Concepto y evolución de los modelos de valor añadido en educación

Concept and evolution of educational value-added models

Rosario Martínez Arias

Universidad Complutense de Madrid. Facultad de Psicología. Departamento de Metodología de las Ciencias del Comportamiento. Madrid, España

José Luis Gaviria Soto

María Castro Morera

Universidad Complutense de Madrid. Facultad de Educación. Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE). Madrid, España

Resumen

Los modelos de valor añadido son un conjunto de técnicas estadísticas complejas que utilizan datos de puntuaciones de tests de los estudiantes de varios años, para estimar los efectos de las escuelas individuales. Los modelos intentan aislar la contribución de la escuela al desarrollo del aprendizaje de los alumnos. Existen diversas variaciones de los modelos que se utilizan en trabajos de investigación y en evaluaciones prácticas. En este artículo se presenta el concepto de valor añadido de la escuela y su historia y evolución. Se establece el origen y desarrollo de los modelos en torno a tres líneas principales: la investigación sobre la efectividad de las escuelas, las críticas derivadas de los informes actuales de rendición de cuentas y el desarrollo de los modelos estadísticos multinivel. Se explica la motivación de su uso para evitar algunos problemas frecuentes en las presentaciones de los resultados de las escuelas y en su ordenación, tal como sucede con las «Tablas de Liga en Inglaterra» y los informes de «Progreso Anual Adecuado» consecuencia de la aplicación de la *No Child Left Behind* en USA. Los principales problemas analizados son los relacionados con la metodología transversal de las evaluaciones

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

y los sesgos de selección derivados de las características de los estudiantes en el ingreso, variables sociodemográficas y factores contextuales de las escuelas. En resumen, se exponen las principales aproximaciones existentes para medir los efectos de las escuelas y se concluye que la aplicación de los procedimientos de Valor Añadido representa una importante promesa para la evaluación de las escuelas.

Palabras clave: investigación sobre efectividad de las escuelas, rendición de cuentas, evaluación de escuelas, modelos de valor añadido.

Abstract

Value-added modelling is a collection of complex statistical techniques which use multiple years of students' test score data to estimate the effects of individual schools. The models attempt to isolate the contributions of schools to student learning development. Several variations of these models are being applied in the research literature and in practical school assessments.

In this paper we introduce the concept of value-added and its history and evolution. We establish the origin and the development of the models around three main lines: the school effectiveness research, the criticism on actual school accountability reports, and the development of multilevel statistical models. We explain the motivation to use these models in order to prevent some problems with the reporting of school results and school ratings, such as the League Tables in England and the Adequate Yearly Progress from the *No Child Left Behind* in USA. The main analyzed problems are those related to the cross-sectional methology of most of assessments, and the selection biases derived from the student characteristics at the intake, socio-demographic variables and contextual factors of schools. We summarize the principal existing modeling approaches for measuring school effects and we conclude that the application of value-added procedures holds considerable promise for school assessment.

Key Words: school effectiveness research, school accountability, school assessment, value-added models.

Introducción

Muchos gobiernos están en la actualidad insatisfechos con los niveles de rendimiento de sus estudiantes y reciben presiones crecientes para mejorar la eficacia y eficiencia del sistema educativo. Esta preocupación está abonada por la investigación aplicada de la economía de la educación que establece que la prosperidad económica y social

de los individuos y de las naciones radica en la educación y formación (Hanushek, 2005; Hanushek y Wossman, 2006). El nivel de logros educativos de un país suele considerarse como un indicador de sus reservas de «capital humano» o disponibilidad de potenciales trabajadores con educación y destrezas suficientes. Además, la sociedad debe asegurar la igualdad de acceso a las oportunidades educativas para lograr niveles satisfactorios de bienestar individual y social. Es decir, la educación proporcionada debe ser equitativa, para lo que deben reducirse las brechas de rendimiento entre diferentes grupos de estudiantes. Esta preocupación por alcanzar adecuados niveles de rendimiento se ha extendido a la sociedad en general, como pone de relieve el creciente interés mediático por los resultados de las encuestas internacionales.

Tanto los partidarios de la rendición de cuentas de las escuelas como los no partidarios reconocen que un ingrediente fundamental para la mejora de las escuelas y del sistema educativo en general es un buen sistema de evaluación de las escuelas, que permita disponer de datos en los momentos adecuados. Estos datos servirán para emprender acciones de mejora de las escuelas y para la planificación de las reformas educativas.

Es en este contexto en el que la metodología conocida como modelos de Valor Añadido (en adelante VA) se ha recibido con entusiasmo y con grandes expectativas por parte de los responsables de la educación.

Aunque no existe una definición única de lo que es el VA de una escuela, suele considerarse como su contribución al progreso neto de los estudiantes hacia objetivos de aprendizaje establecidos, una vez eliminada la influencia de otros factores ajenos a la escuela que pueden contribuir a dicho progreso. Los modelos de VA son un conjunto de procedimientos estadísticos que se utilizan para hacer inferencias sobre la eficacia de las escuelas y de los profesores que ponen el acento en las ganancias de los estudiantes en el tiempo. Tienen en común el seguimiento de la trayectoria de los estudiantes analizando las medidas de los resultados de dos o más años. Estos datos de la evolución de los estudiantes se transforman en indicadores de la eficacia de la escuela o del profesor. La idea es simple: las escuelas más eficaces son aquellas cuyos estudiantes ganan más, mientras que las menos eficaces son las que ganan menos. Algunos de estos modelos, los denominados contextualizados, ajustan los resultados por variables socioeconómicas y demográficas con objeto de hacer más justas las comparaciones entre escuelas.

La importancia atribuida a los modelos de VA en la actualidad se debe a un gran interés por hacer a las escuelas responsables de sus rendimientos rindiendo cuentas a la sociedad de los aprendizajes de sus estudiantes. Este interés se encuentra fomentado por los procesos de descentralización y de autonomía creciente para las escuelas que

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

se han emprendido en muchos países. La autonomía debe estar acompañada de evaluaciones estructuradas y sistemáticas de los resultados que proporcionen feedback a las escuelas sobre sus fuerzas y debilidades y sobre qué aspectos deben mejorar. Es en este contexto en el que los modelos de VA pueden ser de extraordinaria utilidad.

Aunque el interés por los modelos de VA para la evaluación es reciente, en realidad no son tan nuevos, sino que son el producto de diferentes líneas de investigación que marcan su desarrollo y algunos de los debates sobre su utilización. En este artículo se presentan las líneas de investigación que han llevado a los modelos de VA actuales, las razones o motivación para su implantación y los principales modelos propuestos. Limitaciones de espacio impiden entrar en una explicación detallada de los modelos. Los lectores interesados pueden encontrar información complementaria en algunas fuentes recientes como la revisión realizada por McCaffrey, Lockwood, Koretz y Hamilton (2003), el número monográfico de la revista *Journal of Educational and Behavioral Statistics* (2004, (29), 1), editado por Wainer y en las compilaciones de los symposia recientes sobre el tema de la Universidad de Maryland editadas por Lissitz (Lissitz, 2005, 2006).

El caldo de cultivo para los modelos de VA

En este apartado y sin ánimo de exhaustividad, se presentan los antecedentes que están en la base de los modelos de VA. Se agrupan los antecedentes en tres líneas: la investigación sobre las escuelas efectivas, la política de rendición de cuentas y los avances en las técnicas estadísticas que hicieron posible la aplicación de los modelos. En realidad, es difícil separarlas, ya que están bastante unidas y son coexistentes en el tiempo. La investigación sobre la efectividad de las escuelas está más enraizada en la búsqueda de factores que determinan la calidad y la equidad de los aprendizajes; la línea que denominamos de rendición de cuentas basada en las puntuaciones de los tests puede considerarse más vinculada a los esfuerzos de evaluación de las escuelas, con consecuencia o no, dentro de las políticas de publicidad de resultados del sector público. Los desarrollos de las técnicas estadísticas están en la base de las dos líneas. En el desarrollo de los modelos también han influido la tendencia creciente a las evaluaciones a gran escala mediante tests estandarizados, nacionales e internacionales y la creciente disponibilidad de ordenadores para la creación y almacenamiento de bases de datos y de un software adecuado para su tratamiento.

La investigación sobre la efectividad de las escuelas IEE (School Effectiveness Research)

Sus orígenes se encuentran principalmente en el Reino Unido y Estados Unidos, donde influyentes 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/raza y el estatus socioeconómico (Coleman et al., 1966; Jencks et al., 1972).

Estos tempranos estudios adolecían de un buen número de limitaciones metodológicas y la investigación posterior intentó poner de relieve la existencia de efectos escolares significativos, aún reconociendo la gran influencia del contexto socioeconómico y cultural de los estudiantes (Edmonds, 1979; Mortimore, Sammons, Stoll, Lewis y Ecob, 1988). Se intenta dar respuesta a las siguientes preguntas: *¿Pueden las escuelas ser efectivas?, ¿Qué determina la efectividad de la escuela? y ¿Hasta qué punto la escuela es efectiva en la reducción de las desigualdades en rendimiento debidas al origen social o étnico de los sujetos?*

Desde sus comienzos, la IEE se caracteriza por seguir una metodología fundamentalmente cuantitativa, basada en las puntuaciones de los estudiantes en tests estandarizados y por el uso de procedimientos estadísticos complejos que permitan desenmarañar de los resultados de los estudiantes la influencia de otros efectos ajenos a la escuela y a sus prácticas educativas. En un primer momento se caracterizó por el estudio de lo que se podría denominar el «rendimiento contextualizado» mediante estudios transversales realizados en un momento en el tiempo. Un ejemplo de esta metodología son los estudios PISA, en los que se intenta eliminar de los resultados otros aspectos relacionados con características socioculturales de los estudiantes. Un ejemplo de este enfoque con los datos españoles de PISA 2003 puede encontrarse en Marchesi y Martínez Arias (2006). Esta aproximación por su carácter transversal resulta insuficiente (Willms y Raudenbush, 1989; Goldstein, 1987). La investigación más reciente se caracteriza por estudiar muestras amplias de escuelas y poner el acento en la evaluación del progreso sobre el tiempo, más que en las instantáneas tomadas en un único momento derivadas de los estudios transversales. En la actualidad y siguiendo a Sammons (2006) puede considerarse la IEE como una línea de investigación que intenta separar las características de los estudiantes de los efectos de las escuelas.

No podemos entrar aquí en los principales resultados derivados de la IEE y remitimos a los lectores a Scheerens y Bosker (1997) y Teddlie y Reynolds (2000) y a la revisión de los diferentes meta-análisis realizados por Scheerens (2005). En general,

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

puede decirse que sus resultados son algo más alentadores que los de los tempranos estudios de Coleman y Jenks. Puede concluirse que *las escuelas establecen diferencias* (MacBeath y Mortimore, 2001; Reynolds y Creemers, 1990; Sammons, Hillman y Mortimore, 1995; Sammons y Reynolds, 1997; Scheerens, 2005; Scheerens y Bosker, 1997; Teddlie y Reynolds, 2000). Hay un porcentaje de la varianza entre escuelas (entre el 5-35%, según los estudios) que se explica por políticas y prácticas educativas y por el ambiente y clima del aprendizaje de la escuela. También se ha encontrado una escasa influencia de las variables de *inputs* o recursos en los países desarrollados, excepto en los aspectos relacionados con la formación y experiencia del profesorado, resultado coincidente con los derivados de la línea conocida como «función de producción» (Hanushek, 2003) en la Economía de la Educación. Estos datos han desplazado el acento de los *inputs* a los procesos y a los resultados en la evaluación educativa. Por lo que se refiere a las variables de procesos de las escuelas, los resultados son algo ambigüos (Muijs y Reynolds, 2000; Stevens, 2005; Zvoch y Stevens, 2003). Otro resultado bastante claro es que las escuelas efectivas responden mejor a la meta de la equidad, ya que los sujetos en desventaja progresan más en estas escuelas.

En lo que sí existe un amplio acuerdo es en que las comparaciones entre escuelas no pueden establecerse sobre los resultados brutos, sino que deben basarse en ajustes del rendimiento inicial y de otros factores relevantes y en el progreso de sus estudiantes (Goldstein et al., 1993; Goldstein y Thomas, 1996; Gray, Jesson, Goldstein, Hedger y Rasbash, 1996; Mortimore, Sammons y Thomas, 1994; Sammons, 1996). Se considera una escuela efectiva aquella en la que los estudiantes progresan más allá de lo que puede esperarse, añadiendo valor extra a los resultados de sus alumnos en comparación con otras escuelas que sirven a poblaciones que son similares en el ingreso.

La evaluación mediante tests al servicio de las reformas educativas y la política de rendición de cuentas

La evaluación externa mediante tests estandarizados, ha sido constante en los diversos intentos de reformas educativas en los Estados Unidos desde los años sesenta (Hamilton, 2003; Wang, Beckett y Brown, 2006). Destacan, entre otros, los usos de los tests para el diagnóstico y monitorización del sistema educativo, representados

por el *National Assessment of Educational Progress* (NAEP), que comenzó en 1969 y que introdujo una rigurosa metodología psicométrica y estadística que tuvo una gran repercusión en las evaluaciones posteriores nacionales e internacionales. Además, las alarmas desencadenadas tras la participación de los Estados Unidos en las tempranas evaluaciones internacionales del rendimiento educativo bajo los auspicios de la *International Association for the Evaluation of Educational Achievement* (IEA) influyeron considerablemente en los posteriores desarrollos de la evaluación. Un hito importante fue la publicación en 1983 del influyente texto *A nation at risk* (National Commission of Excellence in Education, 1983). En él se insiste en los riesgos derivados de la pérdida de competitividad futura frente a otras naciones, como consecuencia de la baja calidad de la educación. Las reflexiones a que dio lugar pueden considerarse el desencadenante de diversas reformas educativas para las que resultó esencial la evaluación mediante tests, tales como el cambio de las competencias mínimas a altos estándares de rendimiento y la necesidad de disponer de datos para la monitorización y reforma del sistema. Como consecuencia, en muchos estados se iniciaron evaluaciones a gran escala, para disponer de datos para la rendición cuentas y para la mejora de las escuelas. Todo ello culminó en la legislación federal con *No Child Left Behind* (NCLB, aprobada en enero de 2002), con la exigencia de la rendición de cuentas de las escuelas en todos los estados para poder acogerse a beneficios federales.

Es importante destacar también la introducción a partir de los años noventa de la evaluación basada en *estándares de rendimiento* dentro de las reformas encaminadas a la mejora de las escuelas, sistema que recoge la NCLB. La idea fundamental de esta forma de evaluación es que los gobiernos especifican lo que los estudiantes deben saber y ser capaces de hacer en los distintos cursos y materias y estas especificaciones se recogen en los denominados *estándares de contenido* (O'Day y Smith, 1993). Los *estándares de rendimiento* representan una forma de evaluación referida a criterios que permiten establecer diferentes niveles de logro que se establecen a partir de las puntuaciones de los tests, juicios de expertos y consideraciones de política educativa. Para mas información véanse Cizek y Bunch (2007) y en castellano Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006). En general, esta forma de evaluación con altos estándares y la clasificación por niveles es muy bien aceptada por los profesores y por la opinión pública (Wang et al., 2006).

El uso de los tests está tan generalizado que se habla de la rendición de cuentas basada en tests, aunque evaluación y rendición de cuentas son dos procesos diferentes (Hill, Scout, DePascale, Duna y Simpson, 2006). Un aspecto importante de

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

estas políticas es evaluar si los estudiantes realizan progresos satisfactorios y si alcanzan los estándares de rendimiento establecidos por las autoridades. También, en aras de la equidad, se evalúa el rendimiento de distintos grupos de estudiantes, caracterizados por diferentes variables sociodemográficas (género, etnia, estatus socioeconómico).

La política de rendición de cuentas no llegó con la NCLB como algo nuevo, sino que en los últimos veinte años ya hubo un creciente interés dentro de las políticas generales de los gobiernos y estados en Estados Unidos y en otros países en la evaluación externa de las escuelas como un instrumento de supervisión útil para la mejora continua del sistema educativo (Braun y Kanjee, 2006; Kane y Staiger, 2002; Goldstein y Spiegelhalter, 1996; Hanushek y Raymond, 2004; Taylor y Nguyen, 2006). Aunque hay diferentes formas de rendición de cuentas que van desde la simple publicación de los resultados a incentivos más directos como la aplicación de recompensas y sanciones, en todos los casos, siempre es un instrumento de los políticos para ver el grado de cumplimiento de las metas, nunca una meta en sí misma. Se supone que aporta una información que puede apoyar la mejora continua de las escuelas.

Dentro de esta línea de rendición de cuentas para la mejora de los rendimientos de todos los estudiantes y de los estándares de rendimiento, la NCLB establece que todos deben alcanzar el nivel de competente en 2014, estableciendo metas anuales de progreso. También se establecen metas que reduzcan las brechas de rendimiento entre diferentes subgrupos. Impone la evaluación mediante tests anuales de Lengua y Matemáticas en todas las escuelas en los cursos de 3º a 8º. Establece consecuencias en términos de recompensas y sanciones para las escuelas según el cumplimiento o no de los objetivos. Son precisamente las consecuencias la parte más controvertida de la NCLB (Cronin, Kingsbury, McCall y Bowe, 2005; Carlson, Martínez, O'Day, Stecher, Taylor y Cook, 2007).

Los principales modelos alternativos a los de VA y que han sido objeto de numerosas críticas son el modelo de estatus y el modelo de mejora. El modelo de estatus (Goldschmidt et al., 2005; Raudenbush, 2004a) compara los resultados anuales de la escuela con el objetivo establecido (el nivel de competente), conocido como *Adequate Yearly Progress* (AYP). El modelo de mejora compara los porcentajes de estudiantes de una cohorte que alcanzan el objetivo en un curso y año particular con los porcentajes de la cohorte siguiente en el mismo curso que tiene el problema de que las diferencias en rendimiento pueden deberse a errores de muestreo y cambios en las cohortes (Hill y DePascale, 2003; Linn y Haug, 2002; Linn, Baker y Betebenner, 2002; Rouse, 2005; Zvock y Stevens, 2006).

Muchas de las críticas a los modelos de estatus y de mejora, en sus diversas variantes, se deben precisamente al uso de datos transversales (Linn, 2005; Linn y Haug, 2002; Raudenbush, 2004,b; Rouse, 2005; Stevens, 2005; Thum, 2002; Zvoch y Stevens, 2006), dada su menor susceptibilidad a las características de los estudiantes y otros factores contextuales. Hay una gran cantidad de incertidumbre asociada con las ganancias de cohortes sucesivas debida a errores de muestreo y diferencias de movilidad. El seguimiento de los estudiantes es el que permite separar los efectos sistemáticos de las escuelas de las características de los estudiantes.

Las ganancias promedio en rendimiento basadas en el seguimiento longitudinal individual de los estudiantes son las únicas que pueden proporcionar bases para eliminar (al menos en parte) las explicaciones competidoras de las diferencias en rendimiento entre escuelas (Sanders y Horn, 1994, 1998; Ballou, Sanders, y Wright, 2004; McCaffrey et al, 2003; McCaffrey, Lockwood, Koretz, Louis y Hamilton, 2004).

La rendición de cuentas en Inglaterra

La publicación de la normativa *Every Child Matters* (1988) supuso el comienzo de la publicación de los resultados de las escuelas en tests externos realizados en determinados momentos o estadios clave. En este caso el objetivo de la evaluación es el de la identificación de escuelas efectivas e inefectivas, de cara a la introducción de mejoras y proporcionar información a los padres para la elección de centro. Esta práctica fue continuada por los gobiernos laboristas. Inicialmente los resultados se reportaban en forma de las denominadas *Tablas de Liga,* cuyo uso fue muy controvertido desde sus inicios. Los principales argumentos de sus detractores se basaban en la injusticia de los resultados brutos debido al extraordinario impacto que en ellos tenía el nivel anterior de los estudiantes y variables sociodemográficas (Goldstein y Spiegalhalter, 1996; Nuttall, Goldstein, Prosser y Rasbash, 1989; Saunders, 1999; Yang, Goldstein, Rath y Hill, 1999). Mostraron que después del ajuste de los niveles previos de los estudiantes cambiaba la ordenación de muchas escuelas y que cuando se tomaban los intervalos de confianza en torno a las medias, la mayor parte de las escuelas no se diferenciaban del promedio (Fitz-Gibbon, 1997). Estas consideraciones llevaron a una aproximación a la evaluación basada en modelos de VA para ajustar los efectos del rendimiento anterior y posteriormente al valor añadido contextualizado en 2005, incluyendo además los intervalos de confianza (Ray, Evans y McCormack, 2008, en este volumen; Schagen, 2006). Una revisión de la evolución del VA en Inglaterra puede verse en Saunders (1999).

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Los procedimientos estadísticos

El camino hacia los modelos de VA en la evaluación no habría sido posible sin el desarrollo de modelos estadísticos que permiten descomponer la variación de los resultados de los estudiantes en diferentes fuentes de variación, analizar variables procedentes de distintos niveles y tener en cuenta las dependencias de los datos de las escuelas. La mayor parte de las técnicas estadísticas paramétricas utilizadas tradicionalmente en la evaluación de los efectos de las escuelas suponen la independencia de los errores aleatorios, provocando importantes sesgos cuando no se cumple este supuesto. Por otra parte, los datos educativos tienen una estructura multinivel en la que las escuelas están anidadas en contextos, las clases en las escuelas y los estudiantes en clases y profesores. Esta estructura de los datos provoca dependencias entre las unidades de análisis, que viola el supuesto de independencia. Por otra parte, en la investigación de la eficacia de las escuelas se da el problema de integrar datos procedentes de diferentes niveles en un modelo único, integrando en la predicción del resultado predictores de diferentes niveles (De Leeuw y Meijer, 2008). Los iniciales trabajos de Coleman et al. (1966) y de Jencks et al. (1972) no tuvieron en cuenta ni la estructura de los datos, ni las dependencias. Ya en 1939 Thondike advirtió de los peligros de usar estimaciones derivadas de la correlación y regresión registradas en el nivel de grupo para hacer inferencias relativas a individuos y subgrupos, fenómeno denominado por Robinson (1950) como «falacia ecológica». Se llegó a una crítica metodológica explícita sobre las aproximaciones de nivel agregado (Burnstein, 1980), llegando en los años ochenta a las aproximaciones hoy conocidas como «modelos lineales mixtos», «modelos multinivel» o «modelos lineales jerárquicos», que ya se estaban utilizando en otras áreas (Aitkin y Longford, 1986; DeLeeuw y Kreft, 1986, Goldstein, 1987; Raudenbush y Bryk, 1986). Estos modelos permitieron la partición de la varianza en diversos niveles y la inclusión de variables predictoras que permiten explicar estas varianzas, con la inclusión de efectos de interacción. La metodología propuesta permite evitar numerosos errores de inferencia estadística (errores de tipo I, sesgos de agregación, heterogeneidad de la regresión, etc.) y separar los efectos de la escuela de otros factores de los estudiantes, llevando a los modelos de rendimiento contextualizados. Estos modelos representaron un gran paso, pero no acabaron con el debate, debido a la inestabilidad temporal de los resultados; escuelas ejemplares un año pueden ser mediocres en otro, a pesar de la supuesta estabilidad de las políticas y prácticas de las escuelas). La extensión de estos modelos a datos longitudinales (Wilms y Raudenbush, 1989) supuso

un importante avance para examinar trayectorias de desarrollo y los efectos de las escuelas sobre estas trayectorias.

En la actualidad existe un importante cuerpo de literatura en forma de libros que permite conocer estos modelos (De Leeuw y Meijer, 2008; Goldstein, 2003; Raudenbush y Bryk, 2002; Gaviria y Castro, 2005; Hox, 2002; Gelman y Hill, 2007). El uso de estas técnicas y su aplicación a la evaluación de los efectos de las escuelas no habría sido posible sin los desarrollos de software estadístico que permitan implementar los complejos procesos de estimación. Se dispone en la actualidad de numeroso software en programas específicos como MLWIN, HLM o AML, y en paquetes de programas de uso general, tanto comerciales (SAS, SPSS, SYSTAT, MPLUS, STATA-GLLAMM, SYSTAT, SPLUS), como de libre acceso (WinBUGS, R), por mencionar solamente algunos[1].

Motivación para usar los modelos de VA

Las numerosas críticas derivadas del uso de los modelos utilizados para cumplir con los requisitos de la NCLB, así como otras críticas, entre las que destacan las realizadas a las tablas de liga, condujeron a un considerable entusiasmo por los modelos de VA como una alternativa (Carey, 2004). Aunque se siguen presentando los resultados como se ha descrito en el apartado anterior, el Gobierno de Estados Unidos se hizo eco de las principales críticas y a finales de 2005 invitó a los estados a hacer propuestas de modelos basados en ganancias de los estudiantes, aprobando algunos estudios piloto que tienen en cuenta el valor añadido.

Las críticas vienen de muchos frentes, aunque la mayor parte pueden agruparse en dos grandes bloques:

- Los sesgos de selección en las escuelas.
- El uso de estudios transversales para la evaluación de las mejoras de aprendizaje.

(1) Las siglas se refieren a los nombres de los programas. Una revisión crítica de la mayor parte de los mencionados con ejemplos de aplicación puede encontrarse en el *Centre for Multilevel Modelling* de la Universidad de Bristol: http://www.cmm.bristol.ac.uk

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Los efectos de las escuelas y los sesgos de selección y composición

En la evaluación de las escuelas y sus efectos y su posterior ordenación se trata de comparar los resultados en términos del rendimiento alcanzado por los estudiantes. En la base de todas las críticas se encuentra la complejidad de lo que encierran las puntuaciones brutas de los tests y de lo que los estudios basados en ellas entienden por efectos de la escuela. Raudenbush y Willms (Raudenbush y Willms, 1995; Willms y Raudenbush, 1989) establecieron una distinción ya clásica entre los diferentes tipos de efectos de las escuelas y lo que representan dentro de las puntuaciones de los tests. Los dos tipos de efectos se basan en el análisis de diferencias entre el rendimiento (o ganancia) de un niño en una escuela particular con el esperado si hubiese estado en otro entorno. La elección del entorno de comparación es crítica para los diferentes usos de la información sobre el efecto de la escuela. Definen dos tipos de efectos de interés, Tipo A y Tipo B. El efecto tipo A es la diferencia entre el rendimiento actual de un alumno y el esperable si hubiese asistido a una escuela «típica», equivalente al que se obtendría si los estudiantes de idénticas características fuesen asignadas al azar a las M escuelas bajo evaluación. Normalmente es el efecto que consideran los padres cuando eligen escuela. Este efecto incluye prácticas de la escuela, composición de los estudiantes y el contexto en el que está localizada. El efecto tipo B se refiere exclusivamente al debido a las prácticas de la escuela, que están bajo la responsabilidad de sus profesionales. Este efecto debe ser separado del contexto de la escuela y otros factores externos (contexto, composición del alumnado, etc.). Este efecto es la diferencia entre el rendimiento de un alumno en una escuela particular y el que sería esperable si asistiese a una escuela con contexto idéntico, con unas prácticas de efectividad promedio. Sería como si M escuelas con contextos idénticos fuesen asignadas a niveles de tratamiento con variación en las prácticas. Este es el efecto *justo* o imparcial que se debería considerar en la evaluación de la eficacia de las escuelas. Una discusión sobre la problemática de establecer inferencias causales en la atribución de dichos efectos a partir de estudios observacionales puede verse en Raudenbush y Willms (1995) y Rubin, Stuart y Zanutto (2004).

Los efectos tipo B se refieren al rendimiento de los estudiantes en una clase y escuela comparado con el de los estudiantes de contextos similares. Estos son los de interés para la rendición de cuentas, ya que otros aspectos no están bajo control de la escuela. La estimación de los efectos tipo A implicaría la sustracción de los efectos de las características de los estudiantes y de las correspondientes interacciones. La

estimación de los efectos tipo B implicaría la sustracción de los efectos de las características de los estudiantes y de los efectos contextuales, junto con las correspondientes interacciones.

Los modelos de VA intentan separar las contribuciones relativas de los distintos efectos, para estimar en la medida de lo posible los efectos de tipo B, atribuibles a la escuela, mediante complejos procedimientos estadísticos que intentan paliar la ausencia de aleatorización en la composición de las escuelas. Se considera que la rendición de cuentas y la evaluación de las escuelas deben limitarse a las partes de la varianza que están bajo su control.

Los modelos antes presentados en la rendición de cuentas y en las tablas de liga originales confunden los efectos de las escuelas con los restantes aspectos que no están bajo su control y que no se deben a las políticas y a las prácticas escolares (Choi, Yamashiro, Seltzer, y Herman, 2004; Rumberger y Palardy, 2004; Linn, 2004, 2005; McCall, Kingsbury y Olson, 2004; Zvoch y Stevens, 2006; Rowe, 2000; Aitkin y Longford, 1986; Hanushek, 1979; Raudenbush, 2004b; Ladd y Walsh, 2002; Novak y Fuller, 2003).

Todos estos efectos derivados de características de los estudiantes, contextos de las escuelas y sus posibles interacciones son los referidos como *sesgos de selección,* ya que la distribución de los estudiantes y de los profesores a las escuelas no es aleatoria. No considerar estos efectos puede llevar a grandes injusticias y a la desmoralización de profesores y directores altamente cualificados y que logran ganancias de aprendizaje considerables con sus estudiantes, pero que atienden a poblaciones en desventaja.

Es necesario realizar los ajustes por las características de los estudiantes y contextuales, lo que lleva a medidas más justas y seguras de los resultados de las escuelas, para que las escuelas puedan ser juzgadas de forma justa (Linn, 2004; 2005). La realización de los ajustes teniendo en cuenta los resultados anteriores de los estudiantes y de otras características, implican necesariamente la utilización de modelos de VA.

Funcionamiento y requisitos de los modelos de VA para la evaluación

Como se ha visto anteriormente, los modelos de VA son un caso especial de los modelos de desarrollo que intentan aislar los efectos de las escuelas por medio del uso de datos longitudinales (Lissitz, Doran, Schafer y Willhoft, 2006), de modo que las varianzas

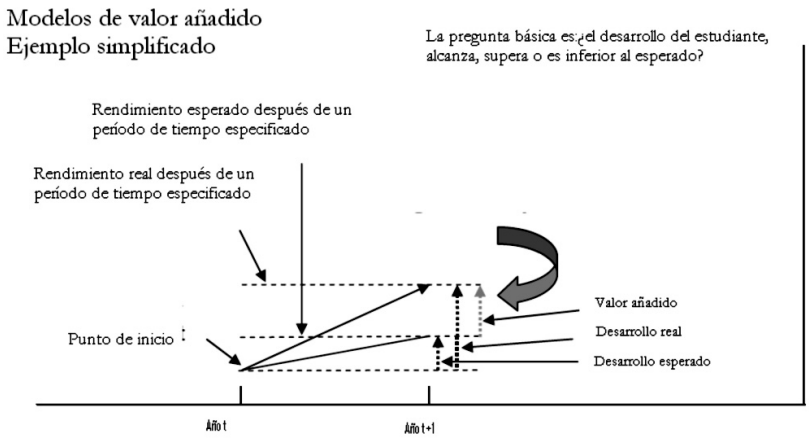
Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

en el rendimiento de los estudiantes puedan atribuirse a las escuelas (o profesores en algunos casos). Intentan responder a la cuestión ¿cuánto valor ha añadido la escuela (o el profesor) al aprendizaje del estudiante? Se diferencian de los modelos de desarrollo simple en que intentan atribuir la ganancia o cambio a las escuelas o profesores. En los modelos de VA el término *efecto* representa normalmente la desviación de la escuela de un resultado potencial esperado. Dentro de la terminología de los modelos multinivel referidos más arriba, son los *residuos no explicados* de la escuela después de ajustar o controlar otros factores o fuentes de variabilidad. Los modelos estadísticos utilizados mediante la partición de la varianza intentan aislar los factores diferentes de las prácticas de la escuela (efectos tipo B) para evaluar su influencia en el aprendizaje o desarrollo de los estudiantes (Drury y Doran, 2004; Hershberg, Simon y Lea-Kruger, 2004; McCaffrey et al., 2003).

Los números que estiman el VA son similares a los residuos de la regresión, ya que representan la parte de los resultados (puntuación promedio del estudiante) que no es explicada por las variables explicativas incluidas en el modelo. Al igual que los residuos, estos números tienen de media 0 y el número ligado a cada escuela particular es interpretado provisionalmente como una medida del rendimiento relativo de la escuela. Son estimadores de la diferencia entre la contribución de la escuela al aprendizaje de sus estudiantes y la contribución promedio al aprendizaje de todas las escuelas de las que se obtuvieron datos. Un valor positivo significa que la escuela parece que ha contribuido más al rendimiento que el promedio, es más efectiva, mientras que un valor negativo que es menos efectiva, aunque los estudiantes de la escuela hayan tenido ganancias brutas positivas durante el período bajo estudio. El modelo ajustado y su éxito al explicar los efectos de las escuelas estará determinado por los datos y las variables empleadas en el modelo, así como por el conjunto particular de escuelas de la muestra. Los diferentes estimadores de efectos están sujetos a diferentes tipos de incertidumbre y sesgo, pudiendo establecerse la incertidumbre en sus errores típicos. La ratio del VA estimado respecto de su error típico puede usarse para ver si es estadísticamente diferente de 0 o valor promedio de todas las escuelas.

La lógica implícita o explícita se basa en la comparación de las ganancias de la escuela con otras similares, que comienzan en los mismos niveles de rendimiento o que sirven a poblaciones similares de estudiantes. En la Figura I se presenta una representación simplificada de los modelos de VA.

FIGURA I. Ejemplo simplificado de un modelo de Valor Añadido



Tipos de modelos estadísticos utilizados en VA

Para la estimación de los efectos de las escuelas pueden utilizarse diferentes modelos estadísticos. La característica fundamental es que incluyen dos o más medidas de rendimiento, pudiendo incluir o no características contextuales de los estudiantes y de las escuelas. Los diferentes modelos parten de distintos supuestos y algoritmos de estimación por lo que los resultados pueden diferir. No obstante, todos tienen en común una serie de características: ser modelos cuantitativos, utilizar las puntuaciones de los estudiantes en tests como medidas de aprendizaje, naturaleza longitudinal con dos o más medidas de los estudiantes e intentar atribuir las ganancias o cambios en desarrollo-aprendizaje a las escuelas o los profesores. Dadas las grandes diferencias entre modelos es difícil unificar la notación. En este apartado seguimos la notación simplificada utilizada en Lissitz et al. (2006). Para un análisis adecuado de la parametrización concreta de los distintos modelos, debe acudirse a las fuentes originales.

La primera distinción entre los modelos es la que se establece entre los efectos como *fijos* o *aleatorios*. En la literatura sobre el diseño experimental se consideran efectos fijos aquellos en los que las inferencias se restringen a los términos específicos incluidos en el modelo. Por el contrario, se habla de efectos aleatorios cuando las unidades se consideran una muestra aleatoria de la población e interesa analizar la variabilidad en la población de la que las unidades son muestreadas. En principio, una aproximación no es

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

necesariamente superior a la otra, sino que está condicionada a las inferencias que se deseen extraer de los datos. La mayor parte de los modelos actuales consideran los efectos de las escuelas como aleatorios, ya que es útil para tratar con estructuras de datos correlacionadas y la heterogeneidad del estatus y cambio de las escuelas.

Modelos univariantes de efectos fijos

En los casos en los que solamente existen puntuaciones de dos años se puede simplificar la estructura multivariante de los datos y tratar los resultados como univariantes. Hay dos aproximaciones, la denominada de ajuste de covariantes y la de la modelización de las ganancias.

En la aproximación de ajuste de covariantes las puntuaciones del año actual se regresan sobre la puntuación del año anterior y sobre otras covariantes adicionales, si se considera necesario:

$$Y_{ti} = \sum_{k=1}^{M} \beta_k (Y_{t-1,i}) + \sum_{l=1}^{J} \beta_l (X_i) + \varepsilon_{ti} \; ; \; \varepsilon_{ti} \sim N(0, \sigma_\varepsilon^2) \qquad (8)$$

Donde Y_{ti} representa la puntuación del estudiante i en el tiempo t, β_k es el efecto de la escuela y los β_l representa el coeficiente del covariante l-ésimo. Este diseño es muy simple, ya que solamente necesita dos puntuaciones de rendimiento que no precisan estar en la misma escala. El valor añadido para el estudiante es la diferencia entre la puntuación actual y la pronosticada y el de la escuela para una materia y curso, el promedio de los residuos de los estudiantes.

Otra aproximación posible es la modelización de la ganancia o de las diferencias entre las puntuaciones de dos años consecutivos de evaluación, $G_t = Y_t - Y_{t-1}$, tomándola como la variable dependiente del modelo de regresión:

$$G_{ti} = \sum_{k=1}^{M} \beta_k (x) + \varepsilon_{ti} \; ; \; \varepsilon_{ti} \sim N(0, \sigma_\varepsilon^2) \qquad (9)$$

Donde β_k es la ganancia promedio de los estudiantes en la escuela k. Al igual que el modelo anterior es muy fácil de estimar, pero requiere que las dos puntuaciones de los tests estén en una escala comparable. El valor añadido para una escuela y materia es el promedio de los resultados.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Estos modelos de efectos fijos han sido usados preferiblemente en el ámbito de las funciones de producción de la economía de la educación (Lockwood y McCaffrey, 2007).

Modelos univariantes de efectos aleatorios

Los dos modelos anteriores pueden tratarse asumiendo los efectos de las escuelas como variables aleatorias con modelos de efectos mixtos. Son más frecuentes que los anteriores. Los resultados son similares a los de efectos fijos en cuanto a la variabilidad de las escuelas, pero proporcionan diferentes estimaciones de los efectos que resultan de formas diferentes de tratar con el error muestral. En los modelos de efectos fijos el efecto de la escuela se estima únicamente a partir de los estudiantes de la escuela; en los modelos de efectos aleatorios se utilizan estimadores empíricos de Bayes que contraen el estimador basado en los estudiantes hacia la media global para todos los estudiantes. Aunque sesgados, tienen propiedades estadísticas óptimas excepto para las escuelas cuyos efectos están lejos de la media. Los de efectos fijos están muy afectados por el error muestral cuando las clases o las escuelas son pequeñas.

La extensión del modelo de ajuste de covariante a los efectos aleatorios tendría la forma siguiente, mostrada en la ecuación (6)

$$\begin{aligned} Y_{ti} &= \mu + \beta(Y_{t-1,i}) + \theta_{j(i)} + \varepsilon_i \\ \theta_{j(i)} &\sim N(0,\sigma_\theta^2); \varepsilon_i \sim N(0,\sigma_\varepsilon^2) \end{aligned} \qquad (10)$$

Donde μ es el parámetro media fijado, β es el aumento esperado en Y asociado con cada cambio unitario en $Y_{t-1,i}$, t indica el tiempo, i el estudiante individual y θ el efecto de la escuela. La notación j(i) indica que el estudiante i está anidado en la escuela j. Los efectos de la escuela son tratados como variación aleatoria en torno a β.

También se puede reformular el modelo de la ganancia, como en la ecuación 11

$$\begin{aligned} G_i &= \mu + \theta_{j(i)} + \varepsilon_i \\ \theta_{j(i)} &\sim N(0,\sigma_\theta^2); \varepsilon_i \sim N(0,\sigma_\varepsilon^2) \end{aligned} \qquad (11)$$

Ahora el efecto de la escuela denota la desviación de la ganancia promedio. En todos los modelos pueden introducirse covariantes adicionales que representen características de los estudiantes.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Puede haber variaciones en los resultados según que se estimen ajustes de covariantes o ganancias (Thum, 2003), ya que puede darse la «paradoja de Lord» (Lord, 1967).

Todos los modelos anteriores comparten algunas limitaciones. Requieren datos completos de los estudiantes, por lo que deben eliminarse sujetos con valores perdidos en alguno de los tests, lo que puede llevar a estimadores sesgados, planteando problemas especiales en escuelas con alta movilidad de los estudiantes y elevadas tasas de repetición. No suelen tener en cuenta de forma explícita los errores de medida, lo que es problemático en los casos de uso de las puntuaciones anteriores como predictores, ya que la regresión asume predictores sin error (McCaffrey et al., 2003). Finalmente, el uso de solamente dos puntuaciones introduce limitaciones en la estimación del desarrollo de los estudiantes (Rogosa, 1995).

Una variante de los modelos anteriores es el desarrollado en el *Dallas Value-Added Assessment System* (Webster y Mendro, 1998; Webster, 2005), que realiza ajustes mediante regresión por mínimos cuadrados ordinarios de las puntuaciones actuales y anteriores, utilizando predictores de características de los estudiantes. Las ganancias o diferencias consideran efectos aleatorios de las escuelas y estos efectos son ajustados en un marco multinivel en el que se consideran de nivel dos y para cuyo ajuste se utilizan variables contextuales de las escuelas.

Otras variaciones incluyen el uso directo de modelos multinivel en los que directamente se modelizan los resultados en un modelo de dos niveles, sin realizar previamente los ajustes de las puntuaciones anteriores y actuales mediante regresión. Los modelos de valor añadido contextualizados de Inglaterra usan este tipo de aproximación.

Los modelos anteriores utilizan como variable dependiente el vector entero de puntuaciones de los sujetos en los tests. Analizan las trayectorias de los estudiantes utilizando para cada sujeto las puntuaciones disponibles de los tests. Suelen incorporar al modelo estructuras complejas de varianza/covarianza de las puntuaciones y permiten incluir diversas trayectorias de desarrollo, tanto lineales como no lineales (Goldstein, 2003; Raudenbush y Bryk, 2002). Pueden darse dos situaciones:

- Modelos completamente anidados, en los que la secuencia de puntuaciones está anidada en los sujetos, y éstos anidados en las clases o en las escuelas, y se denominan modelos anidados de efectos aleatorios, que pueden tratarse con los convencionales modelos multinivel con dos o tres niveles.
- Modelos que permiten analizar diferentes cohortes de estudiantes, la movilidad o cambio de escuelas y el tratamiento de los efectos de escuelas o profesores anteriores.

Existen diferentes procedimientos estadísticos para tratar los modelos anteriores por ejemplo, TVAAS/EVAAS aplica una matriz de covarianza no estructurada a los residuos del estudiante (Ballou, Sanders y Wright, 2004). Otros pueden introducir interceptos y pendientes aleatorias en el nivel del estudiante para explicar esta covarianza (Doran y Lockwood, 2006; Raudenbush y Bryk, 2002).

También existen diseños multivariantes complejos que se suelen diferenciarse en dos bloques, aunque pueden caracterizarse ambos bajo el mismo modelo general (Wright, Sanders y Rivers, 2006).

En primer lugar hay dos modelos cuyo interés radica en la evaluación de los efectos del profesor que son el modelo estratificado *(layered model)* utilizado en los sistemas TVAAS/EVAAS (Ballou et al., 2004; Sanders, 2006; Wright et al., 2006) y el denominado *modelo de la persistencia* (McCaffrey et al., 2003, 2004; Lockwood, 2006). Estos modelos introducen complejidades de cálculo que dificultan su estimación con software convencional.

El segundo bloque, desarrollado en el marco de los modelos multinivel, se conoce como modelo de clasificación cruzada *(cross-classified model)* (Choi, Goldschmidt & Yamashiro, 2006; Choi & Seltzer, 2005; Goldstein, 1987; 2003; Goldstein, Burgess y McConnell, 2007; Ponisciak y Bryk, 2005; Raudenbush y Bryk, 2002).

También se ha propuesto modelos para variables dependientes ordinales (Fielding, Yang y Goldstein, 2003).

Los modelos lineales mixtos utilizados en la mayor parte de las aplicaciones VA son una forma excelente de describir las escuelas efectivas, pero algunos autores encuentran algunos problemas y deficiencias en su utilización, especialmente si se enmarcan dentro de sistemas de rendición de cuentas basadas en estándares como la NCLB. Para responder a la pregunta de si el rendimiento y el desarrollo son adecuados, se han propuesto algunos modelos que están recibiendo cierta consideración en la literatura especializada: sistema REACH (Doran e Izumi, 2004), el modelo híbrido (Kingsbury y McCall, 2006; McCall, Kingsbury y Olson, 2004), las tablas de resultados (Hill et al, 2006) y las matrices de transición (Betebenner, 2005).

Conclusiones

Los modelos de VA son un conjunto de procedimientos estadísticos que han despertado un gran interés entre los investigadores en educación, administradores escolares

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

y políticos, ya que proporcionan medios para separar los efectos derivados de las prácticas de las escuelas de variables de los estudiantes y contextuales. Se consideran un desarrollo muy prometedor para la identificación de las buenas escuelas y de las que necesitan ayuda. Su aplicación mide los efectos de las escuelas (y de los profesores) con mayor precisión y justicia que otros procedimientos utilizados con frecuencia en la evaluación, ya que en el aprendizaje de los estudiantes intervienen numerosos factores, tanto escolares como no escolares.

Durante los últimos diez años se han hecho importantes avances en el desarrollo de estos modelos. El uso de métodos estadísticos complejos ha permitido seguir las trayectorias de rendimiento de los estudiantes con mediciones longitudinales múltiples y la incorporación de medidas de ajuste para separar los efectos verdaderos de la escuela en las ganancias de aprendizaje de los estudiantes.

Los resultados se han mostrado útiles en varias actividades: rendición de cuentas, con o sin consecuencias para las escuelas, elección de escuelas basada en resultados más objetivos que las tablas de liga, y uso de la información en los procesos de mejora y desarrollo de las escuelas, con importantes aplicaciones para los procesos de autoevaluación. En general, sus datos pueden considerarse una buena fuente para el diálogo con las escuelas y la reflexión para emprender acciones de mejora, que mejorarán el rendimiento de los estudiantes.

No obstante, aunque parece haber ganado una extraordinaria popularidad, algunas características metodológicas y de su aplicación en la práctica están todavía bajo escrutinio. En Martínez Arias (2008, en este volumen) se presenta una revisión del estado del arte y de la problemática planteada por los modelos. Cuestiones sobre los supuestos, el tratamiento de los valores perdidos, las diferencias derivadas del uso de uno u otro modelo, la incertidumbre de los estimadores, la dificultad de comunicar los resultados a las partes interesadas de forma comprensible, la necesidad de disponer de tests que permitan la presentación de los resultados en una escala única y la atribución de los efectos a las escuelas a partir de datos observacionales y no experimentales, así como su carácter normativo, que dificulta establecer valoraciones de si el aprendizaje es adecuado, junto con la posibilidad de que rebajen las expectativas para ciertos grupos de sujetos, están bajo discusión en la actualidad.

A pesar de todo, las investigaciones recientes están ayudando a la superación de algunos de los anteriores problemas y los resultados combinados con procesos sistemáticos para la interpretación y aplicación de los datos, pueden ayudar a políticos, administradores, inspectores, directores de centros y profesores a proporcionar una educación de calidad para todos los estudiantes.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Presentación del monográfico

Como hemos presentado a lo largo de estas páginas, los modelos de valor añadido, más que una metodología estadística para tratar los datos procedentes de la evaluación educativa, constituyen una nueva forma de entender el papel que la evaluación desempeña o puede desempeñar en el seno de los sistemas educativos.

Es muy difícil exagerar la importancia que esta estrategia evaluativa va a adquirir en los próximos años en el funcionamiento ordinario de la educación, aunque hoy todavía se trate fundamentalmente de un tema de investigación, al que se incorporan destacadas y bien conocidas experiencias aplicadas.

Todos estos temas y los problemas prácticos y metodológicos asociados son tratados en este número monográfico. En él encontramos diez contribuciones en las que se dibuja el amplio y complejo panorama del valor añadido en educación. Se pueden agrupar en tres grandes bloques temáticos, que muestran la lógica interna de los modelos de valor añadido y la relevancia y vigencia de los mismos.

El primer bloque temático incluye dos artículos. El primero de éstos de los profesores Rosario Martínez Arias, José Luis Gaviria y María Castro (Universidad Complutense de Madrid), se centra en la conceptualización y evolución de los distintos modelos de valor añadido. No se podía comenzar este monográfico sin una adecuada definición de esta medida, mostrando que su dimensión teórica no es ajena a las distintas aproximaciones a la evaluación de la contribución específica de las escuelas a los aprendizajes escolares. De ahí que en esta aportación, además de mostrar los principales modelos de valor añadido en uso, también se repasan los principales retos y dificultades conceptuales y metodológicas de su medición.

Muy vinculado a este artículo, la contribución de la profesora Martínez Arias en el artículo *Usos, aplicaciones y problemas de los modelos de valor añadido en educación* tiene un carácter más metodológico, pues se revisan los modelos estadísticos en uso y se presenta una revisión de los problemas estadísticos, psicométricos y prácticos relacionados con el uso de los modelos de valor añadido para favorecer una interpretación prudente y transparente de los mismos.

El segundo bloque temático se dedica a la descripción y análisis de algunas de las experiencias prácticas vigentes de aplicación de modelos de evaluación de sistemas basados en el valor añadido. Los tres siguientes artículos muestran los sistemas de valor añadido inglés y norteamericano que representan los ejemplos reales de aplicación de estos modelos.

El artículo de Ray, Evans y McCormack muestra la experiencia inglesa desarrollada desde el año 2002 hasta la actualidad con todas las escuelas primarias y secundarias.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

El punto clave de su medición del valor añadido es el desarrollo de una medida sencilla a la vez que técnicamente precisa que es utilizada tanto por las escuelas para el desarrollo de sus propios planes de mejora, como por la Administración educativa inglesa para la introducción y seguimiento de cambios en el sistema educativo que ayuden a elevar los estándares de rendimiento.

El artículo del profesor Thum de la Universidad de Michigan State (USA) recoge una reflexión crítica y profunda, así como un conjunto de pautas de carácter práctico para y sobre el desarrollo de los modelos de valor añadido que pueden acogerse dentro de la legislación norteamericana *No Child Left Behind,* que sin duda ha supuesto una revolución y un estímulo para estos modelos desde 2002.

El último bloque conceptual está relacionado con las *cuestiones metodológicas,* importantes si se quiere desarrollar un sistema de rendición de cuentas basado en modelos de valor añadido, e incluye seis artículos. Si bien es cierto que existe consenso en la literatura de investigación sobre la definición y relevancia práctica de los sistemas de rendición de cuentas de los sistemas educativos, tampoco hay duda de que la discusión científica de los modelos de valor añadido se centra en cómo traducir operativamente las ideas sobre lo que debería ser el valor añadido en la práctica, de forma que se somete a constante escrutinio y mejora el mecanismo técnico subyacente a estos modelos. Sin una adecuada especificación metodológica, no se puede establecer la verdadera utilidad de estos modelos.

El artículo de Eva L. Baker, se centra en la validez de los sistemas de evaluación usados en las escuelas, bien sean éstos formativos o de rendición de cuentas. El inicio de los sistemas de evaluación está en la medida de los logros académicos; de ahí que la autora demande que los tests tengan la denominada *sensibilidad instructiva.*

Vinculados también a cuestiones de medida del rendimiento académico, se incluyen otros dos artículos, relacionados con la necesaria unidimensionalidad de las pruebas para la medida a través del tiempo del valor añadido.

El artículo de los profesores Lizasoain y Joaristi realiza un análisis de la dimensionalidad de un conjunto de pruebas de Matemáticas utilizadas en una evaluación longitudinal del valor añadido realizada con alumnos desde 5º de Educación Primaria hasta 4º de Educación Secundaria Obligatoria de la Comunidad de Madrid, de 2005 a 2007. La clave de este trabajo está en la triangulación de los resultados obtenidos por técnicas factoriales clásicas, por los procedimientos de análisis factorial de información total y por métodos no paramétricos basados en la Teoría de Respuesta al Ítem.

Muy unido a esta temática está el artículo de González, Blanco y Ordóñez. En él, se parte de los patrones de correlaciones que presentan las sucesivas medidas de

rendimiento necesarias para obtener una medida de valor añadido. Cuando se comparan estos patrones con los que aparecen cuando se correlacionan las variables estructurales, se comprueba cómo la fiabilidad y la complejidad estructural de las medidas de rendimiento afectan a esos patrones observables. Una conclusión importante de este artículo es que la creciente complejidad que presentan las medidas de rendimiento académico de cursos sucesivos es más importante, a la hora de explicar los patrones de correlación, que la propia fiabilidad de los instrumentos utilizados.

El concepto de valor añadido y el concepto de crecimiento y cambio en los aprendizajes de los alumnos están íntimamente relacionados. El artículo de Castro, Ruiz de Miguel y López, estudia los efectos del principal predictor de la tasa de crecimiento individual: el nivel inicial de conocimientos de los alumnos. La inclusión de este predictor permite el control de artefactos estadísticos no deseados como la regresión estadística. También se estudian distintas formas de crecimiento distintas a la más utilizada clásicamente que es la lineal. De esta manera, se puede establecer cuál es la forma básica de la función de crecimiento en los modelos de valor añadido.

El artículo de la profesora Ferraõ, de la Universidad da Beira Interior (Portugal), describe el grado en que dos medidas diferentes del estatus socioeconómico, una variable clásica en los modelos de valor añadido, o incluso la exclusión de una variable, producen cambios importantes en las estimaciones del valor añadido de cada escuela.

Por último, el artículo escrito por los profesores Gaviria, Biencinto y Navarro estudia la posibilidad de mantener la misma estructura longitudinal de base con los alumnos de Primaria y Secundaria. La relevancia de este artículo estriba en que es en la transición entre estas dos etapas donde se identifican las mayores variaciones en los indicadores del logro académico, y una estructura común permitiría un análisis de los factores diferenciales que afectan al rendimiento desde la misma estructura de varianza-covarianza.

El conjunto de estos diez artículos representa desde nuestro punto de vista un repaso en profundidad a las cuestiones más candentes relacionadas con el concepto, la medida, los modelos, la metodología y las experiencias aplicadas en el ámbito del valor añadido. Nuestra intención como coordinadores de este volumen ha sido ofrecer al lector un balcón privilegiado y actualizado desde el cual asomarse al complejo y apasionante mundo de la evaluación de los sistemas educativos mediante el uso de modelos de valor añadido. Esperamos que este acercamiento pueda satisfacer y estimular la curiosidad y las diversas necesidades de los lectores de la Revista de Educación.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Referencias bibliográficas

Aitkin, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A,* 149, 1-43.

Ballou, D., Sanders, W. & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics,* 29, 37-65.

Betebenner, D.W. (2005). *Performance standards in measures of educational effectiveness.* Boston College. Department of Educational Research, Measurement and Evaluation.

Braun, H. & Kanjee, A. (2006). *Using assessment to improve education in developing nations*. En H. Braun, A. Kanjee, Bettinger, R. & Kremer, M. (eds.), Improving education through assessment, innovation, and evaluation (1-46). Cambridge, MA: American Academy of Arts and Sciences.

Burnstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education,* 8, 158-233.

Carey, K. (2004). The real value of teachers: Using new information about teacher effectiveness to close achievement gap. *Thinking K-16,* 8, 1-42.

Carlson, K., Martínez, F., O'Day, J., Stecher, B., Taylor, J. & Cook, A. (2007). *State and local implementation of the No Child Left Behind Act. Vol. III. Accountability under NCLB Interim report.* Santa Mónica, CA: The RAND Corporation.

Choi, K., Goldschmidt, P. & Yamashiro, K. (2006). *Exploring models of school performance: From theory to practice* (CSE Rep: No. 673). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Choi, K. & Seltzer, M. (2005). *Modeling Heterogeneity in Relationships between Initial Status and Rates of Change: Latent Variable Regression in a Three-Level Hierarchical Model.* (CSE Rep. No 647). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Choi, K., Seltzer, M., Herman, J. & Yamashiro, K. (2004). *Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP.* (CSE Rep. No 637). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: a guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Cronin, J., Kingsbury, G. G., McCall, M. S. & Bowe, B. (2005). *The impact of No Child Left Behind act on student achievement and growth.* Portland, OR: Northwest Evaluation Association.

De Leeuw, J. & Meijer, E. (2008a). Introduction to multilevel analysis. En J. de Leeuw & E. Meijer (eds.), *Handbook of multilevel analysis* (pp. 1-75). New York: Springer.

— (2008b). *Handbook of multilevel analysis.* New York: Springer.

Doran, H. & Lockwood, J. R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics,* 31, 205-230.

Doran, H. C. & Izumi, L. T. (2004). *Putting Education to the Test: A Value-Added Model for California.* San Francisco, CA: Pacific Research Institute.

Drury, D. & Doran, H. (2003). The Value of Value-Added Analysis. *NSBA Policy Research Brief,* 3, 1-4.

Edmonds, R. R. (1979). Effective schools for the urban poor. *Educational Leadership,* 37, 15-27.

Fielding, A., Yang, M. & Goldstein, H. (2003). Multilevel ordinal models for examination grades. *Statistical Modelling,* 3, 127-153.

Fitz-Gibbon, C. T. (1997). *The value-added national project: Final report: feasibility studies for a national system of value added indicators.* London: School Curriculum and Assessment Authority.

Gaviria, J. L. y Castro, M. (2005). *Los modelos jerárquicos lineales.* Madrid: La Muralla.

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

Goldschmidt, P., Choi, K. & Martinez, F. (2004). *Using Hierarchical Growth Models To Monitor School Performance Over Time: Comparing NCE to Scale Score Results (CSE Rep. No 618).* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Goldstein, H. & Spiegelhalter, D. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society, A,* 159, 385-443.

Goldstein, H. & Thomas, S (1996). Using Examination Results as Indicators of School and College Performance. *Journal of Royal Statistical Society, A,* 159, 149-163.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education,* 19, 425-433.

Goldstein, H. (1987a). *Multilevel models in educational and social research.* New York: Oxford University Press.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

— (2003b). *Multilevel models.* London: Arnold.

Goldstein, H., Burgess, S. & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of The Royal Statistical Society,* A, 170, 941-954.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. & Rasbash, J. (1995). A multilevel analysis of school improvement: changes in schools' performance over time. *School Effectiveness and School Improvement,* 10, 97-114.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education,* 27, 25-68.

Hanushek, E. A. (1979a). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources,* 14, 351-388.

— (2003). The failure of input-based schooling policies. *Economic Journal,* 113, 64-98.

— (2005). The economics of school quality. *German Economic Review,* 6, 269-286.

Hanushek, E. A. & Raymond, M. E. (2004). The Effect of School Accountability Systems on the Level and Distribution of Student Achievement. *Journal of the European Economic Association,* 2, 406-415.

Hanushek, E. A. & Wossmann, L. (2006). Does Early Tracking Affect Educational inequality and performance? Differences-in-differences evidence across countries. *Economic Journal,* 116, 63-76.

Hill, R. & DePascale, C. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practices,* 22(3), 12-20.

Hill, R., Scott, M., De Pascale, C., Dunn, J. & Simpson, M.A. (2006). Using value tables to explicitly value student growth. En R. Lissitz (ed.), *Longitudinal and value added models of student performance* (255-290). Maple Grove, MN: JAM Press.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Jencks, C., Smith, M. S., Ackland, H., Bane, M. J., Cohen, D., Grintlis, H., Heynes, B. & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.

Kane, T. J. & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives,* 16, 91-114.

Kingsbury, G. G. & McCall, M. S. (2006). *The hybrid success model: Theory and practice.* En R. Lissitz (ed.), Longitudinal and value added models of student performance (346-379). Maple Grove, MN: JAM Press.

Ladd, H. F. & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review,* 21, 1-17.

Linn, R. L., Baker, E. V. & Betebenner, D. W. (s.f.). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher,* 31, 3-16.

Linn, R. L. & Haug, C. (2002). Stability of school building accountability scores and gains. *Educational Evaluation and Policy Analysis,* 24, 29-36.

Linn, R. L. (2004). Assessment and accountability. *Educational Researcher,* 29, 4-14.

Lissitz, R., Doran, H., Schafer, W. & Willhoft, J. (2006). Growth modeling, value-added modeling and linking: An introduction. En R. Lissitz (ed.), *Longitudinal and Value-Added Models of Student Performance* (pp. 1-46). Maple Grove, Minnesota: JAM Press.

Lockwood, J. R. (2006). *A case study of some practical challenges of longitudinal student achievement modeling: The RAND Mosaic II Study.* En R. Lissitz (ed.), Longitudinal and Value-Added Models of Student Performance (230-254). Maple Grove, Minnesota: JAM Press.

Lockwood, J. R. & McCaffrey, D. F. (2007). Controling for individual heterogeneity in longitudinal models with applications to student achievement. *Electronic Journal of Statistics,* 1, 223-252.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin,* 68, 304-305.

MacBeath, J. & Mortimore, P. (2001). *Improving school effectiveness.* Buckingham: Open University Press.

Marchesi, A. y Martínez Arias, R. (2006). *Escuelas de éxito en España. Sugerencias e interrogantes a partir del informe PISA 2003.* Madrid: Fundación Santillana.

Martínez Arias, R., Hernández Lloreda, V. y Hernández Lloreda, M. J. (2006). *Psicometría.* Madrid: Alianza.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M. & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Mónica, CA: The RAND Corporation.

McCaffrey, D. M., Lockwood, J. R., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics,* 29, 67-101.

McCall, M.S., Kingsbury, G. G. & Olson, A. (2004). *Individual growth and school success.* Portland, OR: Northwest Evaluation Association.

Mortimore, P., Sammons, P. & Thomas, S. (1994) School Effectiveness and Value Added Measures, *Assessment in Education: Principles, Policy y Practice,* 1, 315-332.

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.

Muijs, R. D. & Reynolds, D. (2000). School effectiveness and teacher effectiveness: some preliminary findings from the evaluation of the Mathematics Enhancement Programme. *School Effectiveness and School Improvement*, 11, 273-303.

No Child Left Behind Act of 2001 (2002). Public Law No. 107-110, 115 Stat. 1425.

Nuttall, D. L., Goldstein, H., Prosser, R. & Rasbash, J. (1989). Differential School Effectiveness. *International Journal of Educational Research,* 13, 769-776

O'Day, J. A. & Smith, M. S. (1993). *Systemic reform and educational opportunity.* En S. H. Fuhrman (ed.), Designing coherent education policy: Improving the system (250-312). San Francisco: Jossey Bass.

Ponisciak, S. M. & Bryk, A. S. (2005). Value-added analysis of the Chicago public schools: An application of hierarchical models. En R. Lissitz (ed.), *Value-Added models in education: Theory and applications* (pp. 40-79). Mapple Grove, MN: JAM Press.

Raudenbush, S. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics,* 29, 121-129.

Raudenbush, S. W. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education,* 59, 1-17.

— (2002). Hierarchical Linear Models (2ª ed.). Thousand Oaks, CA: Sage.

Raudenbush, S.W. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics,* 20, 307-335.

Ray, A., Evans, H. & McCormack, T. (2008). The use of national value added models for school improvement in English schools. *Revista de Educación,* 348.

Reynolds, D. & Creemers, B. (1990). School Effectiveness and School Improvement: a Mission Statement. *School Effectiveness and School Improvement,* 1,1-3

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review,* 15, 351-357.

Rogosa, D. (1995). *Myths about longitudinal research.* En J. M. Gottman (ed.), The analysis of change (3-66). Mahwa, NJ: Erlbaum.

Rowe, K. J. (2000). Assessment, league tables and school effectiveness: Consider the issues and «Let's get real». *Journal of Educational Enquiry,* 1, 73-98.

Rubin, D. B., Stuart, E. A. & Zanutto, E. E. (2004). A potencial outcomes view of value-added assessment in education. *Journal of Educational and behavioural Statistics,* 29, 103-116.

Rumberger, R. W. & Palardy, G. J. (2003). *Multilevel models for school effectiveness research.* En D. Kaplan (ed.), Handbook of Quantitative Methodology for the Social Sciences (235-258). Thousand Oaks, CA: Sage.

Sammons, P. (1996). Complexities in the judgement of school effectiveness. *Educational Research and Evaluation,* 2, 113-49.

Sammons, P., Hillman, J. & Mortimore, P. (1995). *Key Characteristics of Effective Schools: A review of school effectiveness research.* London: Office for Standards in Education.

Sammons, P. & Reynolds, D. (1997). A partisan Evaluation John Elliot of school effectiveness. *Cambridge Journal of Education,* 27, 123-36.

Sanders, W. & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation,* 9, 299-311.

— (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education,* 12, 247-256.

Sanders, W. L., Saxton, A. M. & Horn, S. P. (1997). The Tennessee value-added assessment system: a quantitative, outcomes-based approach to educational assessment. En J. Millman (ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Saunders, L. (1999). A brief history of educational «Value Added»: How did we get to where we are? *School Effectiveness and School Improvement,* 10, 233-256.

Schagen, I. (2006). The use of standardized residuals to derive value-added measures of school performance. *Educational Studies,* 32, 119-32.

Scheerens, J. (2005). The quality imperative. Paper commissioned for the EFA Global Monitoring report 2005. *Review of school and instructional effectiveness research.*

Scheerens, J. & Bosker, R. J. (1997). *The Foundations of Educational Effectiveness.* Oxford: Elsevier.

Stevens, J. (2005). *The study of school effectiveness as a problem in research design.* En R. Lissitz (ed.), Value-Added models in education: Theory and applications (pp. 166-208). Maple Grove, MN: JAM Press.

Taylor, J. & Nguyen, A. N. (2006). An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value? *Oxford Bulletin of Economics y Statistics,* 68, 203-224.

Teddlie, C. & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research.* New York: Falmer Press.

Thum, Y. M. (2002). *Measuring student and school progress with the California API.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

— (2003). Measuring progress toward a goal: estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research,* 32, 153-207.

Wang, L., Beckett, G. H. & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education,* 19, 305-328.

Webster, W. J. & Mendro, R. L. (1997). The Dallas value-added accountability system. En J. Millman (ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.

Webster, W. J. (2005). The Dallas school level accountability model: The marriage of status and value-added approaches. En R. Lissitz (ed.), *Value added models in education: Theory and applications* (pp. 233-271). Maple Grove, MN: JAM Press.

Willms, J.D. & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement,* 2, 209-232.

Wright, S. P., Sanders, W. L. & Rivers, J. C. (2006). *Measurement of academic growth of individual students toward variable and meaningful academic standards.* En R. Lissitz (ed.), Longitudinal and value added models of student performance (385-406). Maple Grove, MN: JAM Press.

Yang, M., Goldstein, H., Rath, T. & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education,* 25, 469-483.

Fuentes electrónicas

Goldschmidt, P., Roschewski, P. Choi, K., Auty, W., Hebbler, S., Blank, R. & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* Washington, DC: Councils of Chief State School Officers, de http://www.ccsso.org/publications/

Hershberg, T., Simon, V. A. & Lea-Kruger, B. (February 2004). Measuring What Matters. *American School Board Journal.* National School Boards Association. Consultado el 25 de octubre de 2007, de http://www.asbj.com/2004/02/0204asbjhershberg.pdf

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives,* 13(33). Consultado el 30 de octubre de 2006, de http://epaa.asu.edu/epaa/v13n33/

National Commission of Excellence in Education (1983). *A Nation at risk.* Washington, DC: Author. Consultado de http://www.ed.gov/pubs/NatAtRisk/index.html

Raudenbush, S. (2004a). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing service. Consultado de http://www.ets.org/research/pic/angoff9.pdf

Sanders, W. L. (2006). *Comparisons Among Various Educational Assessment Value-Added Models.* The Power of Two-National Value-Added Conference, Columbus, OH. Consultado el 14 de junio de 2007, de http://www.sas.com/govedu/edu/services/vaconferencepaper.pdf

Zvoch, K. & Stevens, J. J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives,* 11(20). Consultado el 20 de octubre de 2007, de http://epaa.asu.edu/epaa/v11n20

— (2006). Successive student cohorts and longitudinal growth models: An investigation of elementary school mathematics performance. *Education Policy Analysis Archives,* 14(2). Consultado el 20 de octubre de 2007, de http://epaa.asu.edu/epaa/v14n2/

Dirección de contacto: Rosario Martínez Arias. Universidad Complutense de Madrid. Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid. E-mail: rmnez.arias@psi.ucm.es

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08