analysis of Repeated Measures with Mixed-Model Adjusted *F* Test. *Educational and Psychological Measurements, 64,* 224-242.

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

Kuehl, R. O. (2001). *Diseño de Experimentos. Principios Estadísticos de Diseños y Análisis de Investigación.* México: Thomson.

Laird, N. M., & Ware, H. (1982). Random Effects Models for Longitudinal Data. *Biometrics, 38,* 963-974.

Lee, H., & Ghosh, S. K. (2009). Performance of Information Criteria for Spatial Models. *Journal of Statistical Computation and Simulation, 79*(1), 93-106.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika, 73*(1), 13-22.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed Effects Models for Repeated Measures Data. *Journal of the American Statistical Association, 83,* 1014-1022.

Littell, R. C. (2002). Analysis of Unbalanced Mixed Model Data: A Case Study Comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological and Environmental Statistics, 7,* 472-490.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling Covariance Structure in the Analysis of Repeated Measures Data. *Statistics in Medicine, 19,* 1793-1819.

Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association, 90,* 1112-1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

Mata, P., & Ballesteros, B. (2012). Diversidad Cultural, Eficacia Escolar y Mejora de la Escuela: Encuentros y Desencuentros. *Revista de Educación, 358,* 17-37.

Martínez Arias, R. (2009). Usos, Aplicaciones y Problemas de los Modelos de Valor Añadido en Educación. *Revista de Educación, 348,* 217-250.

Martínez Arias, R., Gaviria, J. L., & Castro, M. (2009). Concepto y Evolución de los Modelos de Valor Añadido en Educación. *Revista de Educación, 348*, 15-45.

MacCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.

McCaffrey, D. L., Lakewood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Mónica, CA: RAND Corporation.

Mac Gilchrist, B., Myers, K., & Reed, J. (2004). *The Intelligent School.* Londres: Sage.

Miller, S. K. (1985). Research on Exemplary Schools: An Historical Perspective. In G.R. Austin & H. Garber (Eds.), *Research on Exemplary Schools* (pp. 3-30). Orlando, Fl: Academic Press.

Mills, J. A., & Prasad, K. (1992). A Comparison of Model Selection Criteria. *Econometric Reviews, 11,* 201-223.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School Matters: The Junior Years.* Somerset, UK: Open Books.

Mortimore, P. (1992). Issues in School Effectiveness. In D. Reynolds & P. Cuttance (Eds.), *School Effectiveness. Research, Policy and Practice* (pp. 154-163). London: Cassell.

Morgenstern, H. (1995). Ecologic Studies in Epidemiology: Concepts, Principles and Methods. *Annual Review of Public Health, 16,* 61-81.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). The Effects of School Membership on Pupils´Educational Outcomes. *Research Papers in Education, 3*(1), 3-26.

Núñez-Antón, V., & Zimmerman, D. L. (2001). Modelización de Datos Longitudinales con Estructuras de Covarianza No Estacionarias: Modelos de Coeficientes Aleatorios Frente a Modelos Alternativos. *Qüestiió, 25,* 225-262.

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*(1), 11-21. http://dx.doi.org/10.1016/j.cedpsych.2012.10.002.

Núñez, J. C., Vallejo, G., Rosário, P., Tuero-Herrero, E., & Valle, A. (in press). Student, Teacher, and School Context Variables Predicting Academic Achievement in Biology: Analysis from A Multilevel Perspective. *Journal of Psychodidactics*. doi: 10.1387/RevPsicodidact.7127.

Ojeda, M. M., & Velasco, F. (2012). Modelación Lineal Jerárquica aplicada a las Finanzas Públicas. En M. Ramos & F. Miranda (Eds.), *Tópicos Selectos de Optimización* (pp. 267-284). España: Universidad Santiago de Compostela, ECORFAN.

Oliver, J. C., Rosel, J., & Murray, L. (2000). Análisis de Medidas Repetidas Mediante Métodos de Máxima Verosimilitud. *Psicothema, 12*(2), 403-407.

Orlich, D. C., Harder, R. J., Callahan, R. C., Trevisan, M. S., & Brown, A. H. (2010). *Teaching Strategies: A Guide to Effective Instruction*. Boston, MA: Wadsworth.

Peña, E. (2011). Modelos Multinivel de los Factores de Eficacia Escolar en el Programa Pisa. Tesis Doctoral Inédita. Universidad de Oviedo.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus.* New York: Springer-Verlag.

Ponisciak, S. M., & Bryk, A. S. (2005). Value-Added Analysis of the Chicago Public Schools: An Application of Hierarchical Models. In R. Lissitz (Ed.), *Value-Added Models in Education: Theory and Applications* (pp. 40-79). Mapple Grove, MN: JAM press.

Prasad, N. G. N., & Rao, J. N. K. (1990). The Estimation of Mean Squared Errors of Small Area Estimators. *Journal of American Statistical Association, 85,* 163-171.

Preece, P. (1989). Pitfalls in Research on School and Teacher Effectiveness. *Research Papers in Education, 4*(3), 47-69.

Purkey, S. C., & Smith, M. S. (1983). Effective Schools: A Review. *The Elementary School Journal, 83,* 412-452.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park: Sage Publications.

Reise, S. P., & Duan, N. (2003). *Multilevel Modeling: Methodological Advances, Issues, and Applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Reynolds, D. (2007). School Effectiveness and School Improvement (SESI): Links with the International Standars/Accountability Agenda. In C. Teddlie & D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 471-485). Nueva York: Springer.

Roberts, K. H., & Burstein, L. (1980). *New Directions in Methodology: Aggregation Issues in Organizational Science.* San Francisco: Jossey-Bass.

Robinson, W. S. (1950). Ecological Correlations and the Behaviour of Individuals. *American Sociological Review, 15*(3), 351-357.

Román, M. (2008). Investigación Latinoamericana sobre Enseñanza Eficaz, ILEE. En UNESCO (Ed.), *Eficacia Escolar y Factores Asociados en América Latina y el Caribe* (pp. 209-255). Santiago de Chile UNESCO.

Rose, G. (1985). Sick Individuals and Sick Populations. *International Journal of Epidemiology, 14,* 32-38.

Rosenberg, B. (1973). Linear Regression with Randomly Dispersed Parameters. *Biometrika, 60,* 61-75.

Sanders, W. L. (2006, October 16). *Comparisons Among Various Educational Assessment Value- Added Models*. Paper presented at The Power of Two- National Value-Added Conference, Columbus, Ohio.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley & Sons, Inc.

Singer, J. D. (2002). Fitting Individual Growth Models Using SAS PROG MIXED. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling Intraindividual Variability with Repeated Measures Data: Methods and Applications* (pp. 135-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, J. D., & Willet, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Snijders, T. A. B., & Bosker, R. T. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* London: Sage.

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, Ser B, 64*, 583-640.

Sterba, S. K., & Pek, J. (2012). Individual Influence on Model Selection. *Psychological Methods, 17*, 582-599.

Stoll, L., & Fink, D. (1996). *Changing Our Schools. Linking School Effectivenessand School Improvement.* Buckingham: Open University Press.

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research.* London, New York: Falmer Press.

Thrupp, M. (1999). *Schools Making a Difference: Lets´be Realistic! School Mix, School Effectiveness and the Social Limits of Reform*. Buckingham: Open University Press.

Townsend, T. (Ed.) (2007). *International Handbook of School Effectiveness and Improvement.* Nueva York: Springer.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, *92*, 351-370.

Vallejo, G., Arnau, J., & Bono, R. (2008). Construction of Hierarchical Models in Applied Contexts. *Psicothema, 20*(4), 830-838.

Vallejo, G., Arnau, J., Bono, R., Fernández, P., & Tuero-Herrero, E. (2010). Nested Model Selection for Longitudinal Data Using Information Criteria and the Conditional Adjustment Strategy. *Psicothema, 22*(2), 323-333.

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of Misspecifying the Error Covariance Structure in Linear Mixed Models for Longitudinal Data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(1), 10-21. doi: 10.1027/1016-9040.12.1.10.

Vallejo, G., Ato, M., & Fernández, M. P. (2010). A Robust Approach for Analyzing Unbalanced Factorial Designs with Fixed Levels. *Behavior Research Methods, 42,* 607-616. doi: 10.3758/BRM.42.2.607.

Vallejo, G., Fernández, J. R., & Secades, R. (2003). Análisis Estadístico y Consideraciones de Potencia en la Evaluación de Programas Mediante Diseños de Muestreo de Dos Etapas. *Psicothema, 15*(2), 300-308.

Vallejo, G., Fernández, J. R., & Secades, R. (2004). Application of A Mixed Model Approach for Assessment of Interventions and Evaluation Programs. *Psychological Reports, 95,* 1095- 1118.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero-Herrero, E. (2011a). Comparison of Modern Methods for Analyzing Unbalanced Repeated Measures Data. *Multivariate Behavioral Research, 46,* 900-937. doi:10.1080/00273171.2011.625320.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero-Herrero, E. (2011b). Selecting the Best Unbalanced Repeated Measures Model. *Behavior Research Methods,* 43, 18-36. doi: 10.3758/s13428-010-0040-1.

Vallejo, G., & Lozano, L. M. (2006). Modelos de Análisis para los Diseños Multivariados de Medidas Repetidas. *Psicothema, 18*(2), 293-299.

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.*

Van der Leeden, R. (1998a). Multilevel Analysis of Longitudinal Data. In C. H. J. Bijleveld & L. J. Th van der Kamp (Eds.), *Longitudinal Data Análisis: Designs, Models and Methods* (pp. 269-317). London: Sage.

Van der Leeden, R. (1998b). Multilevel Analysis of Repeated Measures Data. *Quality & Quantity, 32,* 15-29.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Wang, J., & Schaalje, G. B. (2009). Model Selection for Linear Mixed Models Using Predictive Criteria. *Communications in Statistics – Simulation and Computation, 38,* 788-801. doi: 10.1080/03610910802645362.

Wiley, E. V. (2006). *A Practitioner´s Guide to Value Added Assessment.* Tempe, AZ: Educational Policy Studies Laboratory, Arizona State University.

Wolfinger, R. D. (1996). Heterogeneous Variance-Covariance, Structures for Repeated Measures. *Journal of Agricultural, Biological and Environmental Statistics, 1,* 205-230.

Wu, Y. W., Clopper, R. R., & Wooldridge, P. J. (1999). A Comparison of Traditional Approaches to Hierarchical Linear Modeling when Analyzing Longitudinal Data. *Research in Nursing & Healt, 22,* 421-432.