Revista de Investigación Educativa, 1999, Vol. 17, n.º 2, págs. 429-443

EL ENFOQUE MULTINIVEL EN LA EVALUACIÓN DE SISTEMAS EDUCATIVOS

José Luis Gaviria Soto
Universidad Complutense de Madrid

INTRODUCCIÓN

Dos son los objetivos de este artículo. Por una parte se pretende presentar la relación existente entre los estudios de eficacia escolar y la evaluación del sistema educativo con el meta-análisis[1]. Por otra se afirma que el enfoque multinivel supone una mejora en la presentación de resultados de evaluación, y se presentan dos ejemplos distintos de cómo la evaluación de sistemas educativos puede beneficiarse de dicho enfoque.

LA EVALUACIÓN EN EL SISTEMA EDUCATIVO

Podemos entender sistema social como uno más de los rasgos de la especie humana dirigidos a producir una mejor adaptación al entorno. Ciertamente se ha caracterizado al ser humano como aquella especie capaz no sólo de adaptarse al entorno, sino de modificarlo sustancialmente en su propio beneficio.

Esta dinámica de influencia de doble dirección, el hombre modifica el entorno, el hombre modifica su conducta para adaptarse a ese nuevo entorno, supone un altísimo grado de complejidad, y fundamentalmente de velocidad en el cambio. Los mecanis-

---

1 Este artículo es una reelaboración, manteniendo el nombre, de la ponencia presentada en el symposium 'Los modelos jerárquicos lineales y su aplicación a la investigación educativa' del IX Congreso Nacional de Modelos de Investigación Educativa, Málaga 1999.

mos genéticos tan eficientes en otras especies no son suficientes para garantizar en la especie humana esa capacidad de adaptación. Por esto es el sistema educativo una parte fundamental de esta maquinaria social.

Este carácter adaptativo del sistema educativo es una nota del mismo de la mayor importancia. Si tenemos ésta característica en consideración, nos daremos cuenta de que el propio sistema educativo debe modificarse en alguno o en todos sus elementos cuando se producen modificaciones del propio sistema social.

El mayor o menor éxito de las distintas sociedades, puede en parte explicarse por el grado de flexibilidad de sus instituciones, de su capacidad para reaccionar ante el cambio. Naturalmente los cambios contextuales no son siempre cataclismos revolucionarios. Las tendencias sociales pueden tener variaciones de distinta «longitud de onda». Ciclos largos, ciclos medios, ciclos cortos. La mayor o menor sensibilidad de los subsistemas sociales a los estímulos de cambio, explican las periódicas crisis de ciertas instituciones.

En este sentido no es extraño que el sistema educativo se encuentre en una permanente crisis en las últimas décadas.

La capacidad de reacción de un sistema social depende del sistema de señales con que éste se comunica con su entorno. Así por ejemplo en el sistema económico el sistema de precios y beneficios es un poderoso sistema de señales enviado a cada industria a las que estas reaccionan aumentando o disminuyendo la producción, cambiando los procedimientos de fabricación, estimulando de ésta forma a toda la industria para una utilización óptima de los recursos escasos. En general, son estos mecanismos homeostáticos los responsables de una correcta adaptación de los sistemas al entorno. La carencia de estos sistemas de señales o la poca eficacia de los existentes es lo que determina los desajustes con el entorno.

La adaptación se produce tanto por el feedback de las salidas, como por la capacidad del sistema para reaccionar a la señal recibida.

En este sentido la evaluación de los sistemas educativos puede considerarse como un subsistema de producción de los mensajes relacionados con la principal salida del sistema, que es el rendimiento de los alumnos.

EVALUACIÓN, EFICACIA ESCOLAR Y META-ANÁLISIS

El mecanismo por el cual el sistema introduce cambios en función de la información retroalimentada depende, como es lógico, de la atribución causal que se haga de los efectos observados. Un sistema tan complejo como el educativo consiste en distintos elementos de diferente nivel de complejidad y agregación, que incluyen desde las características individuales de los alumnos, hasta las unidades escolares organizadas en unidades de nivel superior, pasando por las propias escuelas cuyas cualidades influyen en mayor o menor grado sobre rendimientos observado de los alumnos.

Esa necesaria atribución causal pone en estrechísima relación los procesos de toma de decisiones, en los que la evaluación es pieza fundamental, con los estudios de eficacia escolar. Podemos afirmar sin temor a equivocarnos mucho, que en el ámbito educativo las decisiones respecto a la organización del mismo, a la cantidad de recur-

sos necesarios en cada nivel de relación, a los objetivos estratégicos del sistema, a los cambios en general necesarios, se llevan a cabo sin un conocimiento preciso de las consecuencias que esas decisiones tendrán de hecho en la práctica.

Todas las decisiones políticas están basadas en el convencimiento más o menos subjetivo de que ciertas acciones son en general positivas; como aumentar el número de escuelas, incrementar el número de maestros o sus salarios, aumentar el número de horas de cierta materia, disminuir el número de alumnos por clase, comprar computadoras, o desarrollar la enseñanza en una lengua en lugar de otra, etc.

Ciertamente sería muy útil el conocimiento de lo que se denomina la función de producción, que relaciona los inputs y los procesos con los outputs. Pero sin llegar a ese nivel de precisión, el conocimiento de cuantas relaciones causales entre las características de las escuelas o de los individuos y su rendimiento académico pudieran estar documentadas, sería de la máxima importancia a la hora de determinar cuáles son las decisiones políticas más acertadas para mejorar el rendimiento de los alumnos. Podemos asegurar que pocas decisiones respecto del sistema educativo se basan en el conocimiento siquiera aproximado de los efectos esperables sobre el mismo, y que la mayoría de ellas se fundamentan en la creencia subjetiva y casi mágica de que ciertos resultados se siguen necesariamente ciertas medidas.

En este sentido los estudios de eficacia escolar desempeñan un papel fundamental en el futuro de las ciencias de la educación. En el proceso que va desde la emisión y recepción de las señales relativas al ajuste del sistema educativo al sistema social, a la intervención para modificar aquellas variables que permitan disminuir la distancia entre logros y objetivos, los estudios de eficacia permiten determinar cuáles son esas variables sobre las que intervenir.

Murillo (1999) revisa las distintas aproximaciones metodológicas que se han utilizado en los estudios de eficacia escolar, y presenta un clarísimo ejemplo de la superioridad del enfoque multinivel en los estudios individuales. La mayoría de las técnicas estadísticas, incluidas las que se desarrollan bajo este novedoso enfoque, son técnicas de control estadístico. Se trata de la única opción posible cuando no es posible el control experimental directamente a través del diseño. Hanushek (1997) insiste en la necesidad y en la posibilidad de realizar experimentos controlados para verificar la existencia de relaciones causales. Cuando estudios de este tipo no están disponibles, la afirmación de relaciones causales basadas en estudios individuales no experimentales presenta serios riesgos de sesgo. Es muy posible que características idiosincrásicas aparezcan produciendo un importante efecto causal. Basar decisiones organizativas o de inversión en estos estudios aparentemente concluyentes puede dar lugar a graves errores. El metaanálisis se presenta como una muy seria alternativa a los estudios individuales y a los experimentos cuando éstos no son posibles. El metanálisis no excluye ni a los unos ni a los otros. No puede haber meta análisis sin estudios individuales. El meta análisis puede presentar la evidencia empírica de forma lo suficientemente convincente como para justificar un estudio experimental determinado. El estudio meta analítico ofrece una panorámica en la que es posible estudiar cómo las condiciones contextuales en que se ha desarrollado cada estudio afectan a la relación causal de interés. Castro (1999) realiza una muy interesante presentación de cómo el

estudio meta analítico puede considerarse como un caso particular de estructura multinivel, y de cómo la utilización de los procedimientos de estimación asociados a los modelos jerárquicos lineales producen resultados superiores a las estimaciones meta analíticas clásicas.

Dos ejemplos de enfoque multinivel en la evaluación de sistemas educativos.

El énfasis en los estudios de eficacia escolar está en el establecimiento de relaciones causales. El énfasis en los estudios de evaluación está en la precisión de las medidas obtenidas en cada nivel, y en su estabilidad a lo largo del tiempo.

Hay además dos intereses contrapuestos: la información destinada a la población general, la información de carácter más técnico destinada al uso de los profesionales de la educación. El primer tipo información tiene que ser de naturaleza más amplia general, asociada a áreas reconocibles por el público, como son las disciplinas fundamentales. La segunda debe ser más específica, y más relacionada con destrezas y conocimientos mucho más específicos, de forma que pueda planificarse adecuadamente la intervención didáctica a la vista de esos resultados.

En primer lugar conviene aclarar que los distintos nombres que se han utilizado, como 'Modelos jerárquicos lineales', 'Modelos multinivel', 'Modelos de coeficientes aleatorios' etc., difieren básicamente en el grado de generalidad. De todas estas denominaciones la más genérica es la de modelos multinivel que refleja claramente la naturaleza jerárquica de los datos, pero no prejuzga la forma de las funciones que se utilicen. Como veremos más adelante, esta precisión tiene pleno sentido. Estamos, conviene decirlo, más que ante un modelo estadístico específico o una técnica de análisis, ante todo un enfoque de cómo debe abordarse el análisis de datos. En este sentido, modelos muy diversos caben bajo esta denominación. Ciertamente la mayoría de los modelos propuestos y utilizados son de naturaleza lineal, y de una forma u otra están asociados entre sí. Pero no siempre tiene que ser así. En la siguiente sección veremos cómo un modelo multinivel lineal permite mejorar las estimaciones de las medias escolares de rendimiento, en uno de los casos más elementales que pueden presentar esa estructura. Sin embargo veremos cómo la misma estructura multinivel puede resolver un problema en evaluación utilizando como base un modelo no lineal, en este caso un modelo IRT. Son dos ejemplos de cómo pueden resolverse algunos problemas prácticos abordándolos desde una perspectiva en la que se tiene en cuenta la estructuración natural de los datos.

UN MODELO LINEAL MULTINIVEL

Supongamos que deseamos conocer la media general de rendimiento en una materia de un sistema educativo, y al mismo tiempo obtener estimaciones precisas de las medias en esa misma materia de cada una de las escuelas evaluadas.

Denominamos y_{ij} al resultado en matemáticas, por ejemplo, del alumno i de la escuela j y β_{0j} a la media de rendimiento de la escuela j. El modelo multinivel más sencillo que podemos formular es[2]:

2 Éste mismo modelo y muy interesantes ampliaciones del mismo puede verse en Bryk y Raudenbush (1992).

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \qquad \text{con } \varepsilon_{ij} \approx N\,(0,\ \sigma^2_{\varepsilon})$$

$$\beta_{0j} = \beta_0 + \mu_{0j} \qquad \text{con } \mu_{oj} \approx N\,(0,\ \sigma^2_{\mu 0})$$

$$y_{ij} = \beta_0 + \mu_{0j} + \varepsilon_{ij} \tag{1}$$

Donde β_0 es el parámetro que queremos estimar, por ejemplo la media de rendimiento de cierta población, μ_{0j} es la diferencia de la escuela j con la media general, y ε_{ij} es la diferencia del sujeto i con la media de su escuela j.

De la ecuación 1 vemos que $\bar{y}_{.j} = b_{0j} + \bar{\varepsilon}_{.j} = \beta_0 + (\mu_{0j} + \bar{\varepsilon}_{.j})$ y entonces $\bar{\varepsilon}_{.j} \approx N(0, \frac{\sigma^2_{\varepsilon}}{n_j})$

En la estimación de este parámetro tenemos dos casos distintos. Cuando el número de sujetos de todas las escuelas es igual, y cuando es distinto. De cada escuela conocemos $\bar{y}_{.j}$, la media de rendimiento de la escuela j, y n_j, el número de sujetos de esa misma escuela.

En el primer caso, entonces una estimación razonable de nuestro parámetro es la media de las medias, es decir, $\beta_0' = \frac{\sum_J \bar{y}_{.j}}{J}$, ya que la media de cada escuela es una estimación independiente del parámetro, la media general. Además, la media de cada escuela es también un estimador insesgado del parámetro, ya que las medias de las distribuciones de μ_0 y $\bar{\varepsilon}_{.j}$ y son iguales a cero.

Si tenemos el segundo caso, es decir, cuando las escuelas tienen un número muy distinto de sujetos, entonces el estimador más razonable vendrá dado por $\hat{\beta}_0 = \frac{\sum_J n_j \bar{y}_{.j}}{\sum_J n_j}$

Lo buen o mal estimador que cada media de escuela sea del parámetro, depende de la varianza de cada una de ellas (en cada j). Si un estimador tiene mucha varianza, entonces hay mucha incertidumbre respecto al parámetro y viceversa. La varianza del estimador viene dada por

$$V(\bar{y}_{.j}) = V(\beta_0 + \mu_{0j} + \bar{\varepsilon}_{.j}) = V(\mu_{0j} + \bar{\varepsilon}_{.j}) = \qquad \text{(como } \mu_{oj} \text{ y } \bar{\varepsilon}_{.j} \text{ son independientes)}$$

$= V(\mu_{0j}) + V(\bar{\varepsilon}_{.j}) = \sigma^2_{\mu_0} + \frac{\sigma^2_{\varepsilon}}{n_j}$ donde el primer sumando es la varianza entre grupos y el segundo es la varianza dentro de los grupos. Cada grupo por tanto presenta una incertidumbre distinta en la estimación de β_0. Si llamamos $\Delta_j = \sigma^2_{\mu_0} + \frac{\sigma^2_{\varepsilon}}{n_j}$, entonces el inverso es la precisión del estimador $P(\bar{y}_{.j}) = \frac{1}{\Delta_j} = \Delta_j^{-1}$.

Si en cada caso conociésemos la precisión de la media de la escuela como estimador, entonces una estimación muy natural del parámetro consistiría en ponderar la media de cada escuela por su precisión. $\tilde{\beta}_0 = \dfrac{\sum_J \Delta_j^{-1} \bar{y}_{.j}}{\sum_J \Delta_j^{-1}}$ A este estimador se le conoce como *Weighted Least Square Estimator*. Este estimador está siempre comprendido entre β_0' y $\hat{\beta}_0$.

En efecto, cuando los sujetos dentro de las escuelas son muy parecidos entre sí y casi toda la varianza es entre escuelas, entonces $\sigma^2_{\mu_0} >> \dfrac{\sigma^2_\varepsilon}{n_j}$ y $\Delta_j \cong \sigma^2_{\mu_0}$ y $\Delta_j^{-1} \cong \dfrac{1}{\sigma^2_{\mu_0}}$ y

$$\tilde{\beta}_0 = \frac{\sum_J \Delta_j^{-1} \bar{y}_{.j}}{\sum_J \Delta_j^{-1}} = \frac{\sum_J \frac{1}{\sigma^2_{\mu_0}} \bar{y}_{.j}}{\sum_J \frac{1}{\sigma^2_{\mu_0}}} = \frac{\frac{1}{\sigma^2_{\mu_0}} \sum_J \bar{y}_{.j}}{\frac{1}{\sigma^2_{\mu_0}} J} = \frac{\sum_J \bar{y}_{.j}}{J} = \beta_0'$$

Cuando todas las medias de las escuelas son iguales, entonces $\sigma^2_{\mu_0} = 0$ y toda la varianza es varianza entre sujetos dentro de las escuelas $\Delta_j = \dfrac{\sigma^2_\varepsilon}{n_j}$ y $\Delta_j^{-1} = \dfrac{n_j}{\sigma^2_\varepsilon}$

$$\tilde{\beta}_0 = \frac{\sum_J \Delta_j^{-1} \bar{y}_{.j}}{\sum_J \Delta_j^{-1}} = \frac{\sum_J \frac{n_j}{\sigma^2_\varepsilon} \bar{y}_{.j}}{\sum_J \frac{n_j}{\sigma^2_\varepsilon}} = \frac{\sigma^2_\varepsilon \sum_J n_j \bar{y}_{.j}}{\sigma^2_\varepsilon \sum_J n_j} = \frac{\sum_J n_j \bar{y}_{.j}}{\sum_J n_j} = \hat{\beta}_0$$

Como podemos ver, se trata de un estimador superior a los dos que determinan sus límites, ya que tiene mayor capacidad de adaptación a las condiciones específicas de un determinado conjunto de datos.

Tal vez más interesante es la estimación de las medias de cada escuela.

En esta ocasión se trata de estimar el valor de β_{0j} en cada escuela. También como antes tenemos dos formas de estimación. La media observada de cada escuela es un estimador insesgado de β_{0j} con varianza $\dfrac{\sigma^2_\varepsilon}{n_j}$. Pero también $\bar{\beta}_0 = \dfrac{\sum_J \Delta_j^{-1} \bar{y}_{.j}}{\sum_J \Delta_j^{-1}}$ es un estimador común de β_{0j}.

Un estimador bayesiano de β_{0j} es una combinación óptima de los dos anteriores, y viene dado por $\bar{\beta}_{0j} = \lambda_j \bar{y}_{.j} + (1 - \lambda_j)\, \bar{\beta}_0$. El peso λ_j se le denomina fiabilidad de $\bar{y}_j$ como

estimador de β_{0j} y viene dado por $\lambda_j = \dfrac{\sigma^2_{\mu_0}}{\sigma^2_{\mu_0} + \dfrac{\sigma^2_\varepsilon}{n_j}} = \dfrac{\sigma^2_{\mu_0}}{\Delta_j} = \sigma^2_{\mu_0}\Delta_j^{-1}$, siendo sus límites,

$0 \leq \Delta_j \leq \sigma^2_{\mu_0}$ y $0 \leq \lambda_j \leq 1$.

Así cuando dentro de cada grupo la varianza sea cero, es decir, la autocorrelación sea máxima y toda la varianza es varianza entre grupos, entonces el mejor estimador de β_{0j}, es $\bar{y}_{.j}$ y es entonces cuando $\lambda_j = 1$ y $1 - \lambda_j = 0$, y cuando $\dfrac{\sigma^2_\varepsilon}{n_j} >> \sigma^2_{\mu_0}$ entonces $\lambda_j \to 0$, y $1 - \lambda_j \to 0$. Por tanto $\bar{\beta}_{0j}$ está comprendido entre esos dos extremos. A este coeficiente λ_j se le denomina fiabilidad porque es el cociente entre la varianza verdadera (varianza de la puntuación verdadera) y la varianza total. Este estimador tiene el menor error cuadrático medio (Lindley y Smith, 1972). De hecho es ligeramente sesgado hacia β_0, aunque tenderá a estar más cerca de β_{0j} que cualquier otro estimador. Por esto se le llama 'Shrinkage estimator'. Se le ha denominado estimador bayesiano. Cuando no se conocen las varianzas y se sustituyen por estimaciones, entonces se le denomina estimador empírico bayesiano. De hecho $\bar{\beta}_{0j}$ es la media de la distribución a posteriori de β_{0j} condicionada respecto de las varianzas de σ^2_ε, $\sigma^2_{\mu 0}$ y de los datos. Las ventajas de este estimador son evidentes. Podemos ver que se beneficia tanto del hecho de que los datos estén agrupados en clases, como del hecho de que existe cierta unidad entre los datos en la población. Cuando los datos de un grupo determinado son fiables, es lógico aceptar la información que procede de ese grupo sin más objeciones. Pero cuando los datos de ese grupo no son fiables, fundamentalmente por el escaso número de sujetos en el grupo, entonces la fuerza de la estimación se toma de la información proporcionada por otros grupos. De esta forma combinamos muy acertadamente tanto la información proveniente del centro como la información general.

Si del centro disponemos de alguna información relevante, esta puede ser utilizada para mejorar la estimación. Supongamos que de un centro sabemos que es de naturaleza pública. La variable w_j toma el valor 0 si el centro es privado y 1 si es público. Entonces

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \qquad \text{con } \varepsilon_{ij} \approx N(0, \sigma^2_\varepsilon)$$

$$\beta_{0j} = \beta_0 + \beta_1 w_j + \mu_{0j} \qquad \text{con } \mu_{0j} \approx N(0, \sigma^2_{\mu 0})$$

Ahora los parámetros son: β_0 es la media de centros privados, β_1 es la diferencia entre centros públicos y privados, y μ_{0j} es la diferencia entre la media de los centros del tipo al que pertenece j y el centro j.

$$\bar{y}_{.j} = \beta_0 + \beta_1 w_j + \mu_{0j} + \bar{e}_{.j}$$

Condicionando ahora como antes el estimador de β_{0j}, $\bar{\beta}_{0j}$ es $\bar{\beta}_{0j} = \lambda_j \bar{y}_{.j} + (1-\lambda_j)(\beta_0 + \beta_1 w_j)$.

Obsérvese que ahora la estimación de la media del centro j toma valores entre $\bar{y}_{.j}$ y $(\beta_0 + \beta_1 w_j)$ y que ya no es la media general de todos los centros, sino sólo la media de los centros del mismo tipo que *j*, con lo que cabe suponer una mayor aproximación en la estimación, y por tanto una utilización razonable de la información disponible. Cuantas más características conozcamos del centro mayor será la mejora en el proceso de estimación[3].

Un modelo IRT multinivel

En los procesos de toma de datos de rendimiento académico en la evaluación de sistemas educativos, existe el dilema de tratar de obtener la máxima información interrumpiendo lo mínimo las actividades ordinarias de las clases. Por otra parte, recabar información suficientemente detallada de algunas destrezas supondría la aplicación de un gran número de ítems de cada alumno. Esa información de alto detalle resulta de gran utilidad cuando se quiere tomar decisiones didácticas relevantes respecto de la clase.

Sin embargo cuando se proporciona información personalizada de cada alumno, un número excesivamente grande de dimensiones puede tener como resultado una pérdida de la visión de conjunto.

Bock y Mislevy, (1989) propusieron un modelo de toma de datos que tiene las ventajas de la IRT al mismo tiempo que capitaliza la estructura jerárquica de los datos. Se trata este de un ejemplo de modelo multinivel en el que el modelo de base no es lineal.

En este modelo existen dos niveles relacionados. El nivel 1, el de los sujetos, nos proporciona la probabilidad de que un alumno responda correctamente a un ítem de un área global, por ejemplo, matemáticas. En el nivel 2 el modelo nos da la probabilidad de que un alumno elegido al azar en determinada clase responda correctamente a un ítem correspondiente a una subdimensión específica del área global, por ejemplo, resolución de ecuaciones de segundo grado. Los dos submodelos son agregables, en el sentido de que la media esperada de las puntuaciones de los alumnos en el área global, matemáticas, es igual a la media esperada de las puntuaciones de la escuela en las subdimensiones.

El modelo en el nivel 1 nos proporciona para cada alumno una puntuación en el área global, y en el nivel 2 nos proporciona una puntuación para cada *escuela* en cada una de las subdimensiones de interés.

Los datos

El punto inicial del modelo está en la disposición de la toma de datos.

---

3 Las varianzas σ^2_ϵ y $\sigma^2_{\mu 0}$, o en general los términos de la varianza-covarianza de cada nivel se estiman por procedimientos como *'full maximum likelihood'*, *'restricted maximum likelihood'* o *'Bayes estimation'*. Métodos numéricos como EM o *Fisher Scoring*, en combinación con *Newton Raphson* han sido desarrollados para obtener estos estimadores.

TABLA

Formas de la prueba / Subdimensiones	1	2	...	k	...	K
1	Ítem11	Ítem12	...	Ítem1k	...	Ítem1K
2	Ítem21	Ítem22	...	Ítem2k	...	Ítem2K
...	...	...	...	...	...	...
j	Ítemj1	Ítemj2	...	Ítemjk	...	ÍtemjK
...	...	...	...	...	...	...
J	ÍtemJ1	ÍtemJ2	...	ÍtemJk	...	ÍtemJK

Tenemos por tanto K formas distintas de la prueba, todas ellas con J subdimensiones, teniendo, y esto es muy importante, un sólo ítem de cada subdimensión. El objetivo en la construcción de la prueba es lograr ítems tan paralelos como sea posible en cada subdimensión.

Cada alumno recibe una de las formas de la prueba. Supongamos que tenemos H grupos en total. La puntuación que el sujeto i del grupo h recibe en el ítem jk (subdimensión j forma k) es, x_{hijk} que tomará los valores 0 ó 1.

Cómo la variable toma estos valores, depende de una estructura latente que pasamos a describir. Imaginemos que el rendimiento en un ítem determinado, el ítem jk viene dado por una variable latente continua. Si en un determinado individuo esa variable rebasa el valor de uno de umbral, entonces la variable observable x_{hijk} toma el valor 1, mientras que si la variable latente no llega a ese umbral el valor observado es cero. Esta variable latente continua depende a su vez de otras dos variables latentes, que son la habilidad general en el área global de ese individuo, (matemáticas) y la habilidad específica que ese ítem mide (resolución de ecuaciones de segundo grado). Se trata en el fondo de un modelo de factores múltiples, en el que la única particularidad es que la variable explicada es también latente.

El modelo tiene la siguiente forma:

$$z_{hijk} = \alpha_{jk}\theta_{hi} + \delta_{jk}\phi_{hij} + e_{hijk}$$

donde z_{hijk} es la variable latente continua que refleja la habilidad del sujeto i del grupo h para responder correctamente al ítem de la subdimensión j del cuaderno k,

θ_{hi} es la habilidad del sujeto i del grupo h en el área general, en este caso matemáticas,

ϕ_{hij} es la habilidad del sujeto i del grupo h en la subdimensión j, en este caso, por ejemplo resolución de ecuaciones de segundo grado.

α_{jk} y δ_{jk} son las cargas factoriales correspondientes al factor general θ y al factor específico ϕ_j respectivamente,

e_{hijk} es el residuo. El umbral, que está relacionado directamente con la dificultad del ítem jk es γ_{jk}.

Los supuestos de este modelo son los siguientes:

1. $z_{hijk} \approx N(\theta_{h'}, 1)$ en la población de alumnos

2. $\theta_{hi} \approx N(\theta_{h'}, \sigma^2)$ dentro del grupo h

3. $\phi_{hij} \approx N(\phi_{hj'}, \sigma_j^2)$ dentro del grupo h

4. $\text{cov}(\theta_{hi'}, \phi_{hij}) = 0$ y $\text{cov}(\phi_{hij'}, \phi_{hij'}) = 0$ dentro del grupo h

5. $\theta_h \approx N(0, \varsigma^2)$ en la población de grupos

6. $\theta_{hj} \approx N(0, \varsigma_j^2)$ en la población de grupos

7. $\text{cov}(\theta_{h'}, \phi_{hj}) = 0$ y $\text{cov}(\phi_{hj'}, \phi_{hj'}) = 0$ en la población de grupos.

De estos supuestos se sigue que en la población sin condicionar, para el factor general, $\theta \approx N(0,\sigma^2 + \varsigma^2)$ donde la varianza es la suma de la varianza dentro de los grupos y la varianza entre los grupos. Del mismo modo, para cada uno de los factores específicos tendremos $\phi_j \approx N(0, \sigma^2_j + \varsigma^2_j)$. Como se trata de un modelo con variables latentes, no existen unidades naturales, por lo que se puede determinar, sin perder generalidad, que $\sigma^2 + \varsigma^2 = 1$ y $\sigma^2_j + \varsigma^2_j = 1$. Por lo tanto podemos incluir además los siguientes supuestos:

8. $\varsigma^2 = 1 - \sigma^2$

9. $\varsigma_j^2 = 1 - \sigma_j^2$

10. $e_{hijk} \approx N(0,1 - \alpha^2_{jk} - \delta^2_{jk})$ siendo estos términos de error independientes entre las personas, los grupos, las subdimensiones y las formas.

11. $\lambda_i = \dfrac{\delta_{jk}}{\alpha_{jk}}$

Este último supuesto indica que para cada destreza específica *j*, la aportación del factor específico al ítem es la misma para todos los ítems que miden esa subdimensión en todas las k formas. Ciertamente son bastantes supuestos, pero no debemos olvidar que estamos incluyendo supuestos referidos a dos niveles distintos. Y por otra parte no son supuestos muy distintos de los habituales en otros modelos más sencillos de IRT.

A partir de estos supuestos podemos ver cómo obtenemos la probabilidad para el nivel 1 y para el nivel 2.

El modelo para el nivel I

Comenzamos determinando cómo se deriva el modelo IRT para el nivel 1, es decir, para el nivel individual. Se trata por tanto de ver cómo se relacionan las respuestas de

un sujeto a un cuaderno con su puntuación latente θ_{hi} en el área global que estamos evaluando.

$$z_{hijk} = \alpha_{jk}\theta_{hi} + \delta_{jk}\phi_{hij} + e_{hijk} = \alpha_{jk}\theta_{hi} + e^0_{hijk} \text{ donde } e^0_{hijk} = \delta_{jk}\phi_{hij} + e_{hijk} \text{ y } e^0_{hijk} \approx N(0,1 - \alpha^2_{jk}).$$

Obsérvese que lo que estamos modelando es una cualquiera de las columnas de la tabla 1. Podemos ver que en esencia este modelo afirma que la puntuación de un sujeto en la variable latente z depende de su puntuación en un factor latente, en este caso un factor general en matemáticas, y de la relación α de este factor con la variable z. Es cierto que también la capacidad en el factor específico afectará a ese rendimiento, pero en este caso ese factor queda subsumido en el nuevo término de error.

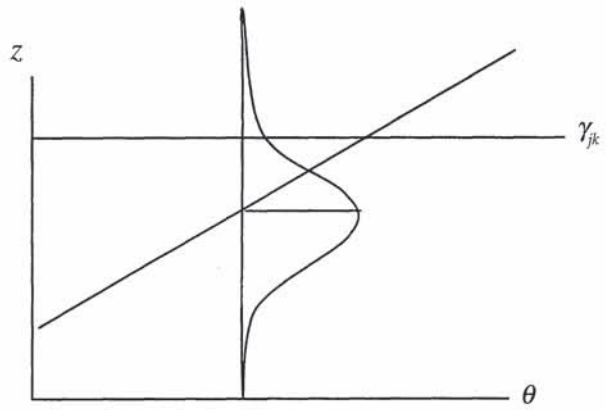

Ilustración 1

En la ilustración 1 vemos cómo se obtiene la probabilidad de que el sujeto responda correctamente al ítem en cuestión. La relación entre θ y z viene dada por la recta de regresión que aparece representada, con una pendiente igual a α_{jk}. La varianza de la distribución condicional será por tanto $1-\alpha^2_{jk}$. Preguntarnos por la probabilidad de que el ítem se responda correctamente es lo mismo que preguntarnos por la probabilidad de que dada la puntuación θ_{hi} del individuo *i* del grupo *h* ese sujeto tenga una puntuación en z por encima de γ_{jk}. Como vemos en la ilustración esa probabilidad es igual al área de la distribución condicional que queda por encima del umbral. Se trata por tanto de

$$P(x_{hijk} = 1 \,|\theta_{hi}) = P(z_{hijk} > \gamma_{jk}) = \frac{1}{\sqrt{2\pi(1-\alpha^2_{jk})}} \int_{-\infty}^{\gamma_{jk}} \exp\left(-\frac{(z-\alpha_{jk}\theta_{hi})^2}{2(1-\alpha^2_{jk})}\right)$$

Haciendo el oportuno cambio de variable tenemos

$$P(x_{hijk} = 1 \,|\theta_{hi}) = \int_{-\infty}^{a_{jk}(\theta_{hi}-h_{jk})} \exp(-\frac{t^2}{2})dt \qquad (2)$$

donde $\alpha_{jk} = \frac{\alpha_{jk}}{\sqrt{1-\alpha_{jk}^2}}$ y $b_{jk} = \frac{\gamma_{jk}}{\alpha_{jk}}$

La ecuación 2 nos da la probabilidad de respuesta al ítem *jk* condicionada respecto de la capacidad θ_{hi} de la persona *i* del grupo *h*. Esto es por tanto un modelo IRT para la capacidad en el área global que medimos, en nuestro ejemplo matemáticas.

Para realizar la estimación de los parámetros de este modelo nos basamos en el hecho de que los errores e^0_{hijk} son independientes entre subdimensiones, y en los patrones de respuestas, de ceros y unos, que nos proporciona cada alumno. Así, dado el patrón de respuestas de un sujeto determinado, la función de verosimilitud de dicho patrón, si llamamos $\Phi_{jk}(\theta_{hi})$ a la ecuación 2, viene dada por

$$P(x_{hi1k},\ldots,x_{hiJk}\ |\theta_{hi}) = \prod_j (\Phi_{jk}(\theta_{hi}))^{x_{hijk}} (1-\Phi_{jk}(\theta_{hi}))^{1-x_{hijk}}$$

Para que esta ecuación sea cierta tiene que haber sólo un ítem por cada una de las subdimensiones, de otra forma no se cumpliría el supuesto de independencia local. A partir de esta ecuación y utilizando los procedimientos de máxima verosimilitud marginal descritos por Mislevy, tal y como están implementados en programas como BILOG pueden obtenerse estimaciones bastante precisas de estos parámetros.

El modelo para el nivel 2

Los mismos supuestos iniciales nos permiten derivar un modelo para las unidades de nivel 2.

$$z_{hijk} = \alpha_{jk}\theta_{hi} + \delta_{jk}\phi_{hij} + e_{hijk} = \alpha_{jk}(\theta_{hi} + \theta_h - \theta_h) + \alpha_{jk}\frac{\delta_{jk}}{\alpha_{jk}}(\phi_{hij} + \phi_{hj} - \phi_{hj}) + e_{hijk}$$

$$= \alpha_{jk}(\theta_h + \frac{\delta_{jk}}{\alpha_{jk}}\phi_{hj}) + \delta_{jk}(\phi_{hij} - \phi_{hj}) + e_{hijk} =$$

$$= \alpha_{jk}(\theta_h + \lambda_{jk}\phi_{hj}) + e^*_{hijk} \qquad (3)$$

donde $e^*_{hijk} = \delta_{jk}(\phi_{hij} - \phi_{hj}) + e_{hijk}$ y $e^*_{hijk} \approx N\left(0, 1-\alpha^2_{jk}(1-\sigma^2) - \delta^2_{jk}(1-\sigma^2_j)\right)$

Obsérvese que en la ecuación 3, a partir de la que obtendremos la probabilidad de que un alumno elegido al azar del grupo h responda correctamente al ítem *jk* es en el fondo una ecuación de regresión entre dos variables latentes, una de las cuales, la independiente, tiene una naturaleza compuesta. En efecto, sus componentes son θ_h que es la media del grupo *h* en el factor general, y ϕ_{hj} que es la media del grupo h en el factor específico *j*. Para simplificar la notación llamaremos ϕ^*_{hj} a esa variable compues-

ta. Pues bien, de forma muy similar a la anterior obtenemos la probabilidad de respuesta correcta al ítem *jk* de un individuo elegido al azar del grupo *h* si este tiene como media en la variable compuesta ϕ^*_{hj}.

$$P(x_{hijk}=1 \mid \phi^*_{hj}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a^*_{jk}(\phi^*_{hj}-b^*_{jk})} \exp(-\frac{t^2}{2})dt \qquad (4)$$

$$\text{donde } a^*_{jk} = \frac{\alpha_{jk}}{\sqrt{1-\alpha^2_{jk}(1-\sigma^2)-\delta^2_{jk}(1-\sigma^2_j)}}$$

$$\text{y } b^*_{jk} = \frac{\gamma_{jk}}{\alpha_{jk}}.$$

Nótese que el parámetro de dificultad es el mismo que en el modelo anterior, lo que nos permitirá poner más adelante en relación los dos submodelos.

En este caso la variable compuesta ϕ^*_{hj} tiene una distribución $\phi^*_{hj} \approx N(0,\sigma^2 + \lambda^2_j\sigma^2_j)$. En esta última expresión de hecho desconocemos todos los elementos que determinan la varianza. Por eso lo que se hace es utilizar $\phi^{**}_{hj} \approx N(0,1)$ y se estiman a^{**} y b^{**}, y en un paso posterior se reescalan.

La relación entre los niveles

Los ítems en el nivel 1 se calibran en una escala que arbitrariamente se define como $\theta \approx N(0,1)$. Sin embargo la escala del nivel 2 no puede definirse arbitrariamente de la misma manera, ya que existe una relación entre la escala del nivel 1 y la del nivel 2, y por eso la escala 'natural' del nivel 2 viene dada por $\phi^*_{hj} \approx N(0,\sigma^2 + \lambda^2_j\sigma^2_j)$. Dado que no conocemos los elementos de la varianza, realizamos un reescalamiento por otros medios.

Para empezar, la relación entre la escala 'real' o 'natural' y la provisional ϕ^{**}_{hj} viene dada por

$$\phi^{**}_{hj} = \frac{\phi^*_{hj}}{C_j} \text{ y } a^{**}_{jk} = a^*_{jk}C_j \text{ con } C_j = \sqrt{(\sigma^2 + \lambda^2_j\sigma^2_j}$$

El hecho importante para el reescalamiento es que como vimos anteriormente $b^{**}_{jk} = b^*_{jk}$.

Por tanto al hacer las estimaciones en el modelo de nivel 1 ya obtuvimos un $\hat{b}_{jk}$ luego una estimación adecuada de C_j será $\hat{C}_j = \frac{1}{K}\sum\frac{\hat{b}_{jk}}{\hat{b}^{**}_{jk}}$ por lo que $\hat{\phi}^*_{hj} = \hat{\phi}^{**}_{hj}C_j$ y

$$\hat{a}^*_{jk} = \frac{\hat{a}^{**}_{jk}}{C_j} \text{ y } \hat{b}^*_{jk} = \frac{\hat{b}^{**}_{jk}}{C_j}.$$

De esta forma vemos en un primer paso que la relación definida entre los niveles nos permite situar las estimaciones en escalas relacionadas.

Un segundo paso consiste en verificar la agregabilidad de los niveles. Esto significa que la puntuación del nivel de grupo en el área global (matemáticas) puede ser calculada en dos formas, bien a través de la obtención de la media de las puntuaciones globales de los individuos del grupo, o bien a través de la media de las medias de subdimensión del grupo.

La media de las subdimensiones j en un grupo, después de reescalar convenientemente, y si no hubiese errores en la estimación y el reescalamiento sería

$$\frac{\sum_j \phi^*_{hj}}{J} = \frac{\sum_j (\theta_h + \lambda_j \varphi_{hj})}{J} = \frac{J\theta_h + \sum_j (\lambda_j \varphi_{hj})}{J} = \theta_h + \frac{\sum_j (\lambda_j \varphi_{hj})}{J}$$

Pero como sí hay error, para una escuela dada lo que hacemos es calcular la esperanza matemática de esa escuela.

$$E(J^{-1}\sum_j \phi^*_{hj} \mid \theta_h) = E\left(\theta_h + \frac{\sum_j (\lambda_j \varphi_{hj})}{J} \mid \theta_h\right) = \tag{5}$$

$$= \theta_h + \frac{\sum_j \lambda_j E(\varphi_{hj} \mid \theta_h)}{J}$$

y como la covarianza de ϕ_{hj} y θ_h es cero, $E(\phi_{hj} \mid \theta_h) = E(\phi_{hj})$ y esto último por definición es cero, luego la ecuación 5 es igual a θ_h. Si tomamos aproximaciones en lugar de parámetros, lo que tenemos es $E(\frac{\sum_j \hat{\phi}^*_{hj}}{J} \mid \theta_h) \cong \hat{\theta}_h$

De esta forma tenemos en relación las estimaciones independientes de los dos niveles. Esta es además una forma muy fácil de comprobar la adecuación del modelo en cada grupo. Naturalmente, sólo si las medias de los grupos coinciden coincidirán las medias en la población.

CONCLUSIONES

Habiendo establecido las relaciones entre el meta-análisis, los estudios de eficacia escolar y la evaluación, se presentan dos ejemplos de cómo el enfoque multinivel puede mejorar las estimaciones de algunos parámetros en la evaluación. (En Murillo (1999) y Castro (1999) se abordan las posibilidades del análisis multinivel en los otros dos campos).

Cuando el enfoque multinivel se utiliza en la estimación de los parámetros fijos (media general, media de las unidades de nivel 2) los estimadores obtenidos son superiores a los estimadores OLS tradicionales.

En otro ejemplo hemos visto cómo el enfoque multinivel en el contexto de los modelos IRT permite satisfacer las necesidades de la evaluación respecto de la información sobre el rendimiento medio de los grupos en ciertas subdimensiones y de la información de las unidades de nivel 1, los sujetos, en áreas más amplias; y de cómo es posible la relación entre los distintos niveles.

REFERENCIAS

Bryk, A.S. y Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Ed SAGE London.

Lindley, D.V., y Smith, A.F.M. (1972). Bayes estimation for linear models. Journal of the Royal Statistical Society, Series B. 34, 1-41.

Mislevy, T.J. y Bock, R.D. (1989). *A Hierarchical Item-Response Model for Educational Testing* en R. Darrell Bock (Ed.). Multilevel Analysis of Educational Data. Academic Press Inc. San Diego (1989).

Castro Morera, M. (1999). Modelos multinivel aplicados al meta-análisis RIE, Volumen 17, nº 2 (1999).

Murillo Torrecilla, F.J. (1999). Los modelos jerárquicos lineales aplicados a la investigación sobre eficacia escolar, RIE, Volumen 17, nº 2 (1999).