Psicothema 2005. Vol. 17, nº 3, pp. 516-521
www.psicothema.com
ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2005 Psicothema

Clases latentes y funcionamiento diferencial del ítem

Paula Elosua y Alicia López Jáuregui
Universidad del País Vasco

La existencia de clases latentes cualitativamente diferentes dentro de una población exige la adaptación de los procedimientos tradicionales de detección del FDI. Este trabajo explora las relaciones entre las clases latentes y el funcionamiento diferencial de los ítems, analizando el FDI en un test de aptitud numérica. Los resultados muestran una disminución en el porcentaje de ítems con funcionamiento diferencial; 28% bajo la concepción tradicional, y 16% cuando se ajusta el modelo de Rasch mixto antes de la detección del FDI. Se concluye que en un estudio de FDI es conveniente estimar previamente las clases latentes y sobre ellas definir los grupos de referencia y focal.

Differential item functioning in heterogeneus populations. Latent classes. The existence of qualitatively different latent classes within a population demand the adaptation of traditional procedures to detect DIF. The aim of this study was to explore the relations between DIF and latent classes. In order to achieve this, a numerical aptitude test was analyzed. The results show a decrease of the number of items with DIF; 28% in the traditional way versus 16% if the mixed Rasch model is used before DIF detection. In DIF analysis studies it would be more appropriate to estimate the latent classes and then to define the reference/focal groups within these classes.

La hipótesis adoptada por la comunidad psicométrica sobre el origen del funcionamiento diferencial del ítem postula la multidimensionalidad de un espacio latente pretendidamente unidimensional, y su diferente distribución entre las poblaciones de interés como posible causa del mismo (Ackerman y Evans, 1994; Elosua, López y Egaña, 2000a, 2000b; Shealy y Stout, 1993). Esta asunción supone la coexistencia de habilidades primarias, o de medición intencionada (θ), con habilidades espurias, o no-intencionadas (η), que generan varianza irrelevante al objeto de medida propuesto (Figura 1a). Como consecuencia, se violaría la condición de invarianza de medida (Lord, 1980), pues diferirían las funciones de respuesta a un ítem estimadas en dos muestras de sujetos (X_R, X_F).

Podríamos traducir las consideraciones anteriores a un lenguaje formal definiendo el funcionamiento diferencial del ítem (X_i) respecto a dos grupos (G_1, G_2), y una variable intencionadamente medida (θ) como,

$$P(X_i = 1|\theta, G_1) \neq P(X_i = 1|\theta, G_2) \qquad (1)$$

donde:

i se refiere al ítem.
X_i es la variable de respuesta.
θ es el objeto intencionado de medida.
G es la variable que indica el grupo de pertenencia.

Fecha recepción: 9-3-04 • Fecha aceptación: 18-10-04
Correspondencia: Paula Elosua
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián (Spain)
E-mail: pspelolp@sc.ehu.es

Según (1) la probabilidad de respuesta al ítem condicionada sobre el nivel de habilidad en la variable medida (sea ésta manifiesta o latente) es diferente en función del grupo de pertenencia.

Esta definición implícitamente asume que dentro de cada uno de los grupos observados (G_1, G_2), que habitualmente se definen en función de factores externos al ítem/test analizado (raza, sexo, nivel educativo, idioma…), la población es homogénea respecto al objeto de medida. Ahora bien, si aceptáramos la hipótesis de heterogeneidad dentro de cada grupo, es decir, si los datos observados reflejaran varias poblaciones latentes o clases (CL1, CL2) no asociadas al criterio externo de grupo, la formulación (1) sería incompleta. Sería necesario incorporar información sobre los grupos latentes en el proceso de definición/detección del funcionamiento diferencial del ítem (ecuación 2) (De Ayala, Kim, Stapleton y Dayton, 2002; Kelderman y Macready, 1990).

$$\begin{gathered} P(X_i = 1|\theta_{CL1}) \neq P(X_i = 1|\theta_{CL2}) \\ P(X_i = 1|\theta_{CL1}, G_1) \neq P(X_i = 1|\theta_{CL1}, G_2) \\ P(X_i = 1|\theta_{CL2}, G_1) \neq P(X_i = 1|\theta_{CL2}, G_2) \end{gathered} \qquad (2)$$

donde:

i se refiere al ítem.
CL1 y CL2 se refieren a la clase latente.
X_i es la variable de respuesta.
θ es el objeto intencionado de medida.
G es la variable que indica el grupo de pertenencia.

En estas situaciones, en las que se presupone multidimensionalidad asociada a la existencia de patrones cognoscitivos o patrones de respuesta diferentes entre los sujetos independientemente de su pertenencia a un grupo sociodemográfico, no sería correcto modelizar las funciones de respuesta con modelos unidimensionales.

Esta práctica violaría sistemáticamente el supuesto de homogeneidad del grupo, lo cual nos llevaría a cuestionar los resultados obtenidos por cualquiera de los procedimientos de detección al uso. Podrían confundirse falta de ajuste y funcionamiento diferencial.

La coexistencia de clases latentes se traduce en una falta de equivalencia de la variable principal dentro de cada grupo, causada por la concurrencia de diferencias cualitativas entre los sujetos. Su origen podría asociarse, entre otras causas, con la utilización de diversas estrategias de resolución o estilos de respuesta. Para describir esta heterogeneidad, es necesario formularla a través de modelos que permitan la coincidencia de clases cualitativamente distintas, dentro de las cuales sea posible cuantificar las diferencias entre los sujetos; éste es el objetivo de los modelos de clase latente, entre los que podríamos citar los modelos de distribución mixtos de Rost (1990), o Mislevy y Verhelst (1990), que son generalizaciones del modelo de Rasch.

Los modelos de clase latente exploran la partición de la población que genere la mayor diferencia entre los parámetros de los ítems, para maximizar así la heterogeneidad entre grupos. Su potencial diagnóstico se presta como instrumento de análisis para indagar en las diferencias cualitativas entre sujetos, y simultáneamente cuantificar sus habilidades respecto a las mismas tareas. Dado que las clases no son conocidas a priori, los modelos son heurísticos en el sentido de que identifican subpoblaciones de individuos, a ser escalados posteriormente por modelos de respuesta al ítem.

En este nuevo marco teórico, la detección del funcionamiento diferencial del ítem exigiría una adaptación del procedimiento habitual, de modo que primero habría que definir las clases latentes (θ_1, θ_2, …, θ_L), y dentro de ellas especificar las clases manifiestas. De este modo, la comparación entre los grupos de referencia y focal se efectuaría dentro de cada clase latente (Figura 1b). El objetivo sería anteponer criterios de validez psicométricos a criterios sociodemográficos.

Sólo cuando las poblaciones fueran homogéneas respecto a θ, coincidirían la perspectiva multidimensional y la de clases latentes. En caso contrario, es decir, si la muestra de calibración contiene clases latentes, y dado que los parámetros a estimar están condicionados por la pertenencia a la clase, sería menester que la estimación de éstos dentro de cada una de las clases precediera a la detección del posible funcionamiento diferencial del ítem.

Dentro de este contexto, el objetivo de este trabajo es ahondar en las posibilidades que ofrece la consideración de clases latentes en el proceso de detección/explicación del funcionamiento diferencial de los ítems. Para ello, estudiamos desde una perspectiva exploratoria las consecuencias que tiene sobre la detección del funcionamiento diferencial del ítem la modelización previa por medio de clases latentes. Los datos analizados pertenecen a una prueba de aptitud numérica administrada a una muestra en la que se han definido los grupos de interés en función de la variable externa, idioma de escolarización.

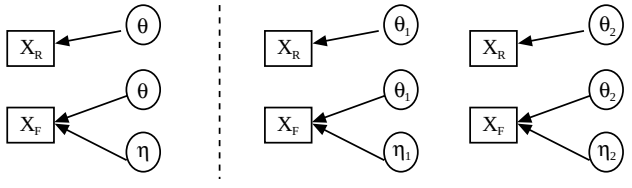

Figura 1a. *FDI y multidimensionalidad; 1b. FDI y clases latentes*

Método

Participantes

La muestra está formada por 1.063 alumnos de Enseñanza Primaria Obligatoria y Enseñanza Secundaria Obligatoria con edades comprendidas entre 9 y 11 años, que cursan estudios en la comunidad autónoma del País Vasco. La selección de los participantes en el estudio ha transcurrido en dos fases. En una primera etapa siguiendo un criterio de doble estratificación en función del tipo centro de enseñanza público/privado/concertado, rural/urbano, se han concretado los centros que podrían participar en el estudio. En una segunda fase, la selección de los centros ha sido circunstancial.

La variable manifiesta que define los grupos de contraste es el idioma de escolarización de los participantes. El grupo de referencia (N_R= 542) está formado por estudiantes que reciben toda su instrucción en castellano (modelo lingüístico A); mientras que el grupo focal se caracteriza por ser escolarizado exclusivamente en euskera (N_F= 521) (modelo lingüístico D).

Instrumentos

La prueba analizada es un test de aptitud numérica que forma parte de la «Batería de aptitudes diferenciales y generales» en su versión E (Yuste, 1988). Es un test compuesto por 25 ítems de elección múltiple, con 6 alternativas de respuesta.

Análisis

Modelo de Rasch Mixto

Entre los modelos de clase latente, en este trabajo utilizamos el modelo de Rasch mixto (Rost, 1990), en el que la probabilidad de respuesta correcta a un ítem (*i*) está condicionada por el nivel en el rasgo, y por la pertenencia a una determinada clase (l).

$$p_{ijl} = \frac{e^{\theta_{jl}+b_{il}}}{1+e^{\theta_{jl}+b_{il}}} \tag{3}$$

donde:

p_{ijl} es la probabilidad de respuesta correcta al ítem i por la persona j en la clase l.
θ_{jl} es el nivel de habilidad de la persona j en la clase latente l.
b_{il} es el parámetro de dificultad del ítem i en la clase latente l.

Al igual que en el modelo de Rasch, para cada una de las clases estimadas se cumple:

$$\Sigma_l b_{il} = 0 \tag{4}$$

Además, asumiendo que las clases son mutuamente excluyentes entre sí,

$$p_{ij} = P\left(X_i = 1 \middle| \theta_j\right) = \sum_l \pi_l P_{jil} = \sum_l \pi_l \frac{e^{\theta_{jl}+b_{il}}}{1+e^{\theta_{jl}+b_{il}}} \tag{5}$$

donde p_{ij} es la probabilidad de respuesta incondicional a la clase. π_l es el parámetro de la clase, que verifica $0<\pi_l<1$ y $\sum_l \pi_l = 1$.

La evaluación del ajuste se basa en índices relativos que comparan modelos en competencia; es decir, modelos que sucesivamente incorporan parámetros de clase. Entre los índices de bondad de ajuste más conocidos se encuentran AIC, BIC o CAIC (Akaike, 1973; Bozdogan, 1987; Read y Cressie, 1988):

$$AIC = -21n(L) + 2k$$
$$BIC = -21n(L) + k\ln(N) \qquad (6)$$
$$CAIC = -21n(L) + k(1+1n(N))$$

donde:

L es el máximo de la función de verosimilitud de los datos.
k es el número de parámetros del modelo.
N es el número total de observaciones.

Resultados

Descriptivos

Las medias aritméticas para los grupos de referencia y focal son 17,22 y 15,05, con desviaciones estándar de 4,30 y 4,40, respectivamente. La diferencia de medias es significativamente mayor para el grupo de referencia que para el focal (t= -8,12; p<0,001). La consistencia interna de la prueba, evaluada con el alpha de Cronbach, es para cada uno de los grupos, 0,80 y 0,79. La diferencia entre estos valores no puede considerarse significativa ($F_{520,541}$= 1,05; p= 0,56).

Unidimensionalidad y estimación de parámetros

Las dos muestras superan la condición de unidimensionalidad esencial definida por Stout (1987) (t= 0,73; p= 0,23; t= 0,91; p= 0,17). La estimación de los parámetros bajo el modelo de Rasch se efectúa de modo independiente en cada una de las muestras (Tabla 1) con el software WinMira (von Davier, 2001). El ajuste de los ítems (índice Q) se evalúa sobre el logaritmo de la verosimilitud de los patrones de respuesta al ítem observados (Rost y von Davier, 1994). El índice Q varía entre 0 (ajuste perfecto) y 1. Los análisis efectuados reflejan problemas para cuatro de los ítems en cada una de las muestras. En la muestra de referencia son los ítems 1 (Q= 0,3547), ítem 3 (Q= 0,2817), ítem 15 (Q= 0,11) e ítem 21 (Q= 0,1187). En la muestra focal son los ítems 1, 3, 6 y 15 los que presentan problemas de ajuste (Q_1= 0,2887; Q_3= 0,2568; Q_6= 0,2785; Q_{15}= 0,1147).

Funcionamiento diferencial del ítem

El funcionamiento diferencial del ítem se evalúa con el estadístico Mantel-Haenszel (Dorans y Holland, 1993) aplicando un procedimiento de purificación del criterio en dos etapas, implementado por las autoras en S-Plus. La tabla 2 muestra los valores de MH-Delta (MH-DELTA= -2,35ln (α_{MH})), estadístico utilizado por el *Educational Testing Service* (ETS), que cuantifica la diferencia de ejecución entre los grupos de referencia y focal. De los 25 ítems analizados, el 28% (negrilla) carece de equivalencia en ambas muestras.

Clases latentes

El estudio del ajuste de los datos a diferentes modelos que definen 1, 2 y 3 clases latentes se apoya en los índices de bondad de ajuste estimados por WinMira (Von Davier, 2001). En la tabla 3 puede apreciarse el descenso en los valores de estos estadísticos en la solución de dos clases respecto a la de una sola clase (modelo de Rasch). Aunque el índice AIC muestra un ligero decremento en la situación de tres clases, tanto BIC como CAIC apuntan a la solución de dos clases como la más plausible.

La mejora en el ajuste de la solución de dos clases respecto al modelo de Rasch puede también concluirse a partir del índice de ajuste al ítem Q. Sólo un ítem en cada una de las clases ofrece valores problemáticos. El ítem 15 en la Clase 1 (Q_{15}= 0,1292) y el ítem 24 en la Clase 2 (Q_{24}= 0,4069). La tabla 1 muestra los valores de los parámetros de dificultad estimados para cada clase.

Tabla 1
Parámetros de dificultad estimados (*p<0,05; **p<0,01)

	Focal	Referencia	Clase 1	Clase 2
1	-1,40*	-1,34**	-0,94	-2,63
2	-1,67	-1,90	-1,50	-2,62
3	-0,77*	-0,65**	-0,18	-2,26
4	-0,97	-0,79	-0,42	-2,25
5	-0,31	-0,66	-0,02	-1,72
6	-0,65**	-0,25	0,04	-1,89
7	-2,70	-2,70	-2,37	-3,79
8	-0,30	-0,29	0,03	-1,28
9	-2,19	-1,69	-1,69	-2,83
10	-1,67	-2,07	-1,58	-2,66
11	1,11	-0,27	0,52	0,04
12	-1,41	-1,23	-1,38	-1,68
13	-0,50	0,12	-0,02	-0,84
14	0,13	0,52	0,50	-0,33
15	-0,39**	-0,21*	-0,48*	-0,38
16	-0,31	-0,18	-0,38	-0,37
17	1,54	1,16	1,36	1,05
18	0,58	0,94	0,64	0,83
19	2,01	1,41	1,41	2,28
20	0,25	0,47	-0,57	1,86
21	1,88	1,81*	1,39	3,41
22	1,80	1,34	0,96	3,74
23	2,00	2,11	1,61	3,91
24	1,53	1,67	0,99	4,12*
25	2,45	2,64	2,07	6,32

Tabla 2
Resultados de la detección del FDI

Ítem	MH-Delta	Ítem	MH-Delta
1	-0,38	14	0,71
2	-0,86	15	0,78
3	-0,12	16	0,28
4	0,15	17	**-1,00**
5	**-1,06**	18	**1,02**
6	0,50	19	**-1,47**
7	0,04	20	0,57
8	-0,11	21	-0,04
9	1,53	22	**-1,20**
10	-1,12	23	0,21
11	**-3,47**	24	0,24
12	0,56	25	0,38
13	**1,29**		

Tabla 3
Estudio del ajuste de diferente número de clases

	1 clase	2 clases	3 clases
Ln(L)	-12973,31	-11448,32	-12603,25
K	26	53	80
AIC	25998,82	25437,83	25366,51
BIC	26128,01	25701,18	25764,01
CAIC	26154,01	25754,18	25844,01

En el modelo de dos clases los parámetros de pertenencia a cada una de las clases son h_1= 66,3 y h_2= 33,7. En el diagrama de cajas y bigotes (Figura 2) pueden observarse las distribuciones de las clases latentes estimadas, y dentro de éstas las distribuciones de los grupos manifiestos. La clase 1, constituida por 705 estudiantes, está integrada por un 54,6% de alumnos pertenecientes al grupo de referencia, y por un 45,4% de alumnos provenientes del grupo focal. Esta distribución de proporciones se convierte en un 43,8% y 56,2%, respectivamente, para la segunda clase latente.

Las diferencias cuantitivas existentes dentro y entre clases se resumen en la tabla 4.

Las medias aritméticas globales para cada una de las clases son 17,33 para la más numerosa, y 13,85 para la segunda. Sus desviaciones estándar respectivas son 4,42 y 3,63. La diferencia de medias entre las dos clases es significativa ((t= 13,67; p<0,0001); al igual que lo es para el resto de comparaciones posibles, con un nivel nominal prefijado de 0,01. La ejecución media de los grupos de referencia y focal pertenecientes a la clase 1 (18,32 y 16,12) es superior a los niveles medios de las muestras (17,22 y 15,05); por el contrario, la clase 2 agrupa a miembros de los grupos de referencia y focal representados por medias aritméticas inferiores (14,5 y 13,3). La clase 1 agrupa a sujetos de ejecución media superior, mientras que los miembros de la clase 2 presentan un rendimiento menor en el test analizado.

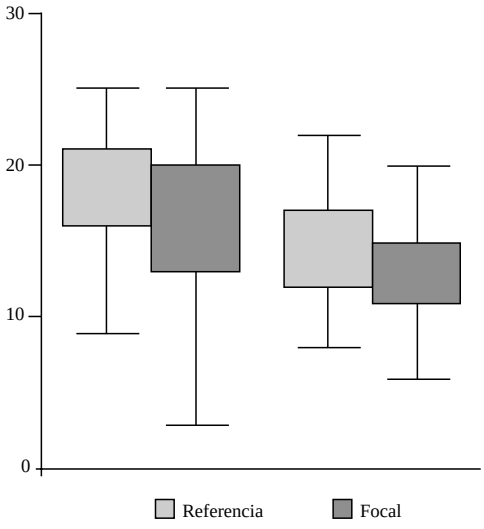

Figura 2. *Descripción de los grupos manifiestos dentro de las clases latentes*

Tabla 4
Descripción de las clases latentes

Clase			Referencia	Focal	
1	N	705	385	320	
	$\bar{X}$	17,3	18,32	16,12	t= 6,7
	σ_X	4,4	4,02	4,58	
2	N	358	157	201	
	$\bar{X}$	13,8	14,5	13,3	t= 3,05
	σ_X	3,6	3,7	3,48	
			t=10,23	t=7,8	

La figura 3 muestra el perfil de los parámetros de dificultad estimados (Tabla 1) para cada una de las clases latentes. Puede observarse que los parámetros correspondientes a la clase 1 son más homogéneos que los de la clase 2. Las desviaciones estándar de las estimaciones de la dificultad de los ítems son, respectivamente, 1,18 y 2,60. Ambos patrones son similares excepto para los últimos cinco ítems de la clase 2, que muestran parámetros excesivamente altos; superiores a 3,41. Estos valores se corresponden con índices de dificultad inferiores a 0,06.

El funcionamiento diferencial de los ítems dentro de cada una de las clases se evalúa con el estadístico Mantel-Haenszel (Tabla 5). Como en el caso anterior, se muestran los valores de MH-Delta, resaltando en negrilla aquellos ítems para los que se rechaza la hipótesis nula con un alpha de 0,01.

Dentro de la primera clase son cuatro los ítems que muestran funcionamiento diferencial; esto supone un 16% del total. En la clase 2 sólo se ha detectado un ítem problemático, que coincide con una de las detecciones de la clase 1.

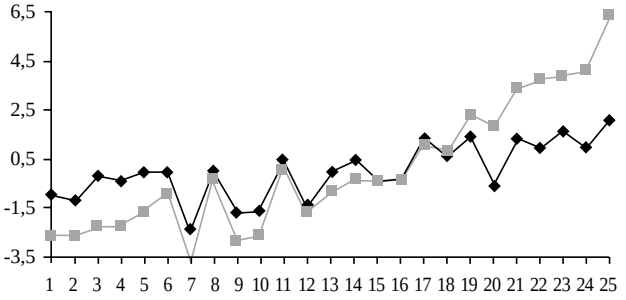

Figura 3. *Perfil de los parámetros de dificultad dentro de cada clase*

Tabla 5
Funcionamiento diferencial del ítem dentro de las clases latentes

	Clase 1	Clase 2		Clase 1	Clase 2
Item	MH-Delta		Item	MH-Delta	
1	-0,17	-1,13	14	0,75	0,68
2	-0,87	-1,07	15	0,40	1,32
3	0,08	-1,13	16	0,63	-0,31
4	-0,03	0,36	17	-1,11	-0,93
5	-1,29	-1,26	18	1,06	1,04
6	0,25	1,23	19	**-1,98**	0,26
7	-0,44	1,18	20	0,94	1,45
8	-0,47	0,32	21	0,11	-1,35
9	**2,38**	-0,92	22	-1,13	-6,04
10	-0,80	-2,19	23	0,55	-1,26
11	**-3,45**	**-3,54**	24	0,62	-2,66
12	0,32	1,01	25	0,49	–
13	**1,88**	0,22			

Entre la detección tradicional y la nueva conceptualización en clases latentes existen diferencias tanto en la cantidad como en la categorización de los ítems clasificados con funcionamiento diferencial. Si bien en el primer caso el 28% de los ítems presentaba problemas, en el segundo ese porcentaje se ha reducido al 16%, siendo el nivel de concordancia entre detecciones del 50%. Como la cuantificación del funcionamiento diferencial a través del estadístico MH-delta permite su valoración numérica, podemos apreciar cómo los ítems que presentan los mayores índices en el procedimiento tradicional (ítems 11, 19 y 13) son también detectados tras la consideración de las clases. Además de estos ítems, la división en clases origina la aparición de un nuevo ítem con funcionamiento diferencial únicamente en la clase 1 (ítem 9) que no fue detectado por el procedimiento tradicional. Por lo demás, y como era esperable, no se produce ninguna alteración en el signo del funcionamiento diferencial asociado a cada uno de lo ítems significativos.

Discusión y conclusiones

Si bien la primera hipótesis relativa a las clases latentes las podría relacionar con la existencia de grupos masters/no masters, los resultados conseguidos en este trabajo nos indican la mayor plausibilidad de otra hipótesis alternativa. La interpretación de las clases obtenidas podría guiarse atendiendo a patrones de comportamiento ante los ítems asociados con un factor de cansancio o falta de tiempo en la ejecución del test completo; esta hipótesis podría corroborarse incidiendo, por un lado, en los valores extremos obtenidos por el estimador del parámetro de dificultad de los cinco últimos ítems (3,41; 3,74; 3,91; 4,12; 6,32), que se corresponden con porcentajes de respuestas correctas extremadamente bajos (0,06; 0,03; 0,04; 0,02; 0,00), y, por otro lado, observando el patrón de respuesta completo. Los parámetros de dificultad para la clase dos son sistemáticamente menores que los obtenidos en la clase uno; lo cual es indicio de que no es una menor «aptitud numérica» media la causante del patrón observado. Si fuera así, quedaría reflejado en el perfil de los parámetros de dificultad, que sería superior al de la clase uno, en todos los ítems del test.

Sobre el primer objetivo del trabajo, el estudio de la relación entre el FDI y las clases latentes, podríamos concluir que en las situaciones en que los modelos mixtos incrementan nuestro conocimiento acerca de las observaciones, su utilidad puede ser doble, por un lado formal y por otro sustantiva. La incorporación de parámetros en un modelo permite siempre una mejor recuperación de los datos, que en este trabajo empírico se ha traducido en una reducción de ítems con funcionamiento diferencial. Desde una perspectiva sustantiva de búsqueda de las causas del FDI, la exploración de clases latentes puede acrecentar nuestras fuentes de información sobre sus causas. En este sentido, la pertenencia a un grupo o clase definida en función de variables cognitivas (u otro tipo de variables) podría eliminar diferencias potencialmente importantes respecto a la variable medida, permitiendo el análisis del FDI en relación con nuevas variables latentes independientes de las variables manifiestas.

Sin embargo, desde un punto de vista aplicado, nos gustaría precisar que la consideración de clases, y la definición dentro de éstas de los grupos de referencia y focal, requiere un incremento en el número de sujetos necesarios para la utilización efectiva de cualquier procedimiento de detección del FDI. En nuestro caso, si bien las muestras de referencia y focal inicialmente eran moderadas (N_R= 542, N_F= 521), tras la agrupación de los sujetos en clases latentes se han reducido considerablemente (N_{RL1}= 385; N_{FL1}= 320; N_{RL2}= 157; N_{FL2}= 201). Es de todos sabido la influencia que tiene el número de sujetos que integran cada uno de los niveles de habilidad sobre los resultados obtenidos.

En definitiva, creemos que el análisis de clases latentes basado en la consideración de diferencias cualitativas entre los sujetos respecto al rasgo medido es una herramienta útil en el análisis de la estructura interna de los datos, que puede aportar información relevante tanto para el estudio de la validez de los tests dentro del análisis de las fuentes de evidencia internas (Elosua, 2003), como para la definición de los grupos normativos sobre los que se interpretarán las puntuaciones empíricas. Es una herramienta de análisis dual porque integra tanto aspectos formales como sustantivos que incide en la doble vertiente de detección/comprensión del funcionamiento diferencial del ítem.

Agradecimientos

Trabajo cofinanciado por la Universidad del País Vasco. UPV1/UPV00109.231-H-14855/2002 y Ministerio de Ciencia y Tecnología MCYT BSO2002-00490.

Referencias

Ackerman, T.A. y Evans, J.A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement 18*(4), 329-342.

De Ayala, R.J., Kim, S., Stapleton, L.M. y Dayton,C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3), 243-276.

Bozdogan, H. (1987). Model Selection and Akaike's Information Criterion (AIC): the General Theory and its Analytical Extensions. *Psychometrika*, *52*(3), 345-370.

Davier, von, M. (2001). *Winmira 1.35. A program systen for analyses with the Rasch model, with the latent class analysis and with the mixed Rasch model*. Kiel: Institute for Science Education (IPN).

Dorans, N.J. y Holland, P.W.(1993). DIF detection and description: Mantel-Haenszel and standarization. En H.Wainer y P.W. Holland (Eds.): *Differential item functioning* (pp. 35-66) Hillsdale NJ: Erlbaum.

Elosua, P., López, A. y Egaña, J. (2000a). Idioma de aplicación y rendimiento en una prueba de comprensión verbal. *Psicothema, 12*(2), 201-206.

Elosua, P., López, A. y Egaña, J. (2000b). Fuentes potenciales de sesgo en una prueba de aptitud numérica. *Psicothema, 12*(3), 376-382.

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

Kelderman, H. y Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifiest and latent examinee groups. *Journal of Educational Meaurement, 27*(4), 307-327.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale,NJ: Erlbaum.

Mislevy, R.J. y Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-216.

Read, N.A.C. y Cressie, T.R.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer Series in Statistics. New York: Springer.

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. y von Davier, M. (1994). A conditional item fix index for Rasch models. *Applied Psychological Measurement, 18*(2), 171-182.

Shealy, R. y Stout, W.F. (1993). A model-based standarization approach that separates true bias/DIF from group differences and detecs test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stout, W.(1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Yuste, C.(1988). *BADYG-E*. Madrid: Ciencias de la educación preescolar y especial.