Usos, aplicaciones y problemas de los modelos de valor añadido en educación

Uses, applications and problems of educational value-added models

Rosario Martínez Arias

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

Resumen

El uso de modelos complejos de valor añadido (VA) que intentan aislar la contribución de las escuelas y profesores al desarrollo del estudiante es creciente. Este artículo presenta varios aspectos relacionados con el estado actual de los modelos de VA. Los documentos utilizados para la revisión fueron artículos de revistas con revisión por pares, libros y documentos de instituciones interesadas en los modelos de VA. En primer lugar se presentan algunos modelos actualmente en uso operativo ampliamente aceptados. En segundo lugar, se exponen algunos usos de los modelos de VA. Se identifican dos objetivos principales que se pueden beneficiar del uso de los resultados: mejora de la escuela y rendición de cuentas, con referencia a la elección de escuela. Finalmente se revisa la investigación actual. Se presenta una revisión de los problemas relacionados con el uso de los modelos de VA, que llevan a una interpretación prudente de los resultados. Los problemas se centran en tres aspectos: estadísticos, psicométricos y prácticos. Las cuestiones estadísticas son: sensibilidad a los supuestos de los modelos, estructura del modelo, datos perdidos, temporalidad de los efectos, atribución causal e incertidumbre de los estimadores. Las cuestiones psicométricas están relacionadas con la construcción de escalas verticales para medir el progreso. Los problemas prácticos tienen que ver con la temporalización y elección de las medidas de resultados y con algunas críticas frecuentes como la reducción del currículo y la enseñanza

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

del test. Se concluye que una aplicación cuidadosa de los modelos de VA, con transparencia y con información adecuada a las partes implicadas, promete ser útil para la evaluación de las escuelas, especialmente para el diagnóstico y mejora y como ayuda en la planificación de las reformas educativas.

Palabras clave: rendición de cuentas de las escuelas, mejora de la escuela, problemas estadísticos de los modelos de valor añadido (VA), escalamiento vertical, modelos de valor añadido.

Abstract

The use of complex value-added models (VAM) that attempt to isolate the contribution of schools or teachers to student development is increasing. This paper presents an overview of several aspects related to the state of art of VAMs. The documents used for the review were peer-referees articles, books and reports from different institutions interested in VAMs. Firstly, it presents some VAMs currently used operationally and very well received by stakeholders. Secondly, the paper provides an overview of the main uses of VAMs. Two broad objectives are identified that can benefit from the use of results: school improvement and school accountability with references to school choice. Finally, the current active research is revised. It provides an overview of the issues related to the use of VAM and leads to caution in the interpretation of results. The main issues are related to three topics: statistical, psychometric and practical issues. The statistical issues are: sensitivity of estimated effects to model assumptions, model structure, missing data, volatility of estimated effects, causal attribution and uncertainty of estimates. The main psychometric issues are related to the construction of vertical scales to measure educational progress. The practical issues are related to the timing of measures, choice of measures of results and some frequent critics such as the narrowing of curriculum and the teaching to the test.

We conclude that a diligent application of VAM, with transparency and adequate information to stakeholders, holds considerable promise for school assessment, especially for diagnostic and school improvement, and as aid in the planning of educational reforms.

Key Words: school accountability, school improvement, statistical issues of value-added models (VAMs), vertical scaling, value-added models.

Introducción

Los modelos de valor añadido (VA en adelante) han ganado en popularidad tanto en la comunidad de investigadores, como entre los políticos y administradores escolares. Sus defensores creen que esta nueva forma de plantear la evaluación permite comparar la efectividad de las escuelas, aunque atiendan a poblaciones diversas de estudiantes (Drury y Doran, 2004; McCaffrey, Lockwood, Koretz, y Hamilton, 2003) y guiar los cambios educativos necesarios, tanto en el nivel de la escuela, como de las reformas políticas. Los estimadores numéricos que proporcionan permiten aislar mejor que otros métodos las contribuciones de las escuelas al aprendizaje de los estudiantes, separándolas de otros factores. En Martínez Arias, Gaviria y Castro (2008, en este volumen) se presentan el origen y los principales modelos estadísticos, con un importante cuerpo de investigación que los avala. También existen algunos sistemas de evaluación mediante modelos de VA implantados en la práctica desde hace varios años, que ponen de relieve la factibilidad y utilidad de su aplicación.

No obstante, algunas revisiones (Bock, Wolfe y Fisher, 1996; Kuppermintz, 2003; Mccafrey, et al., 2003), aunque reconocen el importante avance que suponen, sugieren cautelas frente a una aplicación poco crítica, especialmente si los resultados implican consecuencias para profesores y escuelas (Braun, 2005b). Es decir, los modelos de VA, aunque útiles, no son la panacea que permite atribuir de forma inequívoca los resultados del aprendizaje a las escuelas o profesores. Ningún modelo estadístico, ni ningún método de análisis, por sofisticado que sea, puede compensar completamente la falta de aleatorización de los estudios de observación característicos de la investigación sobre los efectos de las escuelas (Braun, 2005a). No obstante, cuando los resultados se interpretan de forma adecuada, como medidas descriptivas, pueden ser de gran utilidad para la evaluación.

Hay algunos modelos que llevan varios años implantados de los que se pueden extraer lecciones sobre los usos potenciales y que mencionaremos a continuación. Podríamos decir que del análisis de estos modelos y de diferentes estudios, se presenta una breve revisión sobre sus usos potenciales. Por otra parte, su aplicación supone el uso de una compleja metodología estadística, enredadas cuestiones de interpretación y una serie de supuestos clave que deben tenerse en cuenta para su aplicación con la «debida diligencia» (Braun, 2005a). En el artículo también se presentan los principales problemas que son objeto de investigaciones actuales sobre el uso estos modelos que deben considerarse para una adecuada aplicación. Los problemas se han agrupado en tres bloques: estadísticos, medida de los resultados mediante tests y prácticos.

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

Algunos modelos de VA ya han sido aplicados con éxito desde hace años para la monitorización de los rendimientos educativos y para la evaluación de escuelas y profesores. En general han sido valorados positivamente y han tenido una buena aceptación por las partes interesadas (Rose y Gallup, 2007).

Entre los modelos con mayor implantación en la práctica deben mencionarse los siguientes:

- Tennessee Value Added Assessment System (TVAAS, actuamente EVAAS) es el primero en utilizar esta denominación y también en ser utilizado por un estado entero. Ha sido objeto de más estudios y análisis (Amrein-Bearsley, 2008; Ballou, 2002; Bock,Wolfe y Fisher, 1995; Braun,2005a,b; Braun y Wainer, 2007; Kuppermintz, 2003; Mccaffrey, Lockwood, Koretz y Hamilton, 2003). Fue desarrollado por el profesor William Sanders y sus colaboradores (Sanders y Horn, 1994; Sanders, Saxton y Horn, 1997). En la actualidad la aplicación del sistema en los estados que lo requieren se lleva a cabo desde la empresa SAS (que desarrolla el software estadístico), bajo la dirección del profesor Sanders y se denomina *Education Value-Added Assessment System* (EVAAS). Permite obtener perfiles de las escuelas, estableciendo los resultados por quintiles de rendimiento. El método EVAAS proporciona además la denominada *«metodología de proyección»* que proporciona estimadores del nivel de rendimiento del estudiante individual en algún punto futuro, bajo el supuesto de que tendrá una experiencia de escolarización promedio. Wright, Sanders y Rivers (2006) presentan un estudio con dicha metodología. El modelo estadístico es un modelo de los clasificados como de efectos mixtos multivariante y complejo (Martínez Arias et al. 2008), el modelo estratificado. Permite el tratamiento de los datos perdidos, utilizando para cada estudiante todos los datos disponibles. El modelo es muy eficiente en el sentido de que hace uso de toda la información para una cohorte de estudiantes en un período de cinco años. La utilización de tantas medidas y de varias materias minimiza los efectos del error de medida y de la fugacidad o temporalidad de los resultados. No utiliza ajustes de otras covariantes. Esta cuestión ha sido sometida a análisis empíricos y no se ha encontrado como universalmente válida (McCaffrey et al., 2003). No obstante, Ballou, Sanders y Wright (2004) mostraron la posibilidad de inclusión de estos ajustes en el modelo.
- Dallas Value Added Assessment System (Modelo DVAAS): El modelo comenzó en 1984. Parte de una clara filosofía de la equidad, que exige tener en cuenta

para la evaluación las diversas características de los estudiantes y del contexto de la escuela. Desde sus orígenes cuenta con la *Accountability Task Force,* un comité en el que están integradas todas las partes interesadas. Una buena descripción del modelo de evaluación puede encontrarse en Webster y Mendro (1997) y Webster (2005). El modelo sigue longitudinalmente a los estudiantes y utiliza un análisis de datos en dos estadios para los cálculos. En el primer estadio se ajustan las puntuaciones de los dos cursos comparados por variables de los estudiantes, utilizando modelos de regresión de mínimos cuadrados ordinarios. Y los residuos o parte no explicada de las puntuaciones son la base de los cálculos de VA de los profesores y de la escuela seguidos en el segundo estadio, utilizando un modelo lineal jerárquico que controla variables de la escuela.

- Valor añadido de las escuelas públicas de Chicago: El modelo denominado de «productividad» de las escuelas públicas de Chicago evalúa los resultados en competencia lectora y matemáticas de los estudiantes del consorcio entre los cursos 2º y 8º. Se utilizan datos longitudinales de todos los estudiantes para evaluar las ganancias de los estudiantes a lo largo de su permanencia en el sistema educativo. Tiene en cuenta el nivel inicial de los estudiantes y las tendencias de rendimiento. Clasifica las escuelas por medio de un «perfil de productividad». El modelo está descrito en Bryk, Thum, Easton y Luppescu (1998) y es analizado como un modelo multinivel con tres niveles: el desarrollo del estudiante, el estudiante y la escuela. El modelo jerárquico lineal con tres niveles, aunque permite el tratamiento de sujetos con datos perdidos, requiere un anidamiento completo de los estudiantes en la misma escuela para todos los años evaluados, lo que provoca datos perdidos especialmente en las escuelas con alta movilidad. Este hecho les llevó a la introducción de un modelo de clasificación cruzada en los últimos años, lo que permite incorporar la movilidad de los estudiantes. Una descripción del modelo puede encontrarse en Ponisciak y Bryk (2005).
- El modelo de Valor Añadido Contextualizado en Inglaterra: En este volumen Ray, Evans y Mc Cormack (2008) describen la evolución de la experiencia en Inglaterra hasta la actualidad.
- Algunas experiencias españolas: En España, con escasa tradición de evaluación externa de los resultados de los estudiantes, se han desarrollado algunos intentos de evaluación longitudinal con modelos de VA. La experiencia pionera fue la llevada a cabo desde 1997 por el Instituto IDEA a través del Equipo REDES para la evaluación de la Educación Secundaria Obligatoria. Esta experiencia

estuvo limitada a un número reducido de centros voluntarios, en su mayor parte de la Comunidad de Madrid. Una descripción de la experiencia y los principales resultados puede encontrarse en algunos trabajos de miembros del equipo REDES (Marchesi, Martín, Martínez Arias, Tiana y Moreno, 2002; Marchesi, Martínez Arias y Martín, 2004). Una experiencia mucho más ambiciosa y rigurosa es la desarrollada por el equipo investigador de la Facultad de Educación de la Universidad Complutense, dirigido por el profesor Gaviria en la Comunidad de Madrid, con un gran número de centros seleccionados de forma aleatoria. No entramos aquí en su descripción por describirse la experiencia en varios artículos de este volumen (Castro, Ruíz de Miguel y López, 2008; Gaviria, Biencinto y Navarro, 2008; Lizasoaín y Juaristi, 2008).

Principales usos de los modelos de VA

Las experiencias realizadas hasta el momento permiten hablar de diferentes usos de los resultados de los modelos de VA. Básicamente pueden agruparse en dos grandes bloques: la rendición de cuentas y la mejora y desarrollo de las escuelas. Los distintos modelos de VA pueden servir para los dos objetivos, pero dado que las connotaciones y consideraciones vinculadas a cada uno de ellos son muy diferentes, conviene establecer claramente los objetivos antes de la introducción de un modelo. Al final del apartado se presenta una breve guía de consideraciones prácticas para un uso fructífero de los modelos.

Rendición de cuentas (RC)

En los últimos años, la adopción de sistemas de rendición de cuentas es frecuente en muchos países, dentro de la tendencia general a la evaluación de resultados en el sector público. El objetivo suele ser la comparación del uso de recursos, resultados y productividad de las diversas instituciones que reciben financiación pública. Los diversos sistemas utilizados pueden tener diferentes objetivos y grados de ambición, que van desde simplemente proporcionar información al gobierno sobre los resultados, servir para identificar buenas prácticas en las instituciones, proporcionar información al público y que los usuarios de los servicios puedan elegir sobre bases informadas,

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

hasta la aplicación de incentivos directos vinculados a recompensas y sanciones a las instituciones y sus empleados. Sea cual sea la finalidad, conviene tener en cuenta que la rendición de cuentas en el sistema educativo no es un fin en sí misma, sino un medio de lograr las metas educativas. Aunque la rendición de cuentas de una forma u otra es muy antigua, en la actualidad se ha cambiado el foco de atención. En otras épocas se ponía el acento en medidas de recursos y procesos tales como: el control del número de días de instrucción, el tamaño de la clase, las credenciales del profesorado, etc., mientras que los sistemas actuales lo ponen en las medidas de resultados (Hanushek y Raymond, 2004). Un sistema de rendición de cuentas basado en resultados necesita disponer de indicadores simples y equitativos que permitan diferenciar entre escuelas. Los resultados pueden medirse de diferentes formas, pero los modelos de VA utilizan como medida las puntuaciones en tests estandarizados, que juegan un papel dominante debido a su eficiencia de coste-objetividad.

En Martínez Arias et al. (2008) se presentan algunas formas de rendición de cuentas basadas en resultados, especialmente los derivados de la NCLB en Estados Unidos. Los modelos VA pueden usarse para identificar casos extremos, es decir, escuelas que obtienen resultados significativamente mejores o peores de lo esperado. Estas escuelas pueden investigarse posteriormente en relación con sus prácticas docentes, clima escolar, etc. y la información proporcionada puede ser útil para el sistema.

Algunos modelos de VA también evalúan los efectos del profesorado, no obstante, la mayoría de los autores están de acuerdo en que los efectos derivados de los modelos de VA son un indicador muy imperfecto del efecto del profesorado y que en todo caso serán uno de entre múltiples indicadores de su efectividad (Ballou, 2002; Braun, 2005b; McCaffrey et al., 2003).

También existe consenso en que los modelos deben usarse con extraordinaria cautela cuando sus resultados se utilizan en la toma de decisiones con consecuencias para las escuelas o profesores (Ballou, 2002; Braun, 2005b; Kupermintz, 2003; McCaffrey et al., 2003; Raudenbush, 2004a). Los instrumentos y resultados no son lo suficientemente precisos ni fiables para este propósito. Este uso además puede generar fuertes recelos en los profesores y desanimarles a usar la valiosa información que pueden proporcionar los modelos de VA para la mejora de la instrucción (Yeagley, 2007).

Relacionado con la rendición de cuentas está el tema de la publicación de los resultados para facilitar la *elección de escuela*, aunque no suponga incentivos en términos de recompensas o sanciones. Este uso supone proporcionar información a las familias y al público, del rendimiento de las diferentes escuelas para ayudarles en la decisión de la elección.

Los partidarios de la publicidad de los datos y de la elección de escuelas consideran que estas prácticas son un incentivo para proporcionar mejores servicios por parte de las escuelas y elevar el rendimiento de los estudiantes (Glenn y de Groof, 2005; Hoxby, 2000; 2003), entre otras cosas por acoplarse mejor a sus necesidades. Fuchs y Wossman (2007), analizando datos procedentes de evaluaciones internacionales en el nivel de país encontraron que diversas formas de rendición de cuentas, junto con la autonomía y la elección de escuela se relacionaba con niveles más altos de rendimiento de los estudiantes.

En cualquier caso, independientemente de las ventajas y desventajas de la elección, es muy importante cuando se presentan públicamente los resultados de las escuelas que estén en términos de VA contextualizado.

Los resultados del VA para la mejora y desarrollo de la escuela

Se han documentado diversas finalidades para las que las escuelas pueden aplicar los datos de las evaluaciones para la mejora de sus resultados (Supovitz & Klein, 2003). Los datos deben dar respuesta a diversas preguntas, que pueden ejemplificarse en las siguientes: ¿cómo lo estamos haciendo?, ¿proporcionamos una buena educación a los estudiantes? y ¿estamos proporcionando una buena educación a todos los estudiantes? Para poder responder se necesita disponer de información que permita hacer un análisis detallado, comparando el desarrollo de los alumnos en el tiempo, para diferentes asignaturas y grupos de alumnos. Los datos pueden usarse por las escuelas como parte de sus procesos de *auto-evaluación* para diagnosticar las fuerzas y debilidades del centro (Saunders, 2000). En este sentido los resultados de la aplicación del VA pueden ayudar a plantear preguntas sobre las ganancias de los alumnos y a estimular las discusiones informadas entre el personal de los centros sobre la forma de organizar e impartir la enseñanza. Para que estos procesos funcionen bien deben tratarse como innovaciones educativas, proporcionando la preparación adecuada, apoyo y formación para administradores escolares, inspectores, equipos directivos y profesorado.

Dentro de los usos para la mejora de las escuelas es importante destacar el apoyo que los resultados del VA pueden proporcionar a los inspectores mejorando la eficacia de este servicio. Los inspectores, al disponer de esta valiosa información sobre escuelas y estudiantes, pueden guiarles mejor, poniendo el acento en cuestiones clave de la escuela.

Recomendaciones para el uso adecuado de los modelos de VA

La experiencia ha puesto de relieve usos de los modelos de VA que permiten elevar los rendimientos de los estudiantes. No obstante, para que la experiencia sea fructífera, conviene tener en cuenta una serie de recomendaciones que brevemente se presentan a continuación:

- Clara formulación de los objetivos de la implantación del modelo de VA, ya que diferentes objetivos pueden condicionar el modelo a desarrollar.
- Disponer de datos recogidos sobre los alumnos individuales, con muestras amplias y representativas. Estos datos deben recogerse mediante tests estandarizados externos (estatales o regionales para permitir una adecuada comparación).
- Es conveniente recoger datos de resultados que reflejen todos los niveles de rendimiento del alumnado, no limitándose como en algunos casos de la aplicación de la NCLB a los resultados en torno al nivel de «competente».
- Es imprescindible disponer de varias medidas del rendimiento, y si es posible también del rendimiento en el momento de ingreso en la escuela. En general, como se justifica más adelante, la mayoría de los autores recomiendan la utilización de al menos tres medidas. Como es obvio, es preciso disponer de un identificador único de cada estudiante que se mantiene durante su permanencia en el sistema y que permita vincularlo a la escuela (y al profesor, si el objetivo es evaluar profesorado).
- Aunque como se verá más adelante no existe un consenso generalizado al respecto, es importante disponer de información contextual de los alumnos y de la escuela.
- Utilizar procedimientos de análisis estadístico multinivel, ya que son los únicos que permiten tener en cuenta las dependencias intra-escuela. Estos deben ser preferiblemente del tipo que se ha definido como multivariante (Martínez Arias et al., 2008), ya que permiten analizar los cambios a lo largo del desarrollo y son más eficientes para el tratamiento de valores perdidos.
- Disponer de adecuadas bases de datos y de procedimientos de gestión y control de calidad de las mismas.
- Es imprescindible proporcionar una formación adecuada sobre la interpretación de los datos y sus posibles usos a los equipos directivos de los centros, profesorado y otras partes interesadas.

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

- Finalmente, debe desarrollarse un sistema de reportado de los resultados claro y transparente. Se recomienda representación tabular, gráfica y verbal, indicando claramente que diferencias son estadísticamente significativas, para no atribuir significados indebidos a las ordenaciones de las escuelas, cuando se establecen. Suele ser conveniente reportar varios tipos de resultados: brutos o no ajustados, ajustados por características del alumnado, como el rendimiento anterior y resultados ajustados por factores contextuales del alumno y de la escuela. También es muy importante en los informes desagregar los resultados por grupos de alumnos, para analizar específicamente sus dificultades y sus ganancias de aprendizaje.

Las cuestiones estadísticas no resueltas

La aplicación de los modelos de VA requiere de una compleja metodología estadística con fuertes supuestos. En la actualidad muchas cuestiones estadísticas están siendo objeto de investigación y de discusiones entre los científicos. En este apartado se presenta el estado del arte sobre las principales cuestiones objeto de investigación y debate: selección del modelo, la fugacidad o temporalidad de las puntuaciones, ajuste de variables extraescolares, tratamiento de los casos perdidos, la incertidumbre de las estimaciones, cumplimiento de los supuestos de los modelos y las atribuciones causales.

Las diferencias derivadas del uso de diferentes modelos

Los modelos de VA difieren considerablemente en términos de su complejidad, demandas de datos y facilidad para la comunicación de los resultados. Esta pluralidad de modelos implica que una de las primeras decisiones que se deben tomar es sobre el modelo que se utilizará. Dentro de la elección de modelo, la primera decisión será entre modelos de efectos fijos y de efectos aleatorios. Las aplicaciones derivadas de la aproximación de la función de producción de la economía utilizan sobre todo modelos de efectos fijos (Lockwood y McCaffrey, 2007). En la línea derivada de la evaluación de la efectividad de las escuelas, la mayor parte de los autores optan por la aproximación de efectos mixtos, que tratan los efectos de la escuela como aleatorios

(Ballou, Sanders y Wright, 2004; Sanders, Saxton y Horn, 1997; Goldstein, 2003; McCaffrey, Lockwood, Koretz, Louis y Hamilton, 2004). La investigación ha puesto de relieve que los efectos fijos son muy sensibles a los errores de muestreo, puesto que las escuelas y sobre todo los profesores suelen tener un número reducido de estudiantes lo que lleva a grandes errores muestrales y a la inestabilidad de los estimadores de años sucesivos. Parece más aconsejable la utilización de modelos mixtos de efectos aleatorios, que utilizan estimadores empíricos de Bayes. Estos estimadores tienen la propiedad de ser BLUP *(Best Linear Unbiased Prediction)* y contraen *(shrink)* los estimadores de los efectos de escuelas y profesores hacia la media global según el error o fiabilidad del estimador, consiguiendo una mayor estabilidad de las estimaciones de diferentes años. El problema es que la estimación es más compleja y puede ser más difícil explicarlos a personas sin formación técnica.

Dentro de la aproximación de los modelos mixtos, también existen diversas opciones que difieren en complejidad. La investigación sobre diferencias de resultados derivadas del uso de distintos modelos es todavía escasa y no es concluyente (Ponisciak y Bryk, 2005; Sanders, 2006; Tewke et al., 2004; Wright , 2004; Wright, Sanders y Rivers, 2006).

Cuando la serie de datos es limitada (puntuaciones de dos años) surge la duda de si tratarlos mediante aproximaciones univariantes o modelizar el vector de puntuaciones; parece que la segunda aproximación es más flexible, aunque implique una mayor complejidad en los cálculos.

Las variables utilizadas en el ajuste

Los modelos de VA permiten ajustar las puntuaciones eliminando los efectos de otros factores extraescolares, que se confunden con los efectos de los profesores y de las escuelas.

Existe un consenso generalizado en que se deben utilizar los niveles de entrada o iniciales de los estudiantes en los ajustes de los modelos de VA, ya que el rendimiento anterior es el mejor predictor del rendimiento futuro (Gray, Jesson, Goldstein, Hedges y Rasbash, 1995; Sammons, Thomas y Mortimore, 1997). En lo que no existe acuerdo es en si se deben ajustar los efectos por otras variables contextuales y en este último caso, qué variables se utilizarán. La mayor parte de los investigadores abogan por el ajuste de dichas variables, aunque el modelo de mayor tradición (TVAAS/EVAAS) lo considera innecesario porque al utilizar múltiples puntuaciones y de varios años, el estudiante sirve como su propio control (Ballou et al., 2004). Parece que la mayor parte de los modelos estadísticos obtienen resultados bastante parecidos

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

en lo que se refiere a la ordenación de las escuelas (Schafer, Yen y Rahman, 2000; Webster, 2005), pero lo que parece establecer diferencias es la inclusión o no de variables de ajuste y qué variables (Tewke et al., 2004).

No obstante, la postura mayoritaria es la de incluir variables de ajuste, especialmente cuando se establece alguna forma de rendición de cuentas o de difusión de los resultados, ya que la equidad será cuestionable si no se tienen en cuenta características contextuales de los estudiantes y de las escuelas (McCaffrey et al., 2003; McCaffrey et al., 2004).

Partiendo de la necesidad de realizar ajustes, no existe acuerdo sobre qué variables son las importantes, y aún en el caso de algunas consideradas como tales, puede faltar la información adecuada en las bases de datos escolares. La medida de estatus socioeconómico más frecuente de los estudiantes es el disfrute de beca de comedor *(free lunch),* (Ballou et al, 2004; Braun, 2005b; Goldstein, Burgess, y McConnell, 2007; McCaffrey et al., 2004; Sammons et al., 1997), que es una medida bastante imperfecta del estatus, pero frecuentemente es la única disponible. Otras variables presentes en los registros escolares como género, inmigración, etnia, barreras lingüísticas o dificultades de aprendizaje son importantes para el ajuste. La inclusión de dichas variables es importante para analizar el progreso de diferentes subgrupos.

En cuanto a la inclusión de variables contextuales de la escuela, los partidarios de los ajustes lo consideran conveniente. Estas pueden ser agregados de variables de los estudiantes como puntuaciones medias en algunos tests u otras variables definidas en el nivel de la escuela. Entre las más habituales se encuentran: porcentajes de diversos grupos étnicos de interés, estatus socioeconómico de la comunidad, habitat, porcentaje de estudiantes con beca de comedor (utilizado como indicador de pobreza), titularidad de los centros, porcentaje de estudiantes con necesidades educativas especiales, porcentaje de profesores con alta cualificación, niveles de asistencia, etc.

Hay pocos estudios en los que se comparen modelos con diferentes variables de ajuste sobre los mismos datos, pueden consultarse los trabajos de Choi, Goldschmidt y Yamashiro (2006) y de Stevens (2005).

La fugacidad o temporalidad de las puntuaciones ganancia

Cuando los efectos de la escuela se calculan anualmente sobre la base de dos puntuaciones suele haber importantes fluctuaciones en los resultados, lo que ha llevado a advertir sobre el problema de la fugacidad de las ganancias y de las estimaciones

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

de VA (Kane y Steiger, 2002; Linn y Haug, 2002; Rock, 2007; Way, 2006). El problema se ha explicado con frecuencia por los problemas de la baja fiabilidad de las puntuaciones diferencia (Ballou, 2002; Linn y Haug, 2002) que aumentan los errores de medida. El error de medida y la baja fiabilidad también puede provocar efectos de regresión a la media.

El problema suele paliarse con los modelos multinivel multivariantes con varias mediciones o promediando varias materias (Drury y Doran, 2003; Raudenbush, 2004b; Singer y Willett, 2003), con lo que se aumenta la fiabilidad.

Otros factores responsables de la fluctuación ligados a las medidas de resultados son los cambios en el instrumento de evaluación y el marco (Lockwood y McCaffrey, 2007) y la presencia de efectos techo y suelo en los instrumentos por tener ítems de diferentes cursos o niveles (Rock, 2007).

Los cambios en las ganancias estimadas también se ven afectados por cambios en las variables contextuales y de los estudiantes, así como por el tamaño de las escuelas y cambios en la muestra de escuelas participantes, dado que los resultados siempre se expresan con relación a la muestra. Uno de los aspectos mas estudiados ha sido el del tamaño de las escuelas y la mayor variabilidad de las escuelas pequeñas. Este problema se resuelve mejor con el uso de modelos multinivel de efectos aleatorios, ya que su procedimiento de estimación utilizado contraen los estimadores de las escuelas pequeñas, menos fiables, hacia los valores próximos a la media, teniendo en cuenta explícitamente la incertidumbre.

El tratamiento de los datos perdidos

Los modelos de VA requieren muchos datos de los estudiantes. Todos se basan en puntuaciones de los estudiantes en una o más materias durante dos años o más y aquellos que realizan ajustes por características de los estudiantes, necesitan además datos en estas variables. En la práctica, suele haber muchos registros incompletos, lo que produce problemas de datos ausentes en el análisis. Los modelos univariantes requieren la eliminación de todos los casos con datos incompletos; los multivariantes que analizan el vector completo de puntuaciones permiten la inclusión de sujetos con datos perdidos. También se pueden aplicar técnicas de imputación de los valores perdidos antes de los análisis. Sea cual sea la aproximación seguida, la presencia de datos perdidos puede producir importantes sesgos en las estimaciones de los efectos de la escuela.

En la literatura se habla de datos perdidos completamente al azar (MCAR) o al azar (MAR), que en los modelos de VA reflejarían la independencia entre el patrón de los casos perdidos y sus puntuaciones en los tests (Little y Rubin, 2002). Cuando se eliminan los sujetos con datos incompletos, las estimaciones estarán sesgadas a menos que los datos perdidos sigan un patrón MCAR. En los modelos que utilizan todos los casos como TVAAS/EVAAS, parece que no se introducen sesgos si los datos son MAR, lo mismo que cuando se utilizan técnicas de imputación. El problema es que no está claro si los patrones MCAR y MAR se mantienen en los sistemas escolares en los que se aplican los modelos de VA.

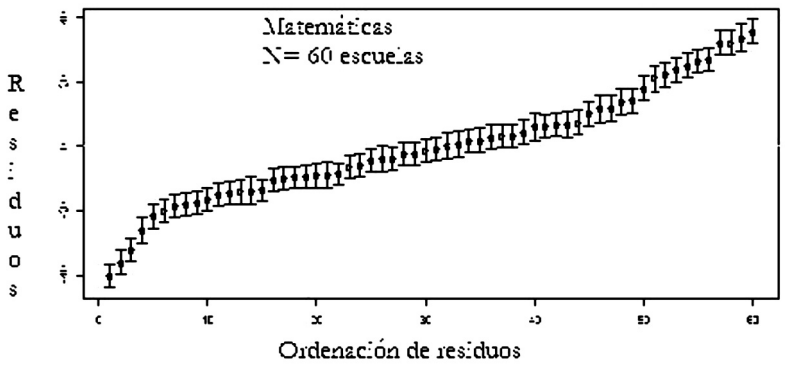
Los investigadores están divididos en cuanto a los efectos de los datos ausentes en las estimaciones de VA, aunque parece que el impacto puede ser importante (Marchesi y Martínez Arias, 2002; Kuppermintz, 2003; McCaffrey et al., 2003; Marchesi et al., 2004; Braun, 2005a; Zvoch y Stevens, 2005a).

Las incertidumbres derivadas del error muestral

Los modelos de VA producen estimadores de los efectos de la escuela en términos del residuo medio, con valores positivos y negativos y media de 0. Estos efectos deben ir acompañados de su varianza y error típico de estimación para una adecuada interpretación. La varianza estimada de un efecto es una medida del grado de incertidumbre que afecta al estimador y que puede llevar a estimaciones imprecisas (McCaffrey et al., 2003). Esta varianza está determinada en gran medida por el modelo de VA utilizado y por la cantidad de datos disponibles de la escuela o del profesor. Si las varianzas son pequeñas en relación a la amplitud de los efectos, las escuelas «atípicas» pueden identificarse fácilmente y es probable que haya más escuelas estadísticamente diferentes de la escuela promedio. No obstante, existe mucha incertidumbre con las escuelas próximas a los valores medios, lo que hace cuestionable el uso de los modelos de VA para la ordenación de las escuelas. Se recomienda representar los resultados de las escuelas ordenados en un gráfico con bandas de error (Goldstein y Healy, 1995) como el representado en la Figura I, para recoger la incertidumbre de los estimadores y no sobreinterpretar las diferencias.

En la figura, las bandas de error reflejan los intervalos de confianza en torno al efecto y solamente se puede hablar de diferencias estadísticamente significativas entre dos escuelas cuando no se solapan sus bandas. La reducción del error típico es muy importante para poder discriminar entre escuelas y se facilita con conjuntos más

FIGURA I. Residuos (valor añadido) de las escuelas con bandas de confianza



amplios de puntuaciones de los tests y con modelos que hacen un uso más eficiente de los datos. Ballou (2005) encontró bastante imprecisión en la estimación de los efectos del profesor.

Las atribuciones causales de los efectos de las escuelas

La meta de los modelos de VA es establecer inferencias sobre los efectos de las escuelas o de los profesores (Raudenbush y Willms, 1995, revisado en Martínez Arias et al., 2008). Los modelos introducen una compleja maquinaria estadística para separar dichos efectos de los innumerables factores de confundido que afectan a los resultados y ganancias. Una de las críticas frecuentes es que la palabra efecto puede tener una connotación causal, en el sentido de que son las escuelas o los profesores la causa de los resultados encontrados en sus estudiantes (McCaffrey et al., 2003; Rubin, Stuart y Zanutto, 2004). Aunque la inferencia causal es la meta última y en la que estarían interesados los políticos y administradores escolares, es evidente que no se puede extraer este tipo de interpretaciones a partir de estudios observacionales como los utilizados en los modelos de VA (Kuppermintz, 2003; Raudenbush, 2004b; Rubin et al., 2004). El reto de la estimación de los efectos es la comparación entre un resultado observado y un resultado potencial. El resultado observado es la puntuación o la ganancia de un estudiante que ha asistido a una escuela j. El resultado potencial es el que se habría observado si el mismo estudiante hubiese sido asignado a una condición alternativa j'. Evidentemente, el estudiante solamente es observado en la condición

j y el problema de los modelos de VA es encontrar el resultado potencial para dicho estudiante. Este resultado ausente suele reemplazarse por el promedio de otras escuelas a las que hubiera podido asistir el estudiante. Este resultado podría ser un adecuado «contrafactual» si la asignación de estudiantes y profesores a las escuelas se realizase de forma aleatoria, de modo que todas estuviesen sirviendo a poblaciones similares, lo que es un supuesto falso. Las inferencias causales solamente pueden extraerse de experimentos aleatorizados con adecuados grupos de control, ya que la aleatorización y grandes muestras reduce la probabilidad de las influencias de las múltiples variables de confundido. El problema de los modelos de VA es que no disponen de diseño experimental que permita evaluar efectos causales. Interpretar un estimador estadístico derivado de un estudio observacional sin aleatorización como una medida directa de un efecto causal es cuestionable (Rosenbaum, 2002; Raudenbush, 2004b; Rubin et al. 2004; Stuart, 2007), ya que puede haber numerosas influencias fuera del control de la escuela. Los modelos de VA intentan captar las virtudes de los experimentos aleatorizados cuando no se han realizado, ya que los datos de la escuela provienen de un estudio de observación y no de un experimento. Son un intento de compensar la ausencia de aleatorización, ajustando por características de los estudiantes y contextuales, estableciendo supuestos de que las covariantes que influyen en los resultados captan adecuadamente las diferencias entre condiciones. El investigador identifica un residuo que permanece en la escuela después de ajustar otras variables que pueden explicar los resultados. Para alejar la tentación de interpretación causal, algunos autores prefieren llamarlos «efectos residuales de la escuela/profesor» (Fitz-Gibbon, 1997; Schatz, Von Secker y Alman, 2005). Por muy complejo que sea el modelo estadístico, no puede compensar la falta de aleatorización. La investigación en ciencias sociales está repleta de ambigüedades en la interpretación, reflejadas claramente en la paradoja de Lord (Lord, 1967). Por este motivo existe un acuerdo generalizado en huir de las atribuciones causales, poniendo el acento en los aspectos descriptivos proporcionados por los estimadores (Raudenbush, 2004b; Rubin et al., 2004; Thum, 2006), aunque estas descripciones son muy útiles para hacer proposiciones sobre el progreso de los estudiantes y de las escuelas.

El cumplimiento de los supuestos de los modelos

La mayor parte de los procedimientos estadísticos utilizados en los modelos de VA son procedimientos paramétricos que requieren el cumplimiento de numerosos supuestos que condicionan la seguridad de los estimadores. En la realidad los supuestos

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

nunca se cumplen completamente y ningún modelo es perfectamente adecuado. La cuestión es determinar la dirección y la magnitud de los sesgos frente a su violación. Los supuestos pueden referirse a la naturaleza de los datos, la estructura del modelo o a ambos. Por lo que se refiere a la estructura de los datos, al igual que en los modelos de regresión suele asumirse la normalidad y la homocedasticidad. También es importante la ausencia de efectos techo y suelo en los datos, ya que su presencia puede conducir a distribuciones asimétricas.

El hecho de trabajar con puntuaciones de tests que presentan errores de medida, puede afectar también a los estimadores. En los modelos en los que se ajustan covariantes como predictores, sean puntuaciones de los tests de años anteriores o de otras variables, un supuesto es que las variables están medidas sin error y el uso de predictores con error afecta a los estimadores de la regresión, pudiendo tener un efecto sustancial en la estimación de los efectos de la escuela. En el caso de las puntuaciones de los tests, el efecto puede reducirse tomando puntuaciones de varios años (Ladd y Walsh, 2002; Lockwood y McCaffrey, 2007; Sanders, 2004). La presencia de errores de medida también afecta a otras variables utilizadas en el ajuste que en muchas ocasiones proceden de autoinformes de los estudiantes y están bastante afectadas por el error. Los errores de medida no uniformes también pueden llevar a la heterocedasticidad de los resultados, contribuyendo a estimaciones sesgadas (McCaffrey et al., 2003).

Por otra parte, en los modelos multivariantes, se requiere más investigación sobre los efectos de los diferentes tratamientos de las estructuras de covarianza (Raudenbush y Bryk, 2002; Singer y Willett, 2003).

En resumen, es necesario realizar más investigación sobre la robustez de los estimadores en presencia de incumplimiento de supuestos, que algunos autores califican como «heroicos» (Rubin et al., 2004).

Cuestiones relacionadas con las medidas de resultados en los modelos de VA

Los modelos de VA tienen como principal soporte las puntuaciones en tests de rendimiento educativo y sus inferencias por lo tanto requieren de tests con adecuadas propiedades psicométricas. La utilidad de su información depende de la calidad de los tests empleados. Desgraciadamente, aunque hay tests mejores que otros, podemos afirmar que «no existe una medida perfecta del rendimiento», proposición con la que están de acuerdo incluso los más ardientes defensores de los tests.

Uno de los requisitos de gran parte de los modelos es el de disponer de las denominadas *escalas verticales* y, tal como señalan McCaffrey et al. (2003), los cambios en los procedimientos de construcción, los métodos usados en la vinculación de los diferentes tests que las componen o el peso relativo dado a sus componentes, pueden influir en las inferencias sobre los efectos de las escuelas o los profesores. En Martínez Arias et al. (2008), este volumen se presentan algunos modelos en los que no se requieren tests verticalmente escalados, como algunos basados en estándares de rendimiento, que utilizan puntuaciones ordinales y que se podrían tratar con los modelos multinivel ordinales Fielding, Yang, Goldstein (2003). No obstante, la evaluación en estos casos, aunque más sencilla, también está basada en puntuaciones de los tests (Lissitz y Huyhn, 2003; Huyhn y Schneider, 2005) y deben gozar de adecuadas propiedades psicométricas. También es posible transformar los datos de los diferentes tests a otros tipos de escalas normativas, como las puntuaciones equivalentes en curva normal (Gong, Perie y Dunn, 2006; McCall, Kingsbury y Olson, 2004; Smith y Yen, 2006), pero las escalas verticales parecen las más adecuadas con los modelos de VA (Choi, Goldschmidt y Yamashiro, 2006).

En general, los modelos parten del supuesto de que las puntuaciones son *bastante buenas,* pero sin entrar en demasiados detalles sobre lo que esta afirmación significa y considerando el tema de las puntuaciones poco problemático.

Lissitz y Huyhn (2003) definen el escalamiento como el proceso mediante el que las puntuaciones directas obtenidas en un test se transforman en un nuevo conjunto de números, con algunos atributos particulares como media y desviación típica. Los mismos autores definen la *escala vertical* como una escala única unidimensional que resume el rendimiento de los estudiantes de diferentes cursos o niveles. Estas escalas permiten establecer comparaciones del rendimiento de los estudiantes a lo largo de diversos cursos para estimar su progreso (Patz, 2007).

La construcción de escalas verticales requiere utilizar procedimientos conocidos en la literatura estadística como de equiparación de formas distintas de tests, con el objetivo de que sus puntuaciones sean comparables. Una breve descripción de los principales métodos y diseños puede encontrarse en Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006) y en Kolen (2006). Un tratamiento completo en Kolen y Brennan (2004) y en Dorans, Pommerich y Holland (2007). Básicamente, hay dos situaciones bien diferenciadas en las que se necesita la conversión de un conjunto de tests a una escala común o equiparada. La equiparación horizontal tiene lugar cuando se usan formas múltiples, de dificultades similares, para evaluar un contenido semejante dentro del mismo curso o edad. La equiparación vertical es denominada

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

vinculación vertical (vertical linking) por no cumplir estrictamente con los requisitos de la equiparación (Holland, 2007) y se utiliza para vincular puntuaciones de tests que evalúan el mismo contenido general, pero con niveles de dificultad diferentes, normalmente destinados a distintos cursos. Las puntuaciones de escala resultantes se denominan escalas verticales. El escalamiento vertical es el proceso de vincular diferentes niveles de un instrumento que miden el mismo constructo, en una escala común de puntuaciones (Kolen, 2006; Kolen y Brennan, 2004; Lissitz y Huyhn, 2003; Patz, 2007; Petersen, Kolen y Hoover, 1989). La validez de las inferencias del VA sobre las ganancias depende en gran medida de la adecuación de las escalas, que generalmente disminuye cuando aumenta la distancia entre cursos o niveles.

Hay numerosas cuestiones prácticas no resueltas relativas a estas escalas y muy pocas recomendaciones claras sobre su construcción. De hecho, los *Standards for Educational and Psychological Tests* (AERA, APA & Nacional Council on Measurement in Education, 1999), las mencionan solamente de pasada. En su construcción deben asumirse diferentes supuestos y tomarse diversas decisiones que pueden llevar a diferentes resultados en los modelos de VA que las utilizan (Harris, 2007; Patz, 2007; Patz y Yao, 2007a; Patz y Yao, 2007b; Tong y Kolen, 2007; Yen, 2007). Su aplicación, cada vez más frecuente en los modelos de VA, ha llevado a una reflexión crítica sobre sus supuestos. Para un análisis más detallado de estos problemas se recomienda la lectura de las referencias citadas, especialmente Harris.

Alternativas a las escalas verticales convencionales

La alternativa más prometedora es utilizar en su construcción modelos psicométricos más complejos, que permitan considerar constructos multidimensionales que reflejen los cambios curriculares y de desarrollo. La opción son los modelos multidimensionales de la teoría de la respuesta al ítem (*Multidimensional Item Response Theory,* MIRT). En este sentido ya existen resultados de algunos de estos modelos (Martineau et al., 2007; Reckase y Li, 2007; Reckase y Martineau, 2004; Patz y Yao, 2007a, 2007b; Roberts y Ma, 2006). No obstante, en las aplicaciones prácticas puede haber algunas dificultades como la identificación de la dimensionalidad adecuada y los procedimientos requeridos para la vinculación. Otra dificultad añadida es la falta de software comercial.

Otra opción es considerar el rendimiento como medida ordinal, utilizando los niveles definidos por los estándares, como en algunos de los modelos presentados en

Martínez Arias et al. (2008). En este caso el problema es el de la comparabilidad entre los niveles de rendimiento de los distintos cursos y para garantizarla se han propuesto los *estándares verticalmente moderados* (Cizek, 2005; Huyhn y Schneider, 2005; Lissitz y Huyhn, 2003) o *verticalmente articulados* (Ferrara, Phillips, Williams, Leinwand, Mahoney y Ahadi, 2007). Básicamente, como señalan Huyhn y Schneider (2005) sus dos elementos básicos son: un conjunto de definiciones políticas de los niveles de rendimiento que se usan para todos los cursos y una línea de tendencia consistente entre cursos que se impone a los porcentajes de estudiantes en las diferentes categorías. Razones de espacio impiden entrar aquí en la descripción de los procedimientos, pero una buena lectura es número monográfico de la revista *Applied Measurement in Education* (2005, vol.v18, núm. 1) coordinado por Cizek.

Cuestiones prácticas ligadas a la aplicación y uso de los tests

La administración de los tests

En cuanto a la aplicación de los tests de rendimiento hay dos cuestiones importantes: el número de medidas de resultados y las fechas de recogida de datos. En relación con la primera cuestión, un sistema de evaluación de VA implica al menos dos mediciones de los resultados, normalmente de dos cursos consecutivos. Los modelos con dos evaluaciones consecutivas analizan la diferencia como ganancia o utilizan la primera como predictora de la segunda. Aunque estos modelos son bastante comunes, no permiten todo el potencial de los estudios longitudinales, ya que no permiten examinar la forma de la función de desarrollo. Como Rogosa (1995) advierte «dos medidas son mejor que una, pero no mucho mejor» (p. 744). Además en estos casos, pueden producirse algunos de los problemas ya tratados en el apartado de los problemas estadísticos, por lo que se recomienda utilizar medidas múltiples. Estas permiten un tratamiento multivariante y multinivel con el que se pueden estimar en el primer nivel los parámetros del cambio y desarrollo y obtener una visión más estable de las ganancias de la escuela Drury y Doran, 2003; Raudenbush, 2004a, 2004b; Raudenbush y Bryk, 2002).

La segunda cuestión se refiere al momento de la recogida de datos. En muchas de las aplicaciones de los modelos de VA se recogen datos una vez al año, normalmente

en primavera. Algunos autores cuestionan estas fechas de aplicación, ya que el intervalo entre aplicaciones incluye partes de dos años académicos con la interrupción del verano, cuyos efectos pueden estar relacionados con el estatus socioeconómico de los estudiantes. McCaffrey et al. (2003) consideran que esto no supone una amenaza a la validez de las inferencias. Otra alternativa considerada a veces es la realización de dos recogidas de datos en el curso, en otoño y primavera. Esta opción, además de más costosa, es desaconsejada por algunos autores (Linn, 2000) puesto que puede introducir sesgos derivados de la selección de los estudiantes, errores de conversión de las escalas y efectos de la práctica. En los modelos basados solamente en dos puntuaciones, es importante la elección de la primera ocasión de medida, ya que el punto inicial puede afectar a la interpretación de los resultados.

Los incentivos «perversos» derivados del uso de los tests

Las evaluaciones mediante tests que tienen como objetivo la rendición de cuentas o la publicidad de los resultados son objeto de críticas frecuentes por parte de algunos investigadores. Aunque el repertorio de críticas es extenso, la mayor parte se centran en dos grandes grupos: la reducción del currículo y la inflación de las puntuaciones de los tests.

El primer tipo de críticas suele referirse a las materias evaluadas y a los tests utilizados en la evaluación. Por lo que se refiere a las materias evaluadas, la crítica se basa en que muchos modelos de VA utilizados en la rendición de cuentas solamente evalúan un número reducido de materias, siendo las más frecuentes las Matemáticas y la Lengua (las requeridas por la NCLB en Estados Unidos). Esta restricción de contenidos puede enfatizar las materias objeto de evaluación y descuidando las restantes. En realidad, esta crítica es aplicable a cualquier sistema de evaluación mediante tests y no solamente a los modelos de VA, ya que, teóricamente pueden construirse tests de cualquier materia, si bien es cierto que las diferencias curriculares entre cursos en Ciencias Naturales y Ciencias Sociales hacen más difícil la construcción de escalas verticales.

Otra crítica frecuente en relación con la anterior es la que insiste en que la evaluación para la rendición de cuentas lleva a los docentes a utilizar estrategias conocidas como «enseñanza para el test» (Kohn, 2000; O'Day, 2002). La crítica adquiere mayor virulencia cuando los tests utilizan formatos de elección múltiple de los que se dice que no captan las destrezas de solución de problemas y procesos superiores. Haciéndose eco de esta crítica, cada vez más evaluaciones incluyen cuestiones de respuesta construida y de ensayo, aunque algunos problemas relacionados con su calidad

técnica y sobre todo con los costos, impiden utilizarlos en las evaluaciones sistemáticas con periodicidad frecuente (Wang et al., 2006). Jacob (2002, 2005) no encontró evidencias de reducción del currículo en las escuelas de Chicago.

Debido a las prácticas docentes vinculadas a los contenidos de los tests, se puede producir la *inflación de las puntuaciones* (Koretz, 2005; Koretz y Hamilton, 2006; Linn, 2006), que lleva a incrementos en las puntuaciones de los tests que no representan auténticas ganancias de aprendizaje y que no se reflejan en otras evaluaciones.

En el fondo de todas estas discusiones se encuentra la denominada Ley de Campbell (Campbell, 1975, p. 35) «cuanto más se usa un indicador social cuantitativo para la toma de decisiones, más sometido estará a presiones de corrupción y habrá mayor tendencia a distorsionar y viciar el proceso social que intenta monitorizar». La generalización de la ley de Campbell está tan extendida que Madaus y Clark (2001) la comparan al principio de incertidumbre de Heisenberg en las Ciencias Sociales. Nichols y Berliner (2005) hacen una excelente revisión de evidencias de la Ley de Campbell en diversas áreas de las ciencias sociales, entre otras en la evaluación educativa, mostrando en este ámbito ejemplos de efectos como el estrechamiento del currículo, la enseñanza del test, la subjetividad en el establecimiento de estándares y la exclusión de las sesiones de tests de algunos grupos de estudiantes.

Otra crítica frecuente, se refiere al carácter de evaluación normativa de los modelos de VA (los resultados o residuos se basan en la comparación con otras escuelas de la muestra). En opinión de algunos investigadores, este tipo de evaluación puede rebajar los niveles de rendimiento de los estudiantes con relación a los que se obtendrían por medio de una evaluación basada en estándares. Se demanda poder responder a la pregunta *¿cuánta ganancia es suficiente?.* En realidad, la respuesta a esta pregunta no es incompatible con los modelos de VA que siguen durante varios años el aprendizaje de los estudiantes y permiten proyectar cuáles serán sus puntuaciones. Algunos de los modelos citados en Martínez Arias et al. (2008) intentan compatibilizar ambas aproximaciones.

Conclusiones y recomendaciones para la investigación futura

Mediante una compleja metodología estadística, los modelos de VA parecen proporcionar un procedimiento adecuado para la evaluación de la eficacia de las escuelas, ya que separa hasta cierto punto la contribución de las escuelas y profesores de otros

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

factores extraescolares. Con la utilización de medidas longitudinales del desarrollo de los estudiantes proporcionan fundamentos más defendibles que los tradicionales métodos transversales basados en el estatus de la escuela en un momento en el tiempo. Estas propiedades de los VA los han llevado a ser considerados como una promesa para realizar evaluaciones de escuelas de forma más justa y segura. Por ello han despertado un fuerte interés tanto entre los investigadores, como entre los políticos y administradores escolares.

Existen ya varios modelos implantados desde hace años cuya experiencia permite reflexionar sobre sus potencialidades de uso, fuerzas y debilidades. Aunque tienen algunas limitaciones técnicas y prácticas, en general han tenido una buena aceptación entre las diferentes partes implicadas en la educación.

Los principales usos de los modelos pueden resumirse en los relacionados con la rendición de cuentas y con el diagnóstico y mejora de las escuelas.

En cuanto al primero de los usos, la aplicación de los resultados del VA no parece adecuada por el momento cuando tiene consecuencias para las escuelas o los profesores. Hay numerosas cuestiones no resueltas sobre la atribución clara de los efectos mostrados en los residuos de los modelos a las escuelas. La complejidad de las escuelas nunca puede ser totalmente captada con los modelos de VA como para atribuirles efectos causales sobre el aprendizaje de sus estudiantes. Los modelos de ganancias simples son aún más problemáticos para esta atribución que los más complejos.

Si la rendición de cuentas consiste simplemente en la publicación de sus resultados para información al público y ayuda a los padres en la elección de escuela, los datos basados en VA son más justos que los resultados brutos no ajustados. No obstante, hay que tener cautela en la difusión de la información, especialmente cuando se establecen ordenaciones, puesto que hay pocas diferencias estadísticamente significativas entre la mayor parte de las escuelas, debido a la incertidumbre o error muestral.

Los modelos de VA pueden resultar muy útiles en los procesos de diagnóstico y mejora de las escuelas. Permiten identificar escuelas que necesitan asistencia, ayuda y formación del profesorado y son un componente esencial en los procesos de autoevaluación de los centros para el diagnóstico de sus fuerzas y debilidades. Combinados con otras metodologías de naturaleza más cualitativa como observaciones, entrevistas, portafolios, etc., ayudan a la identificación de buenas prácticas que se pueden potenciar en las reformas educativas.

Desde el punto de vista técnico, los modelos de VA son un área de investigación activa de la que son esperables nuevos desarrollos metodológicos que permitan mejorar tanto los modelos como los instrumentos de medida utilizados en la evaluación

de resultados. Hay algunos aspectos concretos que necesitan más investigación: estudios empíricos de comparación de diferentes modelos, influencia del uso de diferentes características contextuales, efectos de la no aleatoriedad de los datos perdidos, efectos de la violación de los supuestos estadísticos y sensibilidad de los modelos a los diferentes métodos de construcción de tests y a los efectos de sus propiedades psicométricas.

Se requieren además investigaciones empíricas sobre otros aspectos apenas investigados hasta el momento y sobre los que los modelos de VA son frecuente objeto de críticas: validación externa de los efectos, contrastándolos con otras medidas de eficacia de escuelas y profesores y sobre los supuestos «incentivos perversos» a los que según algunos lleva su aplicación.

Otros aspectos muy importantes que hay que mejorar para el uso eficaz de los datos son los que tienen que ver con la comunicación e interpretación de los resultados por parte de audiencias que no son expertas en estadística. La interpretación de los residuos derivados de complejos modelos estadísticos como efectos de las escuelas puede ser difícil y poco transparente. Es imprescindible llegar a informes que sean de fácil comprensión por las escuelas y profesores, insistiendo en la transparencia de los resultados y en la utilidad para los procesos de decisión de la escuela. Para ello, la implantación de los modelos de VA deberá ir acompañada de planes de formación para el profesorado y otros implicados.

En resumen, a pesar de algunas de las limitaciones presentadas en este trabajo, muchas de las cuales están siendo objeto de investigación para su mejora, los modelos de VA, utilizados de forma adecuada y prudente, sin la pretensión de hacer inferencias para las que por el momento no están capacitados, representan un valioso instrumento al servicio de las escuelas.

Referencias bibliográficas

Aera, Apa & Nacional Council on Measurement in Education (1999). *Standards for Educational and Psychological Tests.* Washington, DC: American Psychological Association.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational researcher,* 37, 65-75.

Ballou, D. (2002). Sizing up test scores. *Education Next,* 2, 10-15.

— (2005). Value-added assessment: Lessons from Tennessee. En R. Lissitz (ed.), *Value-Added models in education: Theory and Applications* (pp. 272-297). Mapple Grove, MN: JAM Press

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.

Bock, R. D., Wolfe, R. & Fisher, T. H. (1996). *A review and analysis of the Tennessee value-added assessment system. Summary and recommendations.* Nashville, TN: Tennessee Office of Education Accountability.

Braun, H. (2005a). Value-Added Modeling: What Does Due Diligence Require? En R. Lissitz (ed.), *Value-Added models in education: Theory and Applications* (19-39). Mapple Grove, MN: JAM Press.

— (2005b). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Princeton, NJ: Educational Testing Service.

Braun, H. & Wainer. H. (2007). Value-added modelling. En S. Sinharay & C. Rao (eds.), *Handbook of statistics, 26: Psychometrics* (pp. 867-891). Amsterdam: North Holland.

Bryk, S.A., Thum, M.Y., Easton, Q.J. & Luppescu, S. (1998). *Academic productivity of Chicago public elementary schools.* Chicago, IL: Consortium on Chicago School Research.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics,* 13, 227-241.

Campbell, D. T. (1975). Assessing the impact of planned social change. En G. Lyons (ed.), *Social research and public policies: The darmouth/OECD Conference* (pp. 3-45). Hanover. NH: Darthmouth College. The Public Affairs Center.

Castro, M., Ruíz de Miguel, C. y López, E. (2008). Forma básica del crecimiento en los modelos de valor añadido: vías para la supresión del efecto de regresión y funciones de crecimiento no lineales. *Revista de Educación*, 348.

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.

Cizek, J. (ed.). (2005). Adapting testing technology to serve accountability aims: The Case of vertically-moderated standard setting. A Special Issue of *Applied Measurement in Education,* 18, 1-9.

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

Cohen, J., Johnson, E. & Angeles, J. (2000). *Variance estimation when sampling in two dimensions via the jackknife with application to the national assessment of educational progress.* Washington, DC: American Institute for Research.

Doran, H. C. (2004). *Value-Added Models and Adequate Yearly Progress: Combining Growth and Adequacy in a Standards-based Environment.* Paper presented at the Annual CCSSO Large-Scale Assessment Conference, Boston, Massachusetts.

Doran, H. C. & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. En R. Lissitz (ed.), *Value-added models in education: Theory and applications* (pp. 80-110). Maple Grove, MN: JAM Press.

Doran, H. C. & Fleischman, S. (2005). Research matters/Challenges of value-added assessment. *Eductional Leadership,* 63, 85-87.

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

Doran, H. C. & Jiang, T. (2006). The impact of linking error in longitudinal análisis: an empirical demonstration. En R. Lissitz (ed.), *Longitudinal and value added models of student performance* (pp. 210-229). Maple Grove, MN: JAM Press.

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

Dorans, N. J., Pommerich, M. & Holland, P. W. (eds.). (2007). *Linking and aligning scores and scales.* New York: Springer.

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

Ferrara, S., Phillips, G. W., Williams, P. L., Leinwand, S., Mahoney, S. & Ahadi, S. (2007). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. En R. Lissitz (ed.), *Assessing and modeling cognitive development in school* (pp. 31-63). Maple Grove, MN: JAM Press.

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.

Fuchs, T. & Wossmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics,* 32, 433-464.

Gaviria, J. L., Biencinto, M. C. y Navarro, E. (2008). Invarianza de la estructura de covarianzas de las medidas de rendimiento académicos en estudios longitudinales en la transición de la educación primaria a secundaria. *Revista de Educación,* 348.

Glenn, C. & De Groof, J. (2005). *Balancing Freedom, Autonomy and Accoutnability in Education.* Nijmegan NL: Wolf Legal Publishers.

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

Goldschmidt, P. & Choi, K. (2007). *The practical benefits of growth models for accountability and the limitations under NCLB. Policy Brief 9.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Goldschmidt, P., Choi, K. & Martinez, F. (2004). *Using Hierarchical Growth Models To Monitor School Performance Over Time: Comparing NCE to Scale Score Results* (CSE Report 618). Los Angeles, CA: Center for Research on Evaluation Standards and Student Testing.

Goldstein, H. & Healy M. J. R. (1995). The Graphical Presentation of a Collection of Means. *Journal of the Royal Statistical Society,* 581, 175-177.

Goldstein, H. (2003). *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.

Gulliksen, H. (1950). *Test theory.* New York. Wiley.

Haertel, E. H. (2004). *The behavior of linking items in test equating* (CSE Report No 630). Los Angeles, CA: Center for Research on Evaluation Standards and Student Testing.

Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item non-equivalent groups equating design. *Applied Psychological Measurement,* 21, 3-24.

Harris, D. J. (2007). Practical issues in vertical scaling. En N. J. Dorans, M. Pomerich, & P. W. Holland (eds.), *Linking and aligning scores and scales* (pp. 233-251). New York: Springer.

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 (pp. 255-290). Maple Grove, MN: JAM Press.

Holland, P. W. (2007). *A framework and history for score linking.* En N. J. Dorans, M. Pomerich & P. W. Holland (eds.), Linking and aligning scores and scales (pp. 5-30). New York: Springer.

Hoxby, C. (2000). Does competition among public schools benefit students and taxpayers. *American Economic Review,* 90, 1.209-1.238.

— (2003). School choice and school competition: Evidence from the United States. *Swedish Economic Policy Review,* 10, 9-65.

Huyhn, H. & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practice. *Applied Measurement in Education,* 18, 99-114.

Jacob, B. (2002). *Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools.* Cambridge, MA.: NBER Working Paper No. 8968.

Kane, T. J. & Staiger, D. O. (2002). *Volatility in school test scores: Implications for test-based accountability systems.* En D. Ravith (ed.), Brooking papers on education policy (pp. 235-260). Washington, DC: Brookings Institution.

Kim, S. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement,* 22, 131-143.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools.* Portsmouth, NH: Heineman.

Kolen, M. J. (2006). Scaling and norming. En R. L. Brennan (ed.). *Educational Measurement* (4ª ed.). (pp. 155-186). Westport, CT: American Council on Education.

Kolen, M. & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2ª ed.). New York: Springer-Verlag.

Koretz, D. M. (2005). Alignment, high stakes and the inflation of test scores. En J. L. Herman & E. H. Haertel (eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education.* vol. 104 Part 2 (pp. 99-118). Malden, MA: Blackwell.

Koretz, D. M. & Hamilton, L. S. (2006). Testing for accountability in K-12. En R. L. Brennan (ed.), *Educational Measurement* (4ª ed.), (pp. 531-578). Westport, CT: Praeger.

Kupermintz, H. (2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis,* 25(3).

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. (2000). Assessments and Accountability. *Educational Researcher,* 29, 4-14.

— (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education,* 19, 5-15.

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

Lissitz, R. (ed.). (2005). *Value added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

— (2006). *Longitudinal and Value added models of student performance.* Maple Grove, MN: JAM Press.

Lissitz, R. W. & Huyhn, H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research, and Evaluation,* 8, 10.

Little R. J.A. & Rubin D. B. (2002). *Statistical Analysis with Missing Data* (2ª ed.) New York: Wiley.

Lizasoain, L. y Juaristi, L. (2008). Análisis de la dimensionalidad en los modelos de valor añadido: estudio de pruebas de matemáticas empleando técnicas factoriales y métodos no paramétricos basados en TRI. *Revista de Educación,* 348.

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.

Madaus, G. & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. En G. Orfield & M. L. Kornhaber (eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 85-106). New York: Century Foundation Press.

Marchesi, A., Martín, E., Martínez Arias, R., Tiana, A. & Moreno, A. (2002). An evaluation network for educational change. *Studies in Educational Evaluation,* 29, 43-56.

Marchesi, A., Martínez Arias, R. y Martín, E. (2004). Estudio longitudinal sobre la influencia del nivel sociocultural en el aprendizaje de los alumnos de la ESO. *Infancia y Aprendizaje*, 27, 307-323.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models.* Unpublished Doctoral Dissertation. Michigan State University.

— (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics,* 31, 35-62.

Martineau, J. A., Subedi, D. R., Ward, K. H., Li, T., Lu, Y., Diao, Q., Pang, F. H., Drake, S., Song, T., Kao, S.C., Zheng, Y. & Li, X. (2007). Non-linear trajectories through multidimensional content spaces: An examination of psychometric claims of unidimensionality, linearity, and interval measurement. En R. Lissitz (ed.), *Assessing and modeling cognitive development in school* (pp. 96-142). Maple Grove, MN: JAM Press.

Martínez Arias, R., Gaviria, J. L. y Castro, M. (2008). Concepto y evolución de los modelos de valor añadido en educación. *Revista de Educación,* 348.

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: *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. & Hauser, C. (2007). Item response theory and longitudinal modeling: The real world is less complicated than we fear. En R. Lissitz (ed.), *Assessing and modeling cognitive development in school* (pp. 143-174). Maple Grove, MN: JAM Press.

Michaelides, M. P., & Haertel, E. H. (2004, May). *Sampling of common items: An unrecognized source of error in test equating* (Technical Report). Los Angeles: Center for the Study of Evaluation y National Center for Research on Evaluation, Standards, and Student Testing.

O'Day, J. (2002). Complexity, Accountability, and School Improvement. *Harvard Educational Review,* 72, 293-329.

Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems.* Washington, DC: The Council of Chief State School Officers.

Patz, R. J. & Yao, L. (2007a). Vertical scaling: Statistical models for measuring growth and achievement. En S. Sinharay & C. Rao (eds.), *Handbook of statistics, 26: Psychometrics* (pp. 955-975). Amsterdam: North Holland.

— (2007b). Methods and models for vertical scaling. En N. J. Dorans, M. Pommerich & P.W. Holland (eds.), *Linking and aligning scores and scales* (pp. 253-272). New York: Springer.

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming and equating. En R. L. Linn (ed.), *Educational Measurement.* (3ª ed.). (221-262). New York: Macmillan.

Pomplum, M., Omar, M. H. & Custer, M. (2004). A comparison of Winstep and Bilog-MG for vertical scaling with the Rasch model. *Applied Psychological measurement,* 28, 247-273.

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. (2004a). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing service.

— (2004b). 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. & 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.

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics,* 29, 117-120.

Reckase, M. D. & Li, T. (2007). Estimating gain in achievement when content specifications change: A multidimensional item response theory approach. En R. Lissitz (ed.), *Assessing and modeling cognitive development in school* (pp. 189-204). Maple Grove, MN: JAM Press.

Reckase, M. D. & Martineau, J. (2004). *The vertical scaling of science achievement tests.* Paper commissioned by the Committee on Test Design for K-12 Science Achievement. Washington, DC: National Research Council.

Roberts, J. S. & Ma, Q. (2006). *IRT models for the assessment of change across repeated measurements.* En R. Lissitz (ed.), Longitudinal and value added models of student performance (100-129). Maple Grove, MN: JAM Press.

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

Rosenbaum, P. R. (2002). *Observational studies.* New York: Springer.

Rose, L. C. & Gallup, A. M. (2007). The 39th Annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi, Delta, Kappan,* 89, 33-48.

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.

Sammons, P., Thomas, S. & Mortimore, P. (1997). *Forging links: Effective schools and effective departments.* London: Chapman y Hall.

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.

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. (2000). Understanding schools' use of value-aded data. The psychology and sociology of numbers. *Research Papers in Education: Policy and Practice*, 15, 241-258.

Schmidt, W. H., Houang, R. T. & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? En R. Lissitz (ed.), *Value-added models in education: Theory and applications* (pp. 145-165). Maple Grove, MN: JAM Press.

Schafer, W. D., Yen, S. J. & Rahman, T. (2000). School effects indices: Stability of one-and-two level formulations. *Journal of Experimental Education,* 68, 239-250.

Schatz, C. J., VonSecker, C. E. & Alban, T. R. (2005). Balancing accountabilit and improvement: Introducing value-added models to a large school system. En R. Lissitz (ed.), *Value added models in education: Theory and applications* (pp. 1-18). Maple Grove, MN: JAM Press.

Smith, R. L. & Yen, W. M. (2006). Models for evaluating grade-to-grade growth. En R. W. Lissitz (ed.), *Longitudinal and value added modeling of student performance* (pp. 82-99). Maple Grove, MN: JAM Press.

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.

Stuart, E. A. (2007). Estimating causal effects using school-level datasets. *Educational Researcher,* 36 (4), 187-198.

Supovitz, J. A. & Klein, V. (2003). *Mapping a course for improved student learning.* Philadelphia: Consortium for Policy Research in Education.

Tekwe, C. D., Carter, R. L., ma C.- X., Algina, J., Lucas, M., Roth, J. et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics,* 29, 11-36.

Thum, Y. M. (2006). Designing gross productivity indicators: A proposal for connecting accountability goals, data and analysis. En R. Lissitz (ed.), *Longitudinal and value added models of student performance* (pp. 436-479). Maple Grove, MN: JAM Press.

Tong, Y. & Kolen, M. J. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education,* 20, 227-253.

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. (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.

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?* (81-99). Thousand Oaks, CA: Corwin Press.

Wright, S. P. (2004). Advantages of a multivariate longitudinal approach to educational value-added assessment without imputation. Paper presented at the 2004 National Evaluation Institute, July 8-10, Colorado Springs, Colorado.

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* (pp. 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.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement,* 23, 299-325.

— (2007). Vertical scaling and Noo Child Left Behind. En N. J. Dorans, M. Pommerich & P. W. Holland (eds.), *Linking and aligning scores and scales* (pp. 273-283). New York: Springer.

Yen, W. M. & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement,* 34, 293-313.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. En R. L. Brennan (ed.), *Educational Measurement* (4ª ed.). (pp. 111-154). Westport, CT: Praeger.

Zvoch, K. & Stevens, J. J. (2005). Sample exclusion and student attrition effects in the longitudinal study of middle school mathematics performance. *Educational Assessment,* 10, 105-123.

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. Consultado de: http://www.ccsso.org/publications/

Gong, B., Perie, M. & Dunn, J. (2006). *Using Student Longitudinal Growth Measures for School Accountability Under No Child Left Behind: And Update to Inform Design Decisions.* Consultado el 9 de enero de 2008, de http://www.nciea.org/publications/GrowthModelUpdate_BGMAPJD07.pdf

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. Consultado de: http://www.nwea.org

McCall, M. S., Kingsbury, G. G. & Olson, A. (2004). *Individual growth and school success.* Lake Oswego, OR: Northwest Evaluation Association. Consultado de: http://www.nwea.org

Nichols & Berliner (2005). *The inevitable corruption of indicators and educators through high-stakes testing.* Tempe, Arizona: Arizona State University. Education Policy studies Laboratory. Consultado de: http://edpolicylab.org

Sanders, W. L. (2006). Comparisons Among Various Educational Assessment Value-Added Models. The Power of Two-National Value-Added Conference, Columbus, OH. Consultado el 20 de mayo de 2007. Consultado de: http://www.sas.com/govedu/edu/services/vaconferencepaper.pdf

Schafer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, research y Evaluation,* 11(4). Consultado de: http://pareonline.net/getvn.asp?v=11yn=4

Yeagley, R. (2007). Separating Growth from Value Added: Two Academic Models Offer Different Tools for Different Purposes. *The School Administrator,* 64(1). Consultado el 18 de octubre de 2007. Consultado de: http://www.aasa.org/publications/saarticledetail.cfm?ItemNumber=7941

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