Psicothema 2008. Vol. 20, nº 3, pp. 497-503
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Una aplicación de la estimación Bayes empírica para incrementar la fiabilidad de las puntuaciones parciales

Paula Elosua Oliden
Universidad del País Vasco

Este trabajo presenta una aplicación de un método estadístico que permite incrementar la fiabilidad de las puntuaciones parciales asociadas a un test. Es una estimación Bayes empírica, generalización de la ecuación de Kelley. Las puntuaciones estimadas se apoyan en la información contenida en el resto de escalas parciales que componen el test. El procedimiento es descrito y discutido aplicándolo a un test intencionadamente multidimensional. El incremento en la fiabilidad de las puntuaciones en cada una de las escalas parciales podría considerarse equivalente a la fiabilidad alcanzable alargando la longitud del test en un 58,06%.

An application of the empirical Bayes estimation to improve the reliability of subscores. In this paper, we show a subscore augmentation procedure that improves the reliability of subscales. The approach uses empirical Bayes estimations. This is a generalization of Kelley's formula. The estimates are based on the information from other related scales on the test. We describe the procedure and we apply it to a multidimensional test. The reliability of the subscores increased, and the improvement could be considered equivalent to a 58.06% increase in the length of the test.

La interpretación de la puntuaciones obtenidas tras la aplicación de un test sólo será formal y éticamente correcta si está basada en estudios que garanticen que tanto su fiabilidad como su validez son óptimas en relación al uso propuesto (Elosua, 2003). En muchas situaciones los tests proporcionan dos tipos de puntuaciones; puntuaciones parciales referidas a diferentes escalas, y puntuaciones totales o generales formadas por una combinación de todos los ítems o escalas que definen el test. En el campo de la evaluación psicológica, por ejemplo la inclusión de dimensiones diferentes pero relacionadas con un mismo constructo es una práctica común; las puntuaciones parciales permiten la construcción de perfiles que aportan información diagnóstica individual. Podríamos citar como muestra, las dimensiones o escalas del autoconcepto (AFA-A, Musitu, García y Gutiérrez, 1997), las escalas o dimensiones relacionadas con los trastornos de los hábitos alimentarios (EDI-2; Garner, 2001), o las dimensiones relacionadas con la personalidad (16PF; Catell, 1989). Todas ellas proporcionan perfiles individuales basados en escalas diferentes a la par que permiten obtener un indicador del nivel general de la persona en el constructo medido. En el campo de la evaluación educativa las puntuaciones parciales vendrían asociadas con la especificación de diferentes áreas de contenido o distintos procesos cognitivos involucrados en la evaluación de un dominio general u objetivo curricular. La información diagnóstica contenida en ellas permitiría una evaluación individual o grupal más profunda que la basada en el análisis de la puntuación total, que por su carácter general en ningún caso discrimina entre contenidos o procesos.

Sin embargo, las características métricas de las puntuaciones totales o las características métricas de las puntuaciones parciales es diferente, y lo es también el propio centro de interés durante el proceso de construcción del test. El uso de la puntuación total como medida indicativa del nivel de una persona en un rasgo psicológico o contenido educativo tiene que venir avalada por una adecuada representación por parte de los ítems del objeto a evaluar; pero, por supuesto, el uso de cada una de las puntuaciones parciales tiene también que estar sujeta a esa necesidad. La posibilidad de utilizar puntuaciones totales y puntuaciones parciales en un mismo test hace necesaria la confluencia de dos requerimientos diferentes, generalidad y concreción; piénsese por ejemplo en la evaluación final asociada a un proceso de formación e instrucción. El test a utilizar debería de cubrir todos los aspectos tratados, sin embargo, si se quiere añadir un carácter diagnóstico al test, a modo de información diferenciada sobre cada uno de ellos, habría que profundizar en las diferentes áreas de contenido que definen el objetivo curricular; por lo tanto cada una de las escalas parciales debería de ser tan específica y sintetizada como fuera posible.

Este doble requerimiento, información general/diagnóstico parcial, afecta directamente a las propiedades formales del test; entre ellas, a la fiabilidad. Una de las características del coeficiente de fiabilidad (véase Muñiz, 1992; Santisteban, 1990) es su dependencia del número de ítems; a medida que se prolonga la longitud del test y siempre y cuando se mantenga la homogeneidad de contenido de los ítems, aumenta el coeficiente de fiabilidad. Debido a esta propiedad las puntuaciones totales son en general más fiables que las puntuaciones parciales. Sin embargo, si queremos utilizar

Fecha recepción: 25-9-07 • Fecha aceptación: 19-11-07
Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
E-mail: paula.elosua@ehu.es

estas últimas su fiabilidad debería de estar garantizada, y aumentar el número de ítems de las escalas parciales con ese fin no es una solución plausible.

La necesidad del incremento de la fiabilidad de las puntuaciones parciales, si bien es un hecho dentro de la evaluación psicológica (ver manuales de los tests citados) está siendo considerada una exigencia en el campo de la evaluación educativa. De hecho, el interés en el incremento de la fiabilidad de las puntuaciones parciales surgió cuando el Departamento De North Carolina (EE.UU.) quiso utilizar las puntuaciones obtenidas en las distintas áreas de contenido de un test con propósitos diagnósticos en lugar de utilizar simplemente la puntuación total. Esta tendencia a utilizar las puntuaciones con objetivos diagnósticos está siendo un imperativo en el campo de la investigación psicométrica actual (Leighton y Gierl, 2007).

Ante este creciente interés se están buscando soluciones que permitan incrementar la fiabilidad de las puntuaciones parciales con el fin de poder utilizarlas con garantías. Dado que la solución del aumento del número de ítems es impracticable, la solución estadística que está siendo evaluada actualmente consiste en incrementar la fiabilidad de las puntuaciones parciales haciendo uso de la información contenida en el resto de escalas que componen la prueba. El procedimiento utilizado por el *Educational Testing Service* es una generalización de la conocida ecuación de regresión de Kelley (1927) para la estimación de la puntuación verdadera en el marco de la teoría clásica de tests (TCT). En el modelo de la TCT la puntuación observada en un test por un sujeto *j* (X_j) es una variable aleatoria compuesta por la puntuación verdadera, o valor esperado de la puntuación observada (V_j= E[X_j]) y un componente de error aleatorio (E_j) de cuya influencia dependerá el coeficiente de fiabilidad de test.

La ecuación de Kelley (ecuación 1) permite estimar la puntuación verdadera de un sujeto (V_j) por medio de una regresión a la media ponderada por el coeficiente de fiabilidad del test ($r_{XX'}$). A medida que mejora el coeficiente de fiabilidad la contribución de la puntuación observada a la puntuación verdadera aumenta; y en los casos en que la fiabilidad del test es baja la estimación se aproxima a la media grupal.

$$\hat{V}_j = r_{xx'}X_j + (1 - r_{xx'})\overline{X} = \overline{X} + r_{xx'}(X_j - \overline{X}) \tag{1}$$

Donde $r_{XX'}$ es el coeficiente de fiabilidad del test
$\hat{V}_j$ es la puntuación verdadera estimada al sujeto j
X_j es la puntuación empírica observada del sujeto j
$\overline{X}$ es la media aritmética de la puntuación observada

La generalización de la ecuación de Kelley para la estimación de la puntuación verdadera a partir de la información contenida en varias escalas parciales se podría expresar del siguiente modo:

$$\hat{\mathbf{V}}_j = \overline{\mathbf{X}} + \mathbf{B}(\mathbf{X}_j - \overline{\mathbf{X}}) \tag{2}$$

Donde **B** es la matriz de pesos
$\hat{\boldsymbol{V}}_j$ es el vector de puntuaciones parciales verdaderas estimadas para el sujeto j
$\mathbf{X}_j$ es el vector de puntuaciones parciales observadas para el sujeto j
$\overline{\mathbf{X}}$ es el vector de medias observadas de las escalas parciales

En el caso de que la matriz **B** sea una matriz identidad (**B= I**) las puntuaciones observadas tendrían una fiabilidad perfecta y serían equivalentes a las puntuaciones verdaderas. Si la matriz de pesos fuera la matriz nula (**B= 0**) la fiabilidad de las escalas parciales sería 0, por tanto las puntuaciones verdaderas serían estimadas por las medias aritméticas parciales obtenidas en el grupo.

La estimación del vector de puntuaciones verdaderas parciales conocidas las puntuaciones observadas parciales (ecuación 2) es similar a una estimación Bayes empírica en la que se desean estimar los valores esperados de las puntuaciones verdaderas condicionados sobre las puntuaciones observadas $E(\mathbf{V_j} \mid \mathbf{X_i})$.

Conocidas las propiedades de la distribución normal multivariada se trata de estimar los parámetros de la distribución condicional de una variable perteneciente a una población multinormal con respecto a otra variable de la misma población.

Siguiendo a Morrison (1967), si X e Y siguen una distribución normal multivariada con parámetros:

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_Y\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY}\\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY}\end{bmatrix}\right) \tag{3}$$

La distribución condicional de una variable respecto a otra de la misma población sigue una distribución normal multivariada con los siguientes parámetros:

$$E\left[\mathbf{X}|\mathbf{Y}\right] = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{XY}^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}_Y\right) = \mathbf{BY} + a \tag{4}$$

$$\boldsymbol{\Sigma}\left[\mathbf{X}|\mathbf{Y}\right] = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}\,\boldsymbol{\Sigma}_{YX} \tag{5}$$

Trasladando estas propiedades al modelo de la teoría clásica de tests, el problema se centraría en la distribución multivariada definida por las puntuaciones verdaderas (**V**) y las puntuaciones observadas (**X**).

$$\begin{bmatrix}\mathbf{V}\\ \mathbf{X}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{\mu}_V\\ \boldsymbol{\mu}_X\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\Sigma}_{VV} & \boldsymbol{\Sigma}_{VX}\\ \boldsymbol{\Sigma}_{XV} & \boldsymbol{\Sigma}_{XX}\end{bmatrix}\right) \tag{6}$$

Dado que según el modelo de la TCT (ver Muñiz, 1992; Santisteban, 1990), a) la media aritmética de las puntuaciones verdaderas es igual a la media aritmética de las puntuaciones observadas ($\mu_V = \mu_X$); b) la covarianza entre las puntuaciones verdaderas es igual a la covarianza entre las puntuaciones observadas ($\sigma_{XX'} = \sigma_{VV'}$); c) la varianza entre puntuaciones verdaderas es igual al producto entre el coeficiente de fiabilidad y la varianza de las puntuaciones observadas ($\sigma^2_V = \rho_{XX'}\,\sigma^2_X$); y d) la covarianza entre las puntuaciones observadas y las puntuaciones verdaderas es igual a la varianza de las puntuaciones verdaderas ($\sigma^2_V = \sigma_{XV}$), la ecuación 6 podría escribirse como:

$$\begin{bmatrix}\mathbf{V}\\ \mathbf{X}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\Sigma}_{VV} & \boldsymbol{\Sigma}_{VV}\\ \boldsymbol{\Sigma}_{VV} & \boldsymbol{\Sigma}_{XX}\end{bmatrix}\right) \tag{7}$$

En estas condiciones la esperanza y la varianza del vector de puntuaciones parciales verdaderas para un sujeto, condicionada

sobre el vector de puntuaciones parciales observadas vendrían dadas por,

$$E\left[\mathbf{V}_j \middle| \mathbf{X}_j\right] = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{VV}\,\boldsymbol{\Sigma}_{XX}^{-1}\left(X_j - \boldsymbol{\mu}_X\right) \tag{8}$$

$$\boldsymbol{\Sigma}\left[\mathbf{V}_j \middle| \mathbf{X}_j\right] = \boldsymbol{\Sigma}_{VV} - \boldsymbol{\Sigma}_{VV}\,\boldsymbol{\Sigma}_{XX}^{-1}\,\boldsymbol{\Sigma}_{VV} \tag{9}$$

Por lo tanto el vector de puntuaciones parciales verdaderas, puntuaciones mejoradas (*augmented scores*) o puntuaciones estimadas, podría obtenerse a través de,

$$\hat{\mathbf{V}}_j = \overline{\mathbf{X}} + \mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}\left(\mathbf{X}_j - \overline{\mathbf{X}}\right) = \overline{\mathbf{X}} + \mathbf{B}\left(\mathbf{X}_j - \overline{\mathbf{X}}\right) = \mathbf{B}\mathbf{X} + \mathbf{a} \tag{10}$$

y la matriz de varianzas-covarianzas condicional se estimaría como:

$$\mathbf{S}_{\hat{\mathbf{V}}|\mathbf{X}_j} = \mathbf{S}_{VV} - \mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}\mathbf{S}_{VV} \tag{11}$$

Las ecuaciones 10 y 11 definen el procedimiento Bayes empírico para la estimación de las puntuaciones parciales verdaderas (Wainer, Sheehan & Wang, 2000). La estimación de la fiabilidad de las nuevas puntuaciones parciales para una escala concreta podría obtenerse como la razón entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones verdaderas estimadas (Wainer, Vevea, Camacho, Reeve, Rosam Nelson, Swygert & Thissen, 2001). Este valor vendría dado por el cociente entre los elementos de la diagonal de $\mathbf{A} = \mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}\mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}\mathbf{S}_{VV}$ (a_{vv}) y los elementos de la diagonal de $\mathbf{C} = \mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}\mathbf{S}_{VV}$ (c_{vv}). Por lo tanto la fiabilidad de las puntuaciones mejoradas que representamos por medio de r_{vv} podría estimarse como:

$$r_{vv} = \frac{a_{vv}}{c_{vv}} \tag{12}$$

Una vez expuesto el modelo Bayes empírico para el incremento de la fiabilidad de las puntuaciones parciales, el objetivo de este trabajo es mostrar su aplicación a través de un ejemplo práctico. En él se aplicará el procedimiento de mejora de las puntuaciones parciales a las escalas de un test de autoconcepto y se describirán las características de las nuevas puntuaciones a través del efecto que ellas causan en la distribución original, las relaciones entre las escalas parciales y la descripción de los perfiles individuales.

Método

Participantes

La muestra está compuesta por 489 estudiantes con edades comprendidas 12 y 18 años. De ellos 262 son chicos chicos, y 277 chicas. La media aritmética de la edad del grupo de los chicos es 13,9 (SD= 1,65), y la media obtenida en el grupo de las chicas es 14,06 (SD= 1,66).

Instrumento

El cuestionario para la medida del autoconcepto, AFA-A, (Musitu, García y Gutiérrez, 1997), es un test de rendimiento típico compuesto por 31 ítems de respuesta ordenada con 3 categorías de respuesta (Siempre, Algunas Veces, Nunca). El test ofrece una puntuación general en la variable autoconcepto, y puntuaciones parciales en las dimensiones de Autoconcepto Académico o Escolar (n= 10), Autoconcepto Social (n= 5), Autoconcepto Emocional (n= 10) y Autoconcepto Familiar (n= 6). Es un cuestionario de naturaleza tetradimensional cuyo ajuste al modelo teórico fue evaluado sometiendo la matriz de correlaciones a un análisis factorial confirmatorio Los valores obtenidos (χ^2=807,65; g.l.= 428; RMSEA= 0,04; GFI= 0,91) permitieron mantener la hipótesis nula de adecuación entre el modelo y los datos (un análisis en profundidad de la estructura del test puede consultarse en Elosua, 2005).

Resultados

Descriptivos

Los estadísticos descriptivos referidos a cada una de las escalas parciales y a la escala total, así como sus correspondientes coeficientes de fiabilidad estimados por el alpha de Cronbach y los errores estándar de medida estimados sobre las puntuaciones estandarizadas pueden consultarse en la tabla 1. Puesto que las escalas analizadas están compuestas por un número diferente de ítems todas las estimaciones fueron efectuadas sobre las puntuaciones estandarizadas.

Estimación de las puntuaciones verdaderas parciales. La matriz de covarianzas entre puntuaciones parciales observadas estandarizadas y la matriz de covarianzas entre puntuaciones parciales verdaderas pueden consultarse en la tabla 2.

Dados los supuestos de la TCT ambas matrices difieren únicamente en sus valores diagonales; los elementos de la diagonal de la matriz de covarianzas observadas son 1, mientras que los valores que aparecen en la matriz de covarianzas verdaderas son los coeficientes de fiabilidad de cada una de las escalas parciales.

Tabla 1
Estadísticos descriptivos correspondientes a las escalas parciales y a la escala total

Escalas	n. ítems	$\bar{X}$	S_X	Asimetría	α	S_v	r_{VV}	$S_{eobservado}$	$S_{eestimado}$
Escolar	10	21,84	2,81	-0,132	0,628	0,43	0,72	0,60	0,35
Emocional	10	20,62	2,87	-0,162	0,622	0,41	0,71	0,61	0,35
Familiar	6	15,43	1,84	-0,676	0,572	0,36	0,68	0,65	0,33
Social	5	13,27	1,64	-1,083	0,634	0,42	0,71	0,60	0,35
Total	31	71,16	6,28		0,766				

La información contenida en la matriz de varianzas-covarianzas verdaderas (S_{VV}) puede ser utilizada para examinar la dimensionalidad del test. El análisis de sus valores propios indicará el carácter unidimensional/multidimensional del test. En nuestro caso (tabla 3) los valores propios muestran la naturaleza multidimensional de la prueba (escalas parciales), así como la presencia de un factor dominante que justificaría la utilización de una puntuación total.

A partir de estas matrices es posible obtener la matriz de coeficientes regresores, matriz $\hat{\mathbf{B}} = \mathbf{S}_{VV}\mathbf{S}_{XX}^{-1}$ (ecuación 10) con la cual podrán estimarse las puntuaciones parciales verdaderas y los nuevos coeficientes de fiabilidad de las escalas. En la matriz **B** (tabla 4) cada una de las filas aporta los pesos por los que deberían de multiplicarse cada una de las puntuaciones parciales observadas para obtener las puntuaciones parciales estimadas o puntuaciones mejoradas.

Estimación de la fiabilidad de las escalas mejoradas. Siguiendo la ecuación 12 estimamos los nuevos coeficientes de fiabilidad de las escalas mejoradas (tabla 1). Los nuevos coeficientes de fiabilidad estimados para las escalas parciales mejoradas son respectivamente 0,72, 0,71, 0,68 y 0,71. En todos los casos las escalas han sufrido un incremento considerable que las sitúa en torno/sobre el valor de 0,7. El efecto de la estimación Bayes empírica afecta también al error estándar de medida. Las desviaciones estándar de los errores de medida asociadas a las escalas parciales observadas tenían un mínimo de 0,60 que se correspondía con las escalas de autoconcepto académico y social, y un máximo de 0,65 obtenido en la escala Familiar. Estas desviaciones son reducidas en el caso de ser estimadas con las puntuaciones parciales mejoradas. Como puede apreciarse en la tabla 1 el rango de errores estándares de medida estimados tiene un valor mínimo de 0,33 y un máximo de 0,35. El valor mínimo estimado corresponde a la escala de Autoconcepto Familiar, que por otro lado, es la escala que ha sufrido una mayor reducción en su variabilidad tras la estimación Bayes empírica (tabla 1); la varianza asociada a esta escala tras la estimación Bayes empírica fue de 0,36.

Efecto sobre la distribución de las puntuaciones. Las nuevas puntuaciones estimadas son regresiones a la media ponderadas por los coeficientes de fiabilidad de las escalas parciales; por lo tanto su varianza es menor que la varianza de las puntuaciones parciales originales. Dado que estamos trabajando con puntuaciones estandarizadas, las varianzas de las escalas parciales originales es en todo los casos igual a 1; sin embargo, las varianzas de las puntuaciones mejoradas han sufrido una reducción tras la cual se sitúan correlativamente en los valores de 0,43, 0,41, 0,36 y 0,42 si bien es cierto que todas ellas están centradas en torno al mismo valor (tabla 1). Los siguientes gráficos de cajas (figura 1) muestran el efecto de concentración que tienen las puntuaciones estimadas sobre las distribuciones originales estandarizadas. La máxima concentración se ha dado en la escala de Autoconcepto Familiar. El coeficiente de fiabilidad original de esta escala fue el más bajo de todas las escalas analizadas (α= 0,572), por lo tanto la regresión hacia la media grupal ha sido máxima comparada con el resto de las escalas, que son algo más fiables. A medida que disminuye el coeficiente de fiabilidad de la escala, aumenta la incertidumbre o el error de medida, como consecuencia las estimaciones que proporciona el procedimiento Bayes empírico se aproximan a la media grupal.

Efecto sobre las puntuaciones individuales. Para mostrar el efecto que las puntuaciones mejoradas ejercen sobre el perfil individual obtenido por medio de las puntuaciones parciales observadas hemos seleccionado sobre la muestra total a seis sujetos al azar. La tabla 5 muestra las puntuaciones parciales observadas es-

Tabla 2
Matrices de varianzas-covarianzas parciales observadas y verdaderas

	S_{XX} Escolar	Emocional	Familiar	Social
Escolar	1,00			
Emocional	0,256	1,00		
Familiar	0,392	0,199	1,00	
Social	0,257	0,365	0,177	1,00
	S_{VV} Escolar	**Emocional**	**Familiar**	**Social**
Escolar	0,628			
Emocional	0,256	0,622		
Familiar	0,392	0,199	0,572	
Social	0,257	0,365	0,177	0,634

Tabla 3
Valores propios de la matriz S_{VV}

	Valor propio	% total	% varianza
1	1,44	58,60	36,00
2	0,55	22,42	13,75
3	0,26	10,78	6,50
4	0,20	08,18	5,00

Tabla 4
Matriz **B**

	Escolar	Emocional	Familiar	Social
Escolar	0,53	0,06	0,15	0,06
Emocional	0,06	0,54	0,04	0,14
Familiar	0,18	0,04	0,48	0,02
Social	0,06	0,13	0,02	0,56

Tabla 5
Puntuaciones parciales empíricas estandarizadas y puntuaciones parciales estimadas

Sujeto	Escolar	Emocional	Familiar	Social
1	-1,72	0,83	-0,23	1,05
	-0,83	*0,48*	*-0,36*	*0,58*
2	0,77	0,13	0,85	1,05
	0,62	*0,30*	*0,59*	*0,68*
3	0,41	-0,56	0,85	-0,17
	0,31	*-0,27*	*0,46*	*-0,12*
4	-0,65	0,13	-0,23	0,44
	-0,34	*0,08*	*-0,21*	*0,22*
5	-0,3	0,13	-1,32	-0,44
	-0,33	*0,06*	*-0,68*	*-0,21*
6	0,77	0,13	-0,23	-0,77
	0,33	*0,00*	*0,01*	*0,37*

Nota: los valores en cursiva representan las puntuaciones parciales estimadas

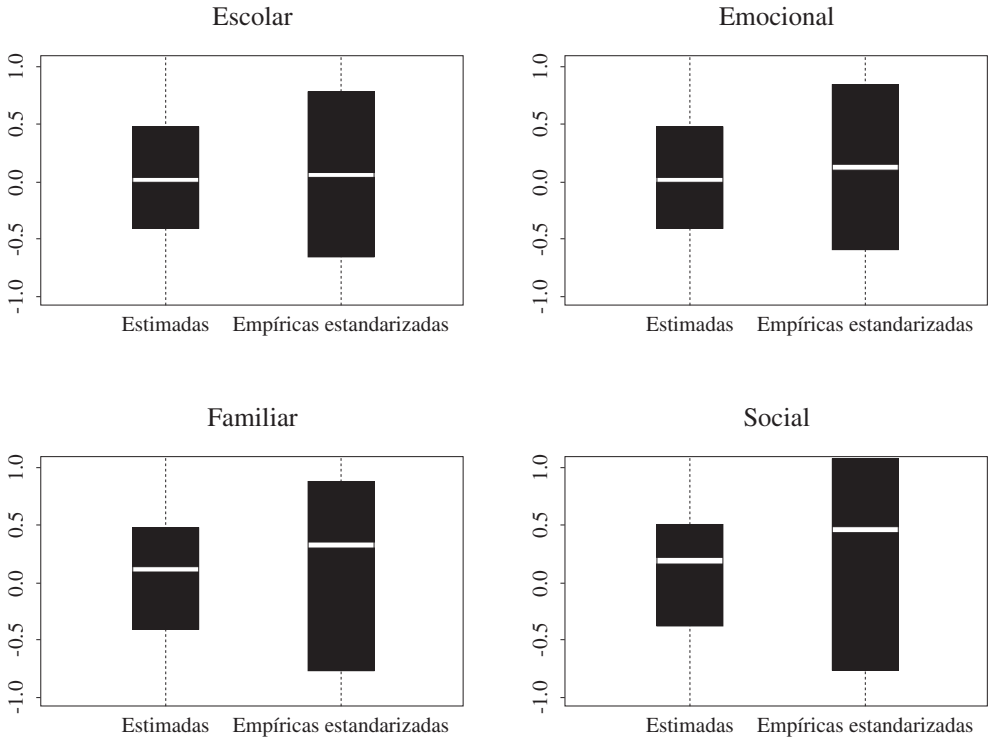

Figura 1. *Distribución de las puntuaciones parciales observadas y estimadas*

tandarizadas y las puntuaciones estimadas para este grupo de estudiantes. Los perfiles individuales correspondientes a ambos tipos de puntuación se han representado en la figura 2, donde se aprecia el efecto de suavizado que ejerce el procedimiento de estimación descrito sobre las puntuaciones empíricas. Como en el caso de la fórmula de Kelley para la predicción de una sola variable, también en el caso multivariado el suavizado es mayor para las escalas con un menor coeficiente de fiabilidad.

Efecto sobre las correlaciones entre escalas parciales. La utilización de puntuaciones parciales originales o estimadas también afecta a la correlación entre escalas. Comparando las correlaciones entre las primeras (tabla 2; dado que se trata de puntuaciones estandarizas las covarianzas y las correlaciones coinciden) y las puntuaciones estimadas (tabla 6) se hace evidente el incremento que han sufrido todas ellas. La correlación previa entre las escalas parciales Escolar-Familiar de 0,392, se ha transformado en una correlación de 0,817. Este efecto es sistemático en el total de correlaciones bivariadas. Las puntuaciones estimadas tienen menos variabilidad que las originales y han sido estimadas utilizando información aportada por el resto de las escalas con lo cual aumenta el grado de interrelación previo existente entre ellas.

Tabla 6
Correlaciones entre las puntuaciones parciales mejoradas

	Escolar	**Emocional**	**Familiar**	**Social**
Escolar	1,00			
Emocional	0,547	1,00		
Familiar	0,817	0,467	1,00	
Social	0,538	0,740	0,422	1,00

Si estimamos los valores propios de la matriz de correlaciones entre puntuaciones observadas y de la matriz de correlaciones entre puntuaciones estimadas puede apreciarse la «pérdida» de multidimensionalidad de los datos analizados, y por tanto la mayor correlación entre escalas parciales. En el primer caso los valores propios estimados fueron 1,82, 0,94, 0,63 y 0,59; en el segundo caso, los valores propios alcanzaron los valores de 2,76, 0,79, 0,26 y 0,16. Estas cantidades indican por un lado un porcentaje de varianza mayor asociado al primer autovalor extraído de la matriz de puntuaciones estimadas (70%) sobre el porcentaje de varianza asociado a las puntuaciones observadas (45%); y por otro una mayor diferencia neta entre los dos primero autovalores en la matriz obtenida a partir de las puntuaciones estimadas (2,76-0,79= 1,97) frente a la diferencia relacionada con las puntuaciones observadas (1,82-0,94= 0,88). Ambos resultados refuerzan las conclusiones sobre la tendencia a la unidimensionalidad de las puntuaciones estimadas.

Discusión y conclusiones

El procedimiento Bayes empírico mostrado permite incrementar la fiabilidad de las puntuaciones parciales por medio de una estimación que hace uso de la información contenida en el resto de las escalas parciales que componen el test. La información aportada por cada una de las escalas parciales dependerá tanto de las relaciones entre ellas como de su coeficiente de su fiabilidad. La estimación ejerce un efecto de suavizado (regresión a la media) que será tanto más abruto cuando la fiabilidad de la escala parcial original sea baja. Este suavizado, incrementa la estabilidad estadística de las puntuaciones, es decir, su fiabilidad.

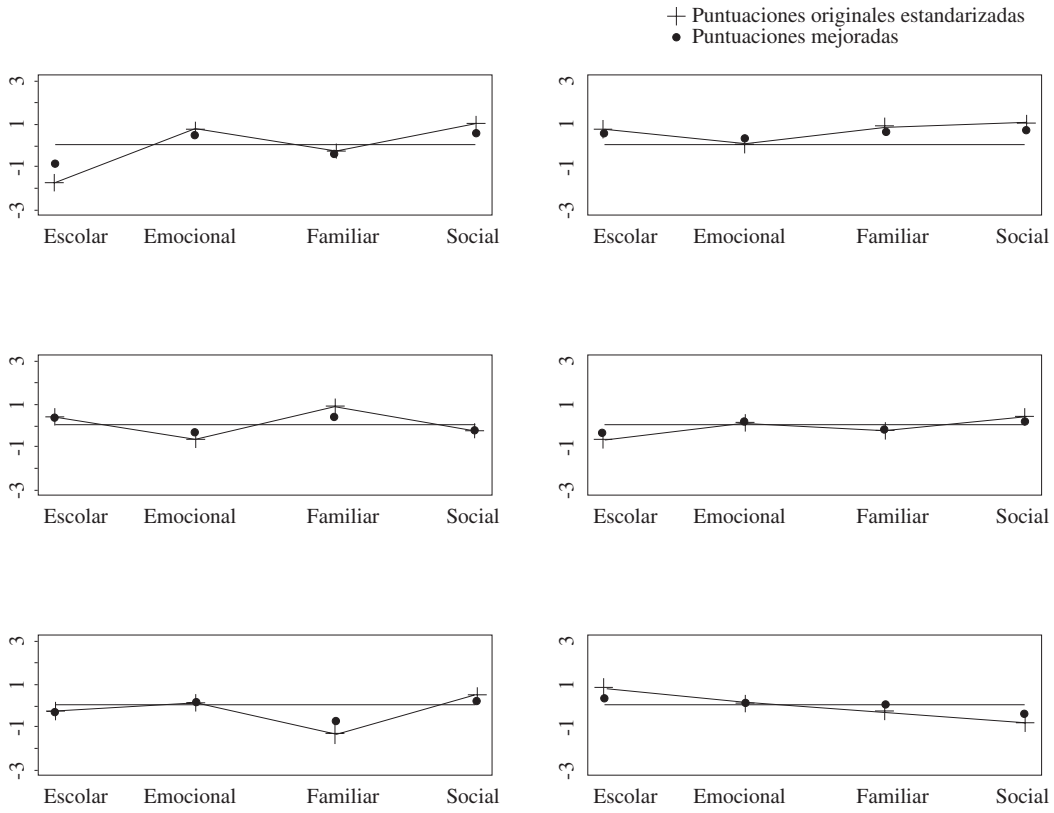

Figura 2. *Perfiles individuales obtenidos con las puntuaciones parciales empíricas estandarizadas y las puntuaciones parciales estandarizadas estimadas*

La valoración del incremento en la fiabilidad de las puntuaciones estimadas podría llevarse a cabo comparándola con el incremento en la fiabilidad obtenible por medio del aumento del número de ítems utilizando para ello la formula profética de Spearman Brown. En nuestro caso, la mejora en la estimación obtenida en cada una de las escalas parciales sería equivalente a incrementar cada una de ellas en un 60,05%, 50,0%, 66,6% y un 60,0%, lo cual se traduciría en un número final de ítems para cada una de las escala de 16, 15, 11 y 7 ítems. El test total contaría con 49 ítems en lugar de 31, es decir un 58,06% más ítems que el test original.

Uno de los problemas asociados con el procedimiento expuesto es que las correlaciones entre las puntuaciones estimadas se incrementan de tal modo que en determinadas circunstancias (p.e. test unidimensional y evaluación de áreas de contenido) los valores pueden aproximarse a uno a la par que la varianza de las puntuaciones disminuye. Esto conllevaría una falta de discriminación entre las distintas escalas parciales y entre los distintos sujetos que responden al test. Esta ausencia de diferenciación no cumpliría los objetivos diagnósticos esperables de este tipo de puntuaciones. Es decir, si una escala puede considerarse esencialmente unidimensional, y por tanto se parte de un alto grado de correlación entre las escalas parciales, los perfiles diagnósticos asociados a la estimación Bayes empírica pueden ser planos y no diferenciar entre escalas parciales. Por ello la estimación Bayes empírica alcanza su máxima eficacia en las situaciones de multidimensionalidad en que las correlaciones entre las escalas parciales son elevadas, el coeficiente de fiabilidad de la escala a mejorar es bajo y los coeficientes de fiabilidad de las escalas parciales en las que se apoya la estimación es elevada (Edwards y Vevea, 2006).

La estimación Bayes empírica afecta a la distribución de las puntuaciones, y a los perfiles individuales. El rango de puntuaciones se reduce, disminuye la varianza y se suavizan los perfiles; por lo tanto también las diferencias. Como contrapartida se incrementa la estabilidad de las puntuaciones. Ante las consecuencias asociadas a la utilización de un tipo de puntuación u otra (original o estimada) la aplicabilidad del método estará sujeta al tipo de información que se quiera obtener. El psicólogo o educador deberán de optar entre centrar su interés en una puntuación asociada a una situación específica (puntuación obtenida en un test concreto) que es menos estable, o en una puntuación con una mayor fiabilidad, la puntuación estimada, que esta definida estadísticamente como la media de la distribución condicional para ese sujeto dada su puntuación observada. La situación sería groseramente comparable a la evaluación de cualquier equipo deportivo o actividad deportiva individual asociada a un sólo evento o al resultado de toda una liga o temporada. En algunas situaciones el centro de interés recaerá sobre el resultado del evento concreto, sin embargo en otras muchas el foco de atención estará definido por el rendimiento medio esperado.

La estimación Bayes empírica está siendo objeto de estudio tanto dentro del marco de la teoría clásica de tests como dentro del marco de la teoría de respuesta al ítem, y se ofrece como una solución estadística y éticamente valida al problema de generar puntuaciones parciales fiables para escalas con un número reducido de ítems y por tanto estadísticamente poco estables.

Agradecimientos

Este trabajo ha sido desarrollado en el marco de un proyecto de investigación subvencionado por el Ministerio de Educación y Ciencia (código es SEJ2005-01694/PSIC) y del Programa de Perfeccionamiento y Movilidad del Personal investigador del Departamento de Educación, Universidades e Investigación del Gobierno Vasco, Orden de 16 de mayo del 2007.

Referencias

Cattell, H.B. (1989). *The 16PF, Personality in depth*. Champain, IL, Institute for personality and ability testing, Inc.

Edwards, M.C., y Vevea, J.L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241-259.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15*(2), 315-321.

Elosua, P. (2005). Evaluación progresiva de la invarianza factorial entre las versiones original y adaptada de una escala de autoconcepto. *Psicothema, 17*(2), 356-362.

Garner, D.M. (2001). *EDI-2. Inventario de trastornos de la conducta alimentaria*. Madrid: TEA

Kelley, T.L. (1927). *The interpretation of educational measurements*. New York: World Book.

Kelley, T.L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Leighton, J.P., y Gierl, M.J. (2007). *Cognitive Diagnostic Assessment for Education*. New York, Cambridge University Press.

Morrison, D.F. (1967). *Multivariate statistical methods*, MacGraw-Hill, New York.

Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid, Pirámide, S.A.

Musitu, G., García,, F., y Gutiérrez, M. (1997). *AFA. Autoconcepto Forma-A.* Madrid: TEA.

Santisteban, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma, S.A.

Shavelson, J., Hubner, J.J., y Stanton, G.C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407-442.

Wainer, H., Sheehan, K., y Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement, 37*(2), 113-140.

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B.B., Rosa, K., Nelson, L., Swygert, K.A., y Thissen, D. (2001). Augmented Scores. «Borrowing Strength» to compute scores based on small number of items. En D. Thissen y H. Wainer (eds.): *Test Scoring* (pp. 343-389). Mahwah, NJ, Lawrence Erlbaum Associates**.**

Ye, F., Stone, C.A., y Lane, S. (2007). Providing subscales scores for diagnostic information. *Annual Meeting of the National Council of Measurement*, Chicago, Il.