Patrones de correlación entre medidas de rendimiento escolar en evaluaciones longitudinales: un estudio de simulación desde un enfoque multinivel

Achievement measurements correlation patterns in longitudinal assessments: a multilevel approach simulation study

Ángeles Blanco Blanco, Coral González Barberá y Xavier G. Ordóñez

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

Resumen

Los modelos de valor añadido operan sobre una estructura de datos longitudinal que plantea importantes retos metodológicos, tanto estadísticos como psicométricos. Este artículo estudia un problema particular asociado a esta estructura desde el enfoque de los modelos jerárquico-lineales con medidas repetidas. Concretamente se estudian los patrones diferenciales de correlación que presentan las puntuaciones latentes, observadas y verdaderas a medida que transcurre el tiempo, y el efecto que ejercen sobre ellos dos factores: la complejidad creciente del rasgo y los patrones de covariación de los residuos en el nivel del alumno y del centro. Para ello, se realiza una simulación que trabaja con una muestra de 25.000 sujetos, para los que se generan puntuaciones latentes, observadas y verdaderas en cuatro evaluaciones sucesivas y seis condiciones distintas. Los resultados permiten describir un patrón de correlaciones en las puntuaciones latentes cuando se asume la unidimensionalidad del rasgo, que difiere del registrado habitualmente para las puntuaciones observadas. Así, la magnitud de la correlación entre dos medidas es mayor cuanto más se alejan del punto inicial de la serie de medidas. Asimismo, los resultados permiten interpretar este patrón mediante el análisis comparativo del efecto de

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

los dos factores considerados sobre los distintos tipos de puntuaciones. Entre las conclusiones del estudio, destacamos dos con especial relevancia en la evaluación longitudinal del rendimiento académico: la necesidad de conocer los patrones de correlación en el nivel de las variables latentes para dar explicaciones plausibles a los patrones empíricos encontrados y la relevancia de la complejidad creciente del rasgo evaluado en la determinación de las estructuras de covariación.

Palabras clave: medida del rendimiento académico, análisis de datos longitudinales, evaluación del valor añadido, modelos multinivel para el estudio del cambio, teoría de respuesta al ítem, multidimensionalidad del rasgo latente, simulación.

Abstract

Value-added models are applied in longitudinal data structures that raise important statistical and psychometrical methodological challenges. In this context, the aim of the study was to explain the differential correlation patterns of *observed, true* and *latent* scores over time. In particular, the effect of the *trait complexity factor* and the second (student) and third (school) levels of the *residual covariate patterns* was analyzed, from the point of view of the lineal hierarchical models with repeated measurements. *Observed, true* and *latent* scores for 25000 subjects were generated using a simulation, in four consecutive assessments and six different conditions. The first result is a description of the correlation pattern in the latent scores when the unidimensionality of the trait is assumed. This pattern is different to the one usually observed in empirical studies. Thus, the magnitude of the correlation between two measures increases as the measures distance to the starting point increases. This pattern can be explained as the result of the effect of the two referred factors. From the results two main findings are proposed. First, the need of understanding of the correlation patterns in the latent variables level to explain the relations in the empirical level. Second, the correlation patterns of achievement measures in educational longitudinal assessment are mainly determined by the growing complexity of the trait.

Key Words: achievement measure, testing, longitudinal data analysis, value-added assessment, multilevel models for change, item response theory, latent variable multidimensionality, simulation study.

Introducción

Bajo la denominación genérica de *modelos de valor añadido* se incluyen en la actualidad diferentes modelos estadísticos aplicados a la evaluación de sistemas educativos que varían muy sustancialmente entre ellos, tanto en complejidad como en los supuestos que subyacen a los mismos (Sanders, 2006; Wiley, 2005). Sin embargo, todas las aproximaciones y propuestas englobadas bajo este término comparten un objetivo: vincular los cambios registrados en el rendimiento individual de los alumnos con las escuelas a las que asisten o con los profesores responsables de la clase a la que pertenecen (Martínez Arias, Gaviria y Castro, 2009). Puesto que son modelos orientados al análisis del cambio, un segundo denominador común a todos ellos, derivado del anterior, es que han de operar sobre datos que permitan el seguimiento individual del crecimiento en el rendimiento a lo largo del tiempo, con el fin de estimar la contribución de la escuela y/o el profesor a dicho crecimiento (Braun, 2005). En consecuencia, la estructura longitudinal de las medidas de rendimiento del alumnado y las implicaciones que de ésta se derivan conforman un objeto de análisis de importancia central en las aproximaciones metodológicas a la estimación de medidas del valor añadido. Porque ciertamente la modelización de datos longitudinales plantea un número importante de cuestiones estadísticas y psicométricas (para una revisión amplia, véase por ejemplo McCaffrey, Lakewood, Koretz y Hamilton, 2003; y, en este mismo número, Martínez Arias, 2009).

En el marco general de estas consideraciones, el presente trabajo se ocupa de analizar un problema particular ligado a la naturaleza longitudinal de las medidas del rendimiento académico típicamente usadas en la evaluación de sistemas educativos. En concreto, el objetivo central del trabajo es ilustrar y explicar la atenuación que se registra en las correlaciones halladas a lo largo del tiempo entre las puntuaciones de rasgo (θ), entre las puntuaciones observadas (X) y entre las puntuaciones verdaderas (T) en las evaluaciones longitudinales. Como se verá, este fenómeno puede entenderse presumiblemente asociado al incremento de la complejidad del constructo *rendimiento académico* en un área determinada a lo largo del tiempo, y por tanto asociado a la desviación de la unidimensionalidad supuesta del rasgo cuando se inicia la serie de medidas.

De acuerdo con lo que constituyen las propuestas recientes más aceptadas, el estudio adopta la aproximación proporcionada por los modelos jerárquico-lineales para la modelización longitudinal (Singer y Willet, 2003), alineándose igualmente con los desarrollos metodológicos más actuales en la medida del valor añadido. Como una derivación de la decisión anterior, también se incorpora como factor explicativo del

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

fenómeno bajo estudio el patrón diverso de covariación (positivo o nulo) asociado a los residuos en el nivel del alumno y del centro educativo.

Planteamiento del problema

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

Sin embargo, la asunción de este patrón frecuente de covariación presumido para las puntuaciones observadas, ¿puede mantenerse cuando tomamos como referente las puntuaciones de rasgo? Es decir, ¿es el que cabe esperar cuando el análisis está referido a las relaciones *reales* que acontecen entre las variables latentes? La respuesta negativa a esta cuestión y sus implicaciones constituyen el centro de este trabajo.

Es importante considerar que hacemos aquí uso del término *real* en el sentido propuesto por Bhaskar (1978) cuando distingue entre tres dominios relevantes en la formalización científica: a) el *empírico,* que consiste en las impresiones de los sentidos y nuestras experiencias directas de las cosas; b) el dominio de *lo actual* que consiste en los eventos que se producen y dan lugar a la realidad efectiva, con toda su complejidad, tanto si es directamente experimentada por nosotros como si no; c) el dominio de *lo real,* que consiste en entidades y estructuras que producen eventos.

Así, desde una perspectiva postpositivista y en un contexto de realismo científico, no son los eventos en sí mismos los que constituyen la realidad, sino que lo actual es una de las posibles manifestaciones de lo real. Por lo tanto, como señala House (1991), desde una perspectiva realista las leyes que se formulan en la ciencia no son proposiciones acerca de los eventos observados o las experiencias habidas, sino más

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

bien proposiciones acerca de las formas en que las entidades causales actúan, produciendo de esa forma eventos y experiencias.

En definitiva, es el dominio de *lo real* el que debe ser conocido y explicado. Por tanto, es de la máxima relevancia el estudio y determinación de las relaciones causales que se establecen entre las variables latentes, puesto que lo que se hace empírico y lo que se hace observable (los patrones correlacionales registrados en las variables observadas) depende de los patrones de correlación que operan a ese nivel.

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

- Primer nivel: Ocasión/Tiempo
- Segundo nivel: Alumno
- Tercer nivel: Escuela

$$y_{tjk} = \beta_{ojk} + \beta_{1jk}(t - t_o) + \varepsilon_{tjk}$$
$$\beta_{ojk} = \beta_{ok} + \mu_{ojk} \quad \beta_{1jk} = \beta_{1k} + \mu_{1jk}$$
$$\beta_{ok} = \beta_{oo} + \mu_{ok} \quad \beta_{1k} = \beta_{10} + \mu_{1k} \qquad (0)$$

Un hecho derivado directamente del modelo mismo es que la varianza total de las puntuaciones en rendimiento aumenta a medida que lo hace el tiempo. Sin embargo, es muy importante notar que, a la vez, el modelo asume que el peso relativo del término residual ε_{tjk} permanece constante a lo largo del tiempo. En consecuencia, es la varianza sistemática la que aumenta manteniéndose constante la varianza no explicada, lo que quiere decir que la varianza compartida entre las variables de respuesta correspondientes a las distintas aplicaciones (cuantificada en términos de covarianza/correlación) aumenta progresivamente. Entonces, es claro que en estas condiciones cabe esperar que las correlaciones entre aplicaciones aumenten a medida que transcurre el tiempo y se van sumando evaluaciones o medidas consecutivas. Lo anterior, que puede ser interpretado en términos de aumento creciente del poder explicativo del modelo, alejará el patrón correlacional registrado para las puntuaciones de rasgo del patrón generalmente asumido para las correlaciones observadas entre evaluaciones sucesivas de rendimiento.

Una ejemplificación de patrones de correlación entre puntuaciones de rasgo

Con el fin de ilustrar la proposición anterior, se desarrolla un ejemplo en el que se atribuyen valores a los términos del modelo multinivel básico [0]. De este modo, se obtendrán las correlaciones para cuatro aplicaciones supuestas de un rasgo dado.

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

La correlación entre dos medidas del rasgo y tomadas en los momentos t y s puede definirse como: $r_{y_t y_s} = \sigma_{y_t y_s} / \sigma_{y_t} \sigma_{y_s}$

Y de acuerdo con el modelo [0], la medida del rasgo y en los momentos t y s toma la forma siguiente:

$$y_t = \beta_{ojk} + \beta_{1jk}(t - t_o) + \varepsilon_{tjk} \; ;$$
$$y_s = \beta_{ojk} + \beta_{1jk}(s - t_o) + \varepsilon_{tjk}$$

siendo por tanto la varianza de la medida, ilustrada aquí para el momento t:

$$\sigma^2_{yt} = \sigma^2_{\mu_{ok}} + \sigma^2 \mu_{ojk} + (t - t_0)^2 \left(\sigma^2_{\mu_{1k}}\right) + (t - t_0)^2 \left(\sigma^2_{\mu_{1jk}}\right) + 2(t - t_0)\left(\sigma_{\mu_{0k}\mu_{ik}}\right) +$$
$$+ 2(t - t_0)\left(\sigma_{\mu_{0jk}\mu_{ijk}}\right) + \sigma^2 \varepsilon_{tjk} \qquad (1)$$

La covarianza entre medidas del rasgo en los tiempos t y s, por su parte, es igual a:

$$\sigma_{y_t y_s} = E(y_t y_s) - E(y_t)E(y_s)$$

Donde:

$$E(y_t) = E(\beta_{ojk}) + (t - t_o)E(\beta_{1jk}) = E(\beta_{ok}) + (t - t_o)E(\beta_{1k}) = \beta_{00} + (t - t_0)\beta_{10}$$

; siendo por tanto $E(y_s) = \beta_{00} + (s - t_0)\beta_{10}$

Así tenemos que:

$$E(y_t)E(y_s) = \beta^2_{00} + (s - t_0)\beta_{10}\beta_{00} + (t - t_0)\beta_{10}\beta_{00} + \beta^2_{10}(t - t_0)(s - t_o) =$$
$$= \beta^2_{00} + \beta_{10}\beta_{00}[(t - t_0) + (s - t_0)] + \beta^2_{10}[(t - t_0)(s - t_0)]$$

Por otro lado,

$$E(y_t y_s) = E\left[\beta^2_{ojk} + \beta_{ojk}\beta_{1jK}[(t - t_0) + (s - t_0)] + \beta_{1jk}[(t - t_0)(s - t_0)]\right] =$$
$$= E\left(\beta^2_{ojk}\right) + [(t - t_0) + (s - t_0)]E\left(\beta_{ojk}\beta_{1jK}\right) + E\left(\beta^2_{1jk}\right)$$

Donde:

$$E\left(\beta^2_{ojk}\right) = E\left[(\beta_{00} + \mu_{0k} + \mu_{ojk})^2\right] = E\left[\beta^2_{ok} + \mu^2_{ojk} + 2\beta_{ok}\mu_{ojk}\right] =$$
$$= E\left[\beta^2_{00}\right] + E\left[\mu^2_{ok}\right] + E\left[\mu^2_{ojk}\right] \; ;$$
$$E\left(\beta^2_{1jk}\right) = E\left[\beta^2_{10}\right] + E\left[\mu^2_{1k}\right] + E\left[\mu^2_{1jk}\right] \; ; \text{ y}$$
$$E\left(\beta_{ojk}\beta_{1jK}\right) = E\left[\beta_{ok}\beta_{1k} + \beta_{ok}\beta_{1jk} + \beta_{1k}\beta_{0jk} + \mu_{ojk}\mu_{1jk}\right]$$

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

Puesto que $E(\beta_{ok}\beta_{1jk}) = E(\beta_{1k}\beta_{ojk}) = 0$, tenemos que:

$$E(\beta_{ojk}\beta_{1jK}) = E[\beta_{ok}\beta_{1k}] + E[\mu_{ojk}\mu_{1jk}] = \beta_{00}\beta_{10} + E[\mu_{0k}\mu_{1k}] + E[\mu_{0jk}\mu_{1jk}]$$

Y sustituyendo los tres términos anteriores en la expresión inicial:

$$E(y_t y_s) = \beta_{00}^2 + E[\mu_{0K}^2] + E[\mu_{0jk}^2] + [(t-t_0)+(s-t_0)][\beta_{00}\beta_{10} + E(\mu_{0k}\mu_{1k})] + \\ + E(\mu_{0jk}\mu_{j1k}) + [(t-t_0)(s-t_0)][\beta_{10}^2 + E(\mu_{1k}^2) + E(\mu_{1jk}^2)]$$

de donde:

$$\sigma(y_t y_s) = \beta_{00}^2 + [(t-t_0)+(s-t_0)]\beta_{00}\beta_{10} + (t-t_0)(s-t_0)\beta_{10}^2 - \beta_{00}^2 - \\ - [(t-t_0)+(s-t_0)]\beta_{00}\beta_{10} - (t-t_0)(s-t_0)\beta_{10}^2 + E[\mu_{0k}^2] + E[\mu_{0jk}^2] + \\ + [(t-t_0)+(s-t_0)][E(\mu_{0k}\mu_{1k}) + E(\mu_{0jk}\mu_{1jk})] + (t-t_0)(s-t_0)\mu_{1k}^2 + \\ + (t-t_0)(s-t_0)\mu_{1jk}^2$$

Simplificando la expresión anterior, tenemos finalmente que la covarianza entre las puntuaciones de dos aplicaciones consecutivas puede expresarse como sigue:

$$\sigma(y_t y_s) = \sigma^2(\mu_{0k}) + \sigma(\mu_{0k}\mu_{1k})[(t-t_0)+(s-t_0)] + \\ + \sigma^2(\mu_{1k})[(t-t_0)(s-t_0)] + \sigma^2(\mu_{ojk}) + \sigma(\mu_{ojk}\mu_{1jk})[(t-t_0)+(s-t_0)] + \\ + \sigma^2(\mu_{1jk})[(t-t_0)(s-t_0)] \qquad (2)$$

Este último resultado analítico es de la máxima importancia. Puesto que la covarianza entre dos medidas dadas del rasgo está fuertemente determinada por la distancia que las separa temporalmente del momento original de medida (t_0), supuesta una covariación no nula de los residuos, la covariación de las puntuaciones se incrementa progresivamente de modo sustancial como un simple resultado del paso del tiempo. Este hecho no es coherente con la experiencia habitual, en la que lo que realmente se observa es una atenuación de las correlaciones.

Con el fin de mantener la ilustración en los términos más sencillos posibles, se adopta una métrica familiar que fija todas las varianzas a 1, valor arbitrario considerado útil a los efectos descritos. Los valores asignados a las covarianzas tratan de reproducir además un contexto plausible en el escenario general de la evaluación del rendimiento. De este modo, se contempla tanto la existencia de una relación

sustancial entre los puntos de partida y la tasa de crecimiento registrada al nivel de escuela y de alumno como la independencia entre ambas características, bien sea parcial o totalmente. De este modo, para las covarianzas de los residuos tanto del nivel alumno ($\sigma_{\mu ojk,\mu 1jk}$) como de centro ($\sigma_{\mu ok,\mu 1k}$) se toman dos valores alternativos iguales, a 0,7 y 0.

Como resultado de tales decisiones, las condiciones o escenarios considerados en el ejemplo son entonces los mostrados en la Tabla I.

TABLA I. Escenarios definidos para la reproducción de patrones de correlación entre medidas sucesivas del rasgo (rendimiento)

Escenario	$\sigma^2_{\mu 0k}$	$\sigma^2_{\mu 1k}$	$\sigma^2_{\mu 0jk}$	$\sigma^2_{\mu 1jk}$.	ε_{tjk}	$\sigma_{\mu 0k,\mu 1k}$	$\sigma_{\mu 0jk,\mu 1jk}$
1	1	1	1	1	1	0,7	0,7
2	1	1	1	1	1	0,7	0,0
3	1	1	1	1	1	0,0	0,0

Para hallar la correlación entre el momento 0 y el momento 1 en el escenario 1, se procede entonces como sigue:

- En primer lugar, obtenemos la covarianza entre las puntuaciones sustituyendo en [2] los valores correspondientes a la varianza de los residuos y a sus covarianzas. Además, t y t_0 son iguales en este caso a 0 y s toma en este ejemplo valor 1. Tenemos entonces que:

$$\sigma(y_0 y_1) = 1 + 0{,}7[(0-0)+(1-0)] + 1[(0-0)(1-0)] + 1 + 0{,}7[(0-0)\ \ (1-0)] + \\ + 1[(0-0)(1-0)] = 3{,}4$$

- A continuación hallamos los valores correspondientes a las varianzas de y 0 e

$$\sigma^2_{y0} = 1 + 1 + (0-0)^2(1) + (0-0)^2(1) + 2(0-0)(0{,}7) + 2(0-0)(0{,}7) + 1 = 3$$

$$\sigma^2_{y1} = 1 + 1 + (1-0)^2(1) + (1-0)^2(1) + 2(1-0)(0{,}7) + 2(1-0)(0{,}7) + 1 = 7{,}8$$

- Entonces podemos establecer que:

$$r_{y0y1} = \sigma_{y0y1} / \sigma_{y0}\sigma_{y1} = \frac{3{,}4}{\sqrt{3}\sqrt{7{,}8}} = 0{,}703$$

Generalizando el procedimiento para los cuatro momentos de medida y los tres escenarios definidos, se obtienen los resultados que se muestran en la Tabla II.

TABLA II. Correlaciones entre medidas sucesivas del rasgo derivadas de las varianzas/covarianzas de los residuos del modelo jerárquico-lineal especificado

	Escenario 1	Escenario 2	Escenario 3
r y0y1	0,703	0,616	0,516
r y0y2	0,680	0,528	0,348
r y0y3	0,660	0,472	0,252
r y1y2	0,896	0,862	0,809
r y1y3	0,898	0,850	0,781
r y2y3	0,951	0,938	0,921

Los resultados muestran que efectivamente el patrón esperado se reproduce en los tres escenarios.

En primer lugar, se constata que inicialmente la correlación entre aplicaciones consecutivas es ligeramente mayor que entre aplicaciones distanciadas en el tiempo, de modo que entre los momentos 0 y 1 la magnitud de la correlación es superior a la registrada entre los momentos 0 y 2 ó 0 y 3. Pero lo que especialmente cabe destacar es que a medida que pasa el tiempo la magnitud de las correlaciones encontradas es mayor, tanto entre aplicaciones sucesivas como entre aplicaciones distanciadas. Esto es, entre los momentos 0 y 1 la correlación es sensiblemente inferior a la encontrada entre los momentos 2 y 3, a pesar de existir la misma *distancia temporal* entre unos y otros. Y otro tanto ocurre si se comparan las correlaciones entre los momentos 0 y 2 con las correspondientes a los momentos 1 y 3. La explicación a este fenómeno, como ya se apuntó, puede hacerse en términos de «varianza disponible», puesto que, si se mantiene el error constante (característica común de los tres escenarios), su importancia relativa decrece a lo largo del tiempo, aumentando la proporción de varianza sistemática y por tanto la magnitud de las correlaciones entre aplicaciones.

Complementariamente, es importante subrayar que el valor absoluto de las correlaciones, manteniéndose constante el patrón descrito, desciende ligeramente al pasar del escenario 1 al 2 y de este último al 3. Es decir, cuando la covarianza de los residuos en los niveles de alumno y escuela es alta y positiva, la covariación entre las puntuaciones de rasgo latente es mayor que cuando los residuos de los dos niveles presentan parcial o totalmente correlaciones nulas.

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

Preguntas de investigación e hipótesis

Las preguntas de investigación a las que se trata de dar respuesta en este trabajo se formulan en los siguientes términos:

- El patrón de correlaciones que frecuentemente observamos en las evaluaciones longitudinales del rendimiento no parece ajustarse a lo que cabe esperar en el nivel en el que acontecen las relaciones causales reales (entre variables latentes). Entonces: ¿se puede explicar dicha desviación como un efecto de la creciente complejidad del rasgo objeto de medida y por tanto de la creciente introducción de incertidumbre o varianza no explicada por el modelo a medida que transcurre el tiempo y se suman evaluaciones sucesivas del rendimiento?
- ¿Cómo afecta en este contexto, cuando se hace uso de modelos multinivel, la naturaleza de la covariación en los residuos al nivel de centro educativo y alumno?

En la práctica de la evaluación del rendimiento, aunque es claro que se pretende medir un único e idéntico rasgo a lo largo del tiempo (rendimiento en una materia o área de competencia), ciertamente ello resulta difícil. Así, basta pensar que en cada aplicación o momento de evaluación adaptamos la prueba a los contenidos del currículo (al que progresivamente se incorporan nuevos aspectos y facetas) y a la capacidad del alumno (presumiblemente creciente). Por tanto, parece razonable asumir que si bien se parte de un rasgo único en la primera aplicación, que nunca deja de ser el principal, al mismo van uniéndose otros rasgos o dimensiones que entran en juego a medida que avanzamos en el tiempo, alejándose así de la unidimensionalidad el objeto de la medida. Y, puesto que conforme pasa el tiempo el rasgo que queremos medir se hace más complejo, el modelo utilizado pierde en igual medida capacidad para determinar el mismo.

Esta última consideración conforma la hipótesis del trabajo: se espera que la simulación propuesta permita identificar la suavización de los patrones crecientes de correlación entre aplicaciones a medida que transcurre el tiempo cuando se considera la complejidad creciente del rasgo.

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

Método

Diseño general de la simulación

La simulación adopta el diseño general de un trabajo de evaluación llevado a cabo en la Comunidad Autónoma de Madrid[1]. En dicho estudio se tomaron medidas de rendimiento a una cohorte de alumnos en el inicio y al término de dos cursos académicos consecutivos.

Para la modelización del rendimiento se define un modelo jerárquico lineal con tres niveles. El modelo incluye como variable predictora, además del tiempo o momento de la medida, el tipo de centro educativo, con cuatro valores distintos. Se trata así de reproducir un contexto plausible para la simulación que atiende al hecho de que el alumnado no se distribuye aleatoriamente entre las distintas escuelas y que el rendimiento medio de éstas, por tanto, varía en cualquier escenario educativo real.

Cuando en el modelo general [0] se incorpora el predictor tipo de centro (W_K) en el nivel macro correspondiente a la escuela, tenemos:

$$\text{Nivel escuela: } \beta_{ok} = \beta_{oo} + \beta_{01}W_K + \mu_{ok} \quad \beta_{1k} = \beta_{10} + \beta_{11}W_K + \mu_{1K}$$

Por lo que, sustituyendo en los niveles anteriores y ordenando la parte fija y aleatoria, el modelo final en el nivel 1 toma la forma siguiente:

$$y_{tjk} = \beta_{00} + \beta_{01}W_K + \beta_{10}(t - t_0) + \beta_{11}W_k(t - t_0) + \mu_{0k} + \mu_{ojk} + \\ + \mu_{1k}(t - t_0) + \mu_{1jk}(t - t_0) + \varepsilon_{tjk} \qquad (3)$$

donde $t\text{-}t_0$ y W_k toman valores entre 0 y 3; es decir, contamos con cuatro momentos de medida y cuatro tipos de escuela diferenciadas en razón de su rendimiento.

Pues bien, la complejidad creciente asociada al rasgo en perspectiva longitudinal puede ser incluida en el modelo [3] multiplicando el término ε_{tjk} (residuo del individuo) por la variable tiempo, de modo que *de facto* se aumenta la importancia o relevancia de posibles dimensiones diferentes de la única e inicialmente considerada. Conviene destacar que la incorporación de este término es un modo de introducir la multidimensionalidad del rasgo en el modelo sin que se prejuzgue ni el

(1) *El valor añadido en educación y la función de producción educativa: un estudio longitudinal;* I+D SEC2003-09742.

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

número ni la naturaleza de las dimensiones involucradas en el mismo. También es claro que, si bien la complejidad creciente del rasgo no puede ser vista como la única fuente de incremento progresivo del error, no es menos cierto que en el contexto de la medida longitudinal del rendimiento escolar constituye la interpretación más plausible. Y en todo caso, al aumento de aquélla se asocia necesariamente el incremento de éste.

El modelo resultante, que presenta la introducción de un término residual modificado, quedaría expresado como sigue:

$$y_{tjk} = \beta_{00} + \beta_{01}W_K + \beta_{10}(t - t_0) + \beta_{11}W_k(t - t_0) + \mu_{0k} + \mu_{ojk} + \mu_{1k}(t - t_0) + \mu_{1jk}(t - t_0) + \varepsilon_{tjk}(t) \quad (4)$$

En consecuencia, se puede definir la capacidad de predicción o determinación del modelo sobre la base de lo que aquí denominamos *factor de complejidad* (FC). Dicho factor expresa la proporción de varianza que un rasgo único es capaz de explicar del total de la varianza registrada:

$$\text{FC} = \frac{\sigma^2_{\theta_j}}{\sigma^2_y}$$

Si FC es igual o cercano a 1, la complejidad del rasgo es nula, es decir, el rasgo es unidimensional y por tanto la importancia relativa del residuo del individuo en el modelo se mantendría constante. Por el contrario, si FC es menor que 1, la complejidad aumenta a medida que el valor de FC es más pequeño, lo que indica la multidimensionalidad del rasgo y tiene su reflejo en el aumento de la importancia relativa del residuo del individuo con el paso del tiempo.

La simulación, por tanto, considera dos modelos para generar puntuaciones de rasgo. A partir de cada uno de ellos, se desarrollan tres escenarios que se distinguen por los valores de las covarianzas de los residuos, que fueron ya definidos y justificados en la ejemplificación de los patrones de correlación que sirvió para plantear el problema. Los tres primeros escenarios, derivados del modelo [3], asumen el carácter constante del residuo del individuo. El modelo [4] se emplea para generar puntuaciones en el contexto de los tres escenarios restantes, caracterizados por el carácter creciente del residuo ε_{tjk}. Por tanto, el esquema general de la simulación es el que muestra la Tabla III.

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

TABLA III. Conjunto de escenarios considerados en la simulación

FC=1				FC<1			
Escenario	$\sigma_{\mu 0k,\mu 1k}$	$\sigma_{\mu 0jk,\mu 1jk}$	ε_{tjk}	Escenario	$\sigma_{\mu 0k,\mu 1k}$	$\sigma_{\mu 0jk,\mu 1jk}$	ε_{tjk}
1	0,7	0,7	Constante	4	0,7	0,7	Creciente
2	0,7	0,0	Constante	5	0,7	0,0	Creciente
3	0,0	0,0	Constante	6	0,0	0,0	Creciente

FC: Factor de complejidad

Muestra y generación inicial de puntuaciones en el rasgo latente

La muestra está conformada por 25.000 sujetos distribuidos en 1.000 escuelas, contándose con 25 sujetos por grupo y un solo grupo en cada escuela. Se definen cuatro tipos de escuelas atendiendo al nivel de rendimiento de su alumnado, siendo las escuelas de tipo I las de menor rendimiento medio y las de tipo IV las de rendimiento superior. Un tercio de la muestra se define como escuelas tipo II y otro tercio, de tipo III . El tercio de escuelas restante se reparte entre las de tipo I y IV .

Fijado lo anterior, las puntuaciones de rasgo se generaron directamente a partir de los modelos [3] y [4].

En lo que se refiere a las *variables predictoras,* ya se indicó anteriormente que $t\text{-}t_0$ y W_k toman valores entre 0 y 3.

Los valores de los coeficientes fueron fijados atendiendo a dos criterios básicos: plausibilidad y sencillez de la escala métrica, determinada por una media igual a 0 y una desviación típica igual a 1. De este modo, se definieron los coeficientes que se describen a continuación:

- La constante o intercepto β_{00} se fija a 0
- Al coeficiente β_{01} se le asigna un valor igual a 0,4, contándose con una diferencia máxima razonable entre las escuelas de menor y mayor rendimiento (igual a 1,2).
- Tanto β_{10} como β_{11} han sido fijados en un valor de 0,2, de nuevo arbitrario pero perfectamente plausible en una escala de media 0 y desviación típica 1. De este modo entre la primera y la cuarta evaluación se considera una tasa total de crecimiento medio igual a 0,8, que es también la diferencia máxima en las tasas de cambio entre las escuelas (la registrada entre las escuelas tipo I y tipo IV).

Por lo que se refiere a la parte aleatoria del modelo, los residuos de los puntos de corte en los niveles alumno y escuela (μ_{0k}, μ_{0jk}) fueron generados directamente como

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

variables aleatorias distribuidas según el modelo N~(0,1). A continuación, se introdujeron como variables independientes en la ecuación de regresión definida para obtener los residuos de las pendientes, de modo que:

$$\mu_{1k} = \mu_{0k} r_{\mu_{1k},\mu_{0k}} + \varepsilon\sqrt{1 - r^2_{\mu_{1k},\mu_{0k}}} \text{ , siendo } \varepsilon \sim N(0,1)$$

$$\mu_{1jk} = \mu_{0jk} r_{\mu_{1jk},\mu_{0jk}} + \varepsilon\sqrt{1 - r^2_{\mu_{1jk},\mu_{0jk}}} \text{ , siendo } \varepsilon \sim N(0,1)$$

Puesto que asumimos que las varianzas de los cuatro tipos de residuos toman valor 1, entonces:

$$r_{\mu_{1k}\mu_{0k}} = \frac{\sigma_{\mu_{1k}\mu_{0k}}}{\sigma_{\mu_{1k}}\sigma_{\mu_{0k}}}; r_{\mu_{1k}\mu_{0k}} = \sigma_{\mu_{1k}\mu_{0k}} \text{, y}$$

$$r_{\mu_{1jk}\mu_{0jk}} = \frac{\sigma_{\mu_{1jk}\mu_{0jk}}}{\sigma_{\mu_{1jk}}\sigma_{\mu_{0jk}}}; r_{\mu_{1jk}\mu_{0jk}} = \sigma_{\mu_{1jk}\mu_{0jk}}$$

Por tanto:

$$\mu_{1k} = \mu_{0k}\sigma_{\mu_{1k},\mu_{0k}} + \varepsilon\sqrt{1 - \sigma^2_{\mu_{1k}\mu_{0k}}} \text{ , siendo } \varepsilon \sim N(0,1)$$

$$\mu_{1jk} = \mu_{0jk}\sigma_{\mu_{1jk},\mu_{0jk}} + \varepsilon\sqrt{1 - \sigma^2_{\mu_{1jk}\mu_{0jk}}} \text{ , siendo } \varepsilon \sim N(0,1)$$

Las covarianzas de los residuos tanto del nivel alumno ($\sigma_{\mu 0jk,\mu 1jk}$) como de centro ($\sigma_{\mu 0k,\mu 1k}$) se fijan en dos valores alternativos (0,7 y 0) dependiendo de los escenarios de simulación considerados.

Finalmente, el término aleatorio del error asociado al sujeto ($\varepsilon_{tjk\ tjk}$) se generó en todos los escenarios de trabajo como distribuido normalmente, con media cero y varianza fijada a 1.

Pues bien, a partir de los modelos y de las especificaciones descritas, se obtuvieron con SPSS 15.0 las puntuaciones generadas iniciales de rasgo para cada sujeto y para cada aplicación en cada uno de los escenarios simulados para su comparación. Se generaron, por tanto, 6 matrices de 25.000 sujetos *4 puntuaciones de rasgo o latentes.

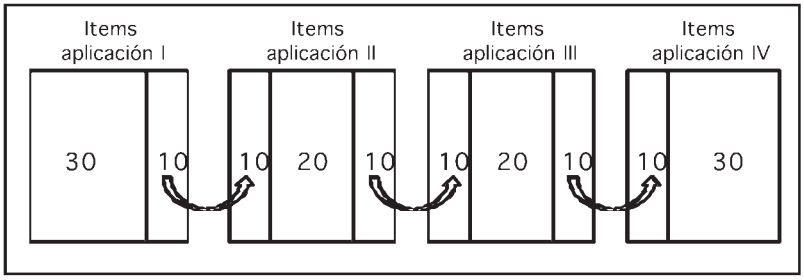
Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Puntuaciones observadas y puntuaciones verdaderas

Para la simulación de las puntuaciones observadas o directas se siguió el proceso inverso al habitual, llevándose a cabo en el marco de la Teoría de Respuesta al Ítem y particularmente empleando un modelo logístico de un parámetro o modelo de Rasch (Rasch, 1960).

La calibración se realizó en el entorno de **R** (Ihaka y Gentleman, 1996) haciendo uso de la librería irtoys. Así, en primer lugar se generaron aleatoriamente valores plausibles del parámetro *b* para un conjunto de 120 ítems aplicados a lo largo de las cuatro aplicaciones simuladas. El esquema definido para cada uno de los tests, incluidas las secciones con ítems de anclaje (cuyos parámetros quedan fijos entre aplicaciones), puede verse en la Figura I, y se tomó de la investigación ya citada.

FIGURA I. Estructura de las pruebas simuladas e ítems de anclaje



La generación aleatoria del parámetro de dificultad se realizó conforme a una Distribución Normal (0,1) con puntuaciones entre -2 y +2, haciendo que aumentasen entre aplicaciones o momentos de medida de forma proporcional al rendimiento.

El parámetro definido en el paso anterior permite determinar la probabilidad de acierto a cada ítem dado el nivel de habilidad de cada sujeto en cada aplicación. Aplicando sobre dichas probabilidades los valores generados para una distribución uniforme, todo valor menor o igual que *p* pasa a ser definido como acierto en un ítem dado, mientras que, si dicho valor es mayor que *p*, el ítem correspondiente pasa a ser codificado como incorrectamente resuelto. Así, se obtiene la puntuación total observada (X) de cada sujeto en cada aplicación y escenario a partir de la suma de aciertos de cada sujeto en cada uno de los 120 ítems.

La suma de las probabilidades de acierto a cada ítem permite la obtención de la puntuación verdadera (T) para cada sujeto en cada aplicación.

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

Resultados

El conjunto de correlaciones obtenidas para cada uno de los seis escenarios definidos en la simulación y para los tres tipos de puntuaciones consideradas se muestra en la Tabla IV.

El primer resultado de interés es el referido a la capacidad de la simulación para reproducir los patrones de correlación de las variables latentes tal y como fueron definidos y ejemplificados a partir del modelo multinivel [0]. Efectivamente, la inspección de las correlaciones obtenidas ahora mediante simulación a partir del modelo [3] en los tres primeros escenarios (en los que el residuo permanece constante) permite constatar que las diferencias son inapreciables cuando se comparan con aquéllas que se estimaron y presentaron en la Tabla II. La conclusión es, por tanto, clara: las correlaciones de las puntuaciones de rasgo generadas se reproducen como cabe esperar por el modelo, es decir, el funcionamiento del modelo [3] es correcto.

TABLA IV. Correlaciones obtenidas en la simulación para los distintos escenarios y tipos de puntuaciones a lo largo de las cuatro aplicaciones supuestas (N=25.000)

	Escenario[a]					
Puntuaciones de rasgo/latentes	**1**	**2**	**3**	**4**	**5**	**6**
r_{y0y1}	,711	,606	,539	,606	,506	,427
r_{y0y2}	,687	,510	,389	,562	,421	,302
r_{y0y3}	,666	,447	,297	,543	,366	,230
r_{y1y2}	,900	,876	,823	,635	,603	,507
r_{y1y3}	,899	,863	,797	,626	,594	,493
r_{y2y3}	,951	,947	,925	,635	,634	,548
Puntuaciones observadas						
r_{y0y1}	,651	,551	,490	,549	,451	,374
r_{y0y2}	,620	,449	,341	,491	,360	,254
r_{y0y3}	,585	,384	,251	,460	,307	,184
r_{y1y2}	,842	,819	,769	,558	,523	,438
r_{y1y3}	,824	,788	,729	,535	,501	,418
r_{y2y3}	,911	,904	,879	,540	,530	,454
Puntuaciones verdaderas						
r_{y0y1}	,674	,574	,509	,567	,469	,390
r_{y0y2}	,638	,464	,351	,508	,371	,264
r_{y0y3}	,601	,396	,257	,474	,317	,192
r_{y1y2}	,859	,838	,790	,566	,533	,445
r_{y1y3}	,839	,804	,747	,544	,508	,425
r_{y2y3}	,922	,916	,892	,544	,536	,458

[a] Escenario 1: $\sigma_{\mu 0k}$, $_{\mu 1k}$ y $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ altas y positivas y ε_{tjk} constante. Escenario 2: $\sigma_{\mu 0k}$, $_{\mu 1k}$ alta y positiva, $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ nula y ε_{tjk} constante. Escenario 3: $\sigma_{\mu 0k}$, $_{\mu 1k}$ y $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ nulas y ε_{tjk} constante. Escenario 4: $\sigma_{\mu 0k}$, $_{\mu 1k}$ y $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ altas y positivas y ε_{tjk} creciente. Escenario 5: $\sigma_{\mu 0k}$, $_{\mu 1k}$ alta y positiva, $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ nula y ε_{tjk} creciente. Escenario 6: $\sigma_{\mu 0k}$, $_{\mu 1k}$ y $\sigma_{\mu 0jk}$, $_{\mu 1jk}$ nulas y ε_{tjk} creciente.

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

Verificado lo anterior, corresponde analizar los resultados que permiten dar respuesta a la pregunta central de la investigación, a saber, la importancia de los efectos del factor de complejidad del rasgo latente así como de las covarianzas de los residuos del modelo en los niveles de alumno y de escuela, en los tres tipos de puntuaciones: de rasgo, observadas y verdaderas.

En primer lugar, conviene considerar las correlaciones de las *puntuaciones de rasgo latente* en los escenarios 4, 5 y 6, con el fin de verificar la hipótesis de su atenuación cuando se comparan con los escenarios 1, 2 y 3. La Tabla IV pone de manifiesto cómo efectivamente, aunque el patrón correlacional se reproduce en lo básico, se atenúan en gran medida tanto las magnitudes de las correlaciones en términos absolutos como las diferencias relativas entre las mismas, lo que apunta al carácter determinante de la única variable diferencial entre ambos bloques: el factor de complejidad. Parece confirmarse por tanto que el modelo utilizado va perdiendo capacidad relativa de determinación a medida que pasa el tiempo; o lo que es lo mismo, que el efecto causal disminuye, tal y como se hipotetizó.

En definitiva, las diferencias entre las correlaciones correspondientes a los escenarios que asumen error constante (1 a 3) y aquéllas derivadas de escenarios que lo asumen creciente (4 a 6) son sensiblemente superiores a las encontradas dentro de cada uno de estos dos bloques. De este resultado puede extraerse como fundamental consecuencia que la importancia o relevancia del factor de complejidad del rasgo latente al determinar las puntuaciones en el mismo parece ser sensiblemente mayor que la presentada por la covarianza entre residuos de ambos niveles. Es decir, como ya se apuntó cuando se analizó la descomposición de las varianzas y covarianzas de los residuos, se comprueba ahora que basta partir de una covarianza entre residuos de segundo y tercer nivel positiva aunque poco elevada o casi nula en la primera aplicación para que la varianza sistemática final, tras mediciones consecutivas,aumente considerablemente, viéndose fuertemente atenuada cuando incluimos un factor multiplicativo del error ligado al tiempo y asociado al efecto que pudieran tener otros rasgos asociados al que deseamos medir (caso de los tres últimos escenarios).

Si se centra el análisis en las *puntuaciones observadas,* las únicas disponibles en un estudio empírico, se verifica cómo los patrones en todos los escenarios son muy similares a los obtenidos en las variables latentes, pero más atenuados. La explicación lógicamente está asociada al error de estimación.

Finalmente, el patrón de correlaciones de las *puntuaciones verdaderas* es prácticamente idéntico al de las puntuaciones observadas. Las escasas diferencias encontradas, que apuntan a una ligera superioridad en la magnitud de las correlaciones en las puntuaciones verdaderas, se deben únicamente al efecto de los errores cometidos al medir.

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

En definitiva, el error de estimación creciente atenúa las diferencias en las correlaciones de los escenarios 4, 5, 6 convirtiendo en irrelevante el factor *momento de aplicación,* variable que determinaba, por el contrario, el patrón de correlaciones en los escenarios 1, 2 y 3. De hecho, para las puntuaciones observadas, en alguno de tales escenarios el error de la medida llega incluso a compensar por completo dicho factor, un efecto por tanto claramente relacionado con la fiabilidad del instrumento.

En resumen, las diferencias encontradas en los patrones de correlaciones de los diferentes escenarios, definidos en función del factor de complejidad del rasgo y de las covarianzas de residuos de los niveles dos y tres del modelo multinivel, se deben principalmente al citado factor, y sólo en segundo lugar, de manera más moderada, a las covarianzas de los residuos. Además, es interesante destacar las diferencias encontradas en la comparación de los tres tipos de puntuaciones en cualquiera de los escenarios. Así, se registra una moderación de las magnitudes de las correlaciones en las puntuaciones observadas y verdaderas debido a la influencia de las características del instrumento utilizado para medir, aunque permanece prácticamente constante el patrón que ya se daba en las puntuaciones de rasgo o latentes.

Conclusiones, limitaciones y prospectiva

De la simulación presentada y de los resultados expuestos, se extraen, a juicio de los autores, algunas importantes aportaciones al campo de la medida en diseños multinivel de medidas repetidas. Se trata, por tanto, de cuestiones que se encuentran en la base de los estudios del valor añadido aplicados a la evaluación educativa, objeto de este monográfico.

En primer lugar, se ha puesto de manifiesto la importancia de conocer y considerar las características diferenciales del patrón de correlaciones de las medidas de rasgo frente a las que frecuentemente se registran en las evaluaciones educativas, que operan sobre la base de puntuaciones observadas. Así, se ha descrito, ejemplificado y simulado, cómo la magnitud de la correlación entre dos medidas es mayor cuanto más se alejan las mismas del punto inicial de la serie de medidas repetidas, siempre y cuando se opere en el nivel de las relaciones causales latentes y se asuma la existencia de un único rasgo. Esto es, a medida que pasa el tiempo y se realizan más aplicaciones o medidas del rendimiento, la correlación entre dos medidas recientes cualesquiera

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

es mayor que entre dos medidas anteriores en el tiempo, a pesar de que exista la misma distancia temporal entre ellas. Y este patrón se reproduce de un modo estable, sólo con muy ligeras variaciones, cuando se definen diferencialmente otros parámetros tales como la covariación de los residuos en el nivel del alumno o la escuela. Quiere decirse que, en este último sentido, basta con partir en la primera aplicación de una correlación positiva aunque poco elevada entre los residuos de nivel dos y tres para que el paso del tiempo multiplique sus efectos.

En segundo lugar, conocido el patrón de correlaciones, se ha indagado en las razones por las que presumiblemente éste no se refleja en los estudios empíricos ocupados de la evaluación longitudinal del rendimiento. Efectivamente, la literatura asume que cabe esperar una correlación de magnitud similar entre medidas que mantienen la misma distancia temporal. Esto es, que, entre las medidas tomadas en la primera y segunda aplicación, la correlación (típicamente no muy alta, entre 0,7 y 0,8) será la misma que entre las tomadas en la cuarta y la quinta aplicación.

Es innegable que las medidas de rendimiento en los estudios empíricos están íntimamente ligadas con los instrumentos utilizados para medirlas. Sin embargo, cabría esperar que, aún asumiendo que no realizamos medidas perfectas del rasgo latente, *rendimiento en matemáticas* por ejemplo, las correlaciones entre dichas medidas repetidas fueran similares a las que presentan las del rasgo latente, aunque más atenuadas, puesto que asumimos error en la estimación. ¿Qué provoca que esto no siempre ocurra? Una respuesta plausible a esta pregunta es la que se deriva de la consideración de lo que aquí se ha denominado *factor de complejidad.*

Es muy probable que, cuando medimos en repetidas ocasiones un mismo rasgo latente, éste varíe en su naturaleza a lo largo del tiempo. Siguiendo con el ejemplo anterior, el rasgo latente *rendimiento en matemáticas* se convierte en más complejo a medida que el currículo escolar mismo gana en complejidad, exige de aptitudes variadas e incluye facetas diferentes a las tomadas como referente en la primera medición. Esto provoca que el modelo de estimación de las puntuaciones de rasgo a partir de las puntuaciones observadas pierda parte de su eficacia. La dimensionalidad del rasgo latente varía y sin embargo seguimos considerando el rendimiento en matemáticas como el mismo y único rasgo. Parece razonable, entonces, que este hecho explique la ausencia de incremento en la magnitud de las correlaciones entre puntuaciones más alejadas del inicio de la serie de medidas, derivada directamente de una estimación del rasgo progresivamente menos precisa.

Una consecuencia clara de lo anterior es la necesidad de profundizar en el estudio de la estructura dimensional de los rasgos latentes y en su reflejo en la dimensionalidad

de las pruebas con el fin de obtener una medida más aproximada a la realidad del rasgo latente y así hacer una correcta y eficaz estimación. En este sentido, un trabajo de este mismo número (Lizasoain y Joaristi, 2009) pone de manifiesto la importancia de construir y usar pruebas que se ajusten a las características del rasgo, tales como su dimensionalidad.

En esta misma línea de este trabajo, también se deriva lo esencial de la equiparación de puntuaciones. Esto es claro al comparar los patrones de correlación en las puntuaciones latentes y observadas. Para estás últimas, la atenuación expresa la estrecha vinculación de la medida con el instrumento utilizado para obtenerla. Así, puesto que en los estudios longitudinales se obtienen puntuaciones en diferentes momentos y con diferentes instrumentos, la evaluación implica necesariamente contar con ítems que sirvan de anclaje entre aplicaciones consecutivas, con el fin de expresar en la misma escala la habilidad del sujeto.

En tercer lugar, cabe destacar la preocupación constante a lo largo de la historia de la psicometría por la fiabilidad clásica, esto es, por los errores de medida. Los resultados de este trabajo reflejan que, asumiendo fiabilidad constante menor que uno, es decir asumiendo que la varianza de los errores de medida permanece constante a lo largo de las cuatro mediciones, las diferencias entre el patrón de correlaciones de las puntuaciones observadas y el de las puntuaciones verdaderas (eliminando el error de medida) son prácticamente inapreciables. Esto quiere decir que lo importante no es tanto eliminar los errores de medida sino que su variabilidad permanezca constante. En definitiva, las diferencias entre unas aplicaciones y otras no dependen tanto del valor del coeficiente de fiabilidad del instrumento, mientras aseguremos su constancia, como de la creciente complejidad del rasgo que se va a medir.

Entre las posibles limitaciones del trabajo, apuntamos aquí el uso de un modelo concreto de generación de las puntuaciones latentes, con unas características específicas: modelo jerárquico lineal de tres niveles y con dos predictores (el momento de aplicación y el tipo de escuela). Esta elección intenta reflejar las tendencias actuales en la estimación del valor añadido pero es evidente que podrían considerarse modelos no lineales o con más predictores (véase por ejemplo, en este mismo número, Castro, Ruiz y López, 2009; Ferraõ, 2009). En este sentido, y también en lo que se refiere a las líneas futuras de trabajo sugeridas por este estudio, destaca el interés de evaluar el efecto de la complejidad creciente de la prueba y de los patrones de covariación de los residuos sobre la estimación misma de medidas del valor añadido.

Referencias bibliográficas

Braun, H. I. (2005). *Using student progress to evaluate teachers: a primer on value-added models.* Princenton, NJ: Educational Testing Service.

Bhaskar, R. (1978). *A realist theory of science.* Harvester Press: Sussex.

Castro, M., Ruíz, C. y López, E. (2009). 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.

Ferraõ, E. (2009). Sensibilidad de las especificaciones de los modelos de valor añadido: la medida del estatus socio-económico. *Revista de Educación,* 348.

House, E. (1991). Realism in Research. *Educational Researcher,* 20 (6), 2-9.

Ihaka, R. y Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299-314.

Lizasoaín, L. y Joaristi, L. (2009). Análisis de la dimensionalidad en los modelos de valor añadido: estudio de las pruebas de matemáticas empleando técnicas factoriales y métodos no paramétricos basados en TRI. *Revista de Educación,* 348.

Martínez Arias, R. (2009). Usos, aplicaciones y problemas de los modelos de valor añadido en educación. *Revista de Educación,* 348.

Martínez Arias, R., Gaviria, J. L. y Castro, M. (2009). Concepto y evolución de los modelos de valor añadido en educación. *Revista de Educación,* 348.

McCaffrey, D. L., Lockwood, J. R., Koretz, D. M. y Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Mónica, CA: RAND Corporation.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research

Sanders, W. L. (2006, October 16). *Comparisons among various educational Assessment Value-Added Models.* Paper presented at The Power of Two–National Value-Added Conference, Columbus, Ohio.

Singer, J. D. y Willet, J. B. (2003). *Applied Longitudinal Data Analysis. Modeling Change and Event Occurrrence.* Oxford: Oxford University Press.

Wiley, E. V. (2006). *A practitioner´s guide to value added assessment.* Tempe, AZ: Educational Policy Studies Laboratory, Arizona State University.

Dirección de contacto: Coral González Barbera. Universidad Complutense de Madrid. Facultad de Educación. Departamento de Métodos de Investigación y Diagnóstico en Educación. Avenida del Rector Royo Villanova s/n. 28040 Madrid. E-mail: cgbarbera@edu.ucm.es

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