Psicothema, 1997. Vol. 9, nº 2, pp. 323-332
ISSN 0214 - 9915 CODEN PSOTEG

EVALUACIÓN DEL RENDIMIENTO EN CIENCIAS DE LOS NIÑOS Y NIÑAS DE 13 AÑOS DE LAS DISTINTAS COMUNIDADES AUTÓNOMAS: IMPACTO O SESGO

Mª Isabel Barbero García* y Pedro Prieto Marañón**
* UNED y ** Universidad de La Laguna (Tenerife)

Este trabajo tiene como objetivo principal estudiar la posibilidad de utilizar los medios del National Assessment of Educational Progress (NAEP) para llevar a cabo evaluaciones del rendimiento en nuestro país. Se ha utilizado una muestra de niños y niñas de 13 años, pertenecientes a las distintas Comunidades Autónomas y se ha tratado de analizar si las diferencias encontradas, en cuanto al rendimiento, se deben a diferencias reales o a un funcionamiento diferencial de los ítems de la prueba en las distintas Comunidades.

Assessment of performance in science of 13 year old children in autonomic spanish communities: impact or bias. In the present work the main aim was to study the posibilities of using the tests of the National Assessment of Educational Progress (NAEP) in Spain in order to carry out an assessment of performance in our country. A sample of 13 year old spanish children, belonging to diffferent spanish Autonomic Communities were chosen. We carried out a statistical analysis in order to establish if performance differences found are due to real differences or to differential item functioning in the different Communities studied.

Cuando en 1983 se designa al Educational Testing Service (ETS) como responsable y coordinador del National Asessment of Educational Progress (NAEP), surge la idea de extender estos estudios a nivel internacional. Así, después de conservaciones con representantes de distintos países se propuso el llevar a cabo un primer estudio internacional «International Assessment of Educational Progress» (IAEP) con dos objetivos fundamentales:

Correspondencia: Mª Isabel Barbero García
Facultad de Psicología
Universidad Nacional de Educación a Distancia (UNED)
Apdo. 50.487
28080 Madrid (Spain)

- Estudiar la posibilidad de reducir en tiempo y dinero los requerimientos de los estudios comparativos a nivel internacional aprovechando los materiales y procedimientos utilizados en el NAEP.
- Permitir a los países interesados investigar, utilizando los medios del NAEP, para determinar hasta qué punto estos medios se adecuaban y podían servir para llevar a cabo proyectos de evaluación local.

En este segundo marco se ha llevado a cabo una investigación que tiene como objetivo estudiar, utilizando los medios del

NAEP, si las diferencias en el rendimiento en Ciencias de los niños y niñas de 13 años de las distintas Comunidades Autónomas españolas se deben a diferencias reales (impacto) o, por el contrario, a un funcionamiento diferencial de los ítems de la prueba debido a factores ajenos al que, realmente, están midiendo (sesgo).

Procedimiento

Selección de la muestra

La muestra utilizada en este trabajo fue la muestra española que participó en el primer estudio internacional (IAEP).

Los distintos países participantes en este primer estudio fueron los responsables del diseño muestral, de la selección de la muestra y de la recogida de los datos. En el caso de la muestra española el responsable fue el Servicio de Evaluación del Centro de Investigación, Documentación y Evaluación (CIDE) del Ministerio de Educación y Ciencia (MEC).

Las únicas condiciones impuestas por el ETS en relación con el procedimiento de muestreo fueron el número total de alumnos y de Centros (2.000 y 100 respectivamente) y la organización de la muestra en pares de Centros con características similares, puesto que el estimador utilizado para la estimación del error típico de los estadísticos fue el de Jacknife que utiliza el método de las pseudo-iteraciones son semimuestras.

Teniendo en cuenta que España está organizada en 17 Comunidades Autónomas de las cuales seis tenían transferidas, en el momento del estudio, todas las competencias en materia educativa, el primer paso en el diseño de muestreo fue definir siete estratos, uno por cada una de las Comunidades Autónomas con las competencias transferidas (Andalucía, Canarias, Cataluña, Galicia, País Vasco y Valencia) y un séptimo estrato en el que se incluirían todos aquellos Centros escolares controlados por el Ministerio de Educación y Ciencia del Gobierno Central (Territorio M.E.C.).

Una vez establecida la variable de estratificación, se utilizó un procedimiento de muestreo bifásico en el que la unidad muestral fue, en la primera fase, el Centro y, en la segunda, el alumno. Para determinar el número de Centros y de estudiantes que deberían ser elegidos en cada una de las Comunidades se utilizó el número de estudiantes matriculados en 8º de E.G.B. en cada una de ellas, ya que este curso aglutina a la mayoría de los alumnos de 13 años. El número total de Centros y alumnos que participaron en el estudio fue de 100 y de 2.000 respectivamente. Sin embargo, en el trabajo sólo se utilizaron los datos de 1.756 alumnos, aquellos que cumplimentaron las pruebas en castellano ya que, aunque se les ofrecía la posibilidad de responder en este idioma o en cualquiera de los correspondientes a sus Comunidades Autónomas: catalán, valenciano, gallego o euskera, dado que se trataba de hacer un estudio comparativo se consideró que era necesario unificar las respuestas.

Para una descripción más detallada de la forma en que se realizó el muestreo se puede consultar Lapointe, Mead y Phillips (1989) y los resultados del estudio en King, Bertrand y Dupuis con la colaboración de Létourneau y Dufour (1989).

Prueba utilizada

Los 60 ítems de elección múltiple que componían la prueba de Ciencias fueron seleccionados de un banco de 188 ítems utilizados en 1986 en el programa de evaluación del sistema educativo americano NAEP, y estaban repartidos en las siguientes categorías en función de sus contenidos: 16 ítems de Ciencias Naturales, 10 de Física, 9 de Química, 9 de Ciencias de la Tierra y del Espacio y 16 de Fundamentos de la Ciencia.

En Ciencias Naturales se incluían aquellos ítems cuyo contenido hacía referencia fundamentalmente a los animales y plantas, características de las distintas especies y la fotosíntesis. Los ítems incluidos en la categoría de Física cubrían los conceptos de fuerza, distancia, peso, volumen, aceleración, cuestiones elementales de óptica y de electricidad. Bajo la denominación de Química se incluían aquellos ítems cuyo contenido hacía referencia a los estados de la materia, reacciones que se producen, naturaleza de las soluciones y nociones muy básicas acerca de las características del átomo. La etiqueta de Ciencias de la Tierra y del Espacio se reservaba para incluir a aquellos ítems relacionados con la historia de la tierra, la atmósfera terrestre, los aspectos físicos de la superficie de la tierra y algunas cuestiones acerca del sistema solar. Finalmente, bajo la categoría de Fundamentos de la Ciencia se agrupan aquellos ítems que hacen referencia a cuestiones de lógica, comprobación de hipótesis, diseño de experimentos utilizando instrumentos científicos e interpretación de resultados.

Aunque en el estudio internacional se eliminaron seis ítems debido a su funcionamiento diferencial en los distintos países, dado que nuestro interés está centrado en la evaluación del rendimiento en las distintas Comunidades Autónomas, hemos utilizado la prueba completa.

La aplicación fue llevada a cabo bien por el personal de los Centros a los que se había instruido en la forma de llevarla a cabo, o bien por el personal ajeno al Centro pero especializado en este tipo de aplicaciones.

La duración de la prueba fue de 45 minutos.

Análisis realizados

En primer lugar se realizaron distintos análisis con el fin de averiguar si había diferencias significativas entre las medias obtenidas en la prueba de Ciencias en las distintas Comunidades Autónomas.

En todos los análisis comparativos se tomó como grupo focal o de comparación el formado por la muestra de niños y niñas pertenecientes al territorio MEC ya que, además de ser la más numerosa y mostrar una mayor variabilidad en cuanto a las respuestas dadas al test, era la más variable en cuanto a las zonas territoriales que representaba (Madrid, Santander, Avila, Palma de Mallorca, Salamanca, Huesca, Zaragoza, Gijón, etc.). Los análisis se efectuaron tanto a nivel de la prueba total como en cada una de las subáreas de la misma, y tanto sobre la muestra general como sobre cada una de las Comunidades Autónomas.

Ante los resultados obtenidos, y antes de concluir que realmente hay diferencias significativas en cuanto al rendimiento en Ciencias de los niños y niñas de 13 años de las distintas Comunidades se creyó conveniente llevar a cabo un estudio acerca de la posible existencia de algún tipo de sesgo en la prueba.

Aunque el hecho de que un ítem esté sesgado implica necesariamente una ejecución diferencial entre grupos, no se puede concluir que siempre que haya una ejecución distinta en un ítem sea porque exista sesgo. Como señala Ackerman (1992) hay que distinguir entre los términos sesgo e impacto. Para Ackerman el término impacto se refiere a una diferencia entre grupos en el desempeño en un ítem causada por una diferencia real en la variable medida. Por el contrario, el sesgo hace referencia a diferencias en el desempeño en un ítem causada por factores ajenos a la variable medida. Asimismo conviene distinguir entre los términos sesgo y funcionamiento diferencial de los ítems (DIF), aunque muchas veces se utilicen como equivalentes. Angoff (1982) considera que la razón para diferenciar ambos términos es que mientras que el segundo se infiere directamente a partir de los

procedimientos estadísticos, para afirmar que un ítem está sesgado contra un determinado grupo es necesario hacer alusión a otro tipo de razones de tipo educativo o psicológicas.

Dado que en este trabajo no estamos teniendo en cuenta ni razones educativas ni psicológicas (que es lógico que las haya) sino simplemente los resultados obtenidos a través de distintos procedimientos estadísticos, de ahora en adelante utilizaremos el término DIF para referirnos a las discrepancias encontradas entre dos grupos respecto a las propiedades psicométricas de un ítem en relación con las del resto de los ítems.

En los últimos años han proliferado los métodos para la detección de DIF y se han hecho muy buenas revisiones sobre la eficacia de cada uno de ellos a través de estudios comparativos (Ironson y Subkoviak, 1975; Shepard, Camilli y Averill, 1981; Shepard, Camilli y Williams, 1985; Thissen, Steinberg y Wainer, 1988; Cohen y Seock-Ho Kim, 1993; Navas, 1993, 1994; Barbero y Prieto, 1995; Prieto y Barbero, 1996).

Cuando un investigador se enfrenta al problema de elegir entre todos estos métodos, es necesario que valore las ventajas de cada uno de ellos, pero también sus limitaciones.

Sin duda la utilización de las técnicas derivadas de la TRI es, hoy día, el método preferido ya que es el que mejor detecta cuándo las diferencias encontradas son debidas a diferencias reales en el rendimiento de los sujetos y cuándo se deben a la presencia de DIF, pero tienen el inconveniente de que su utilización requiere grandes tamaños muestrales (sobre todo si el modelo utilizado es el de 3 parámetros) pues en caso contrario las estimaciones pueden ser inestables.

Otro de los procedimientos preferidos es el de Mantel-Haenszel propuesto por Holland y Thayer (1968), sobre todo con la modificación introducida por Mazor, Clauser y Hambleton (1994) que mejoran su capacidad para detectar ítems con DIF no uniforme.

Teniendo esto en cuenta, en este trabajo se han seguido tres aproximaciones a la medida del DIF: una aproximación factorial, el procedimiento Mantel-Haenszel y una aproximación basada en la TRI.

La aproximación factorial, como señalan Adams y Rowe (1988) puede servir como una primera aproximación al estudio del DIF, ya que permite evaluar hasta qué punto los elementos que componen la prueba miden la misma dimensión. Tiene el inconveniente de que incluso con tests que se sabe de antemano que contienen ítems sesgados, puede producir la misma estructura subyacente en los grupos. El procedimiento requiere llevar a cabo un análisis factorial en cada uno de los grupos, en nuestro caso las distintas Comunidades Autónomas, y comparar las soluciones factoriales encontradas. El método de extracción de factores utilizado en nuestro trabajo ha sido el de factores principales sin ningún tipo de rotación y la comparación de las estructuras factoriales se hizo tomando como grupo de comparación la muestra del territorio MEC y calculando el índice de congruencia de Burt y Tucker.

El procedimiento Mantel-Haenszel suele utilizar como variable criterio para hacer las comparaciones del rendimiento entre los dos grupos la puntuación total del test. De esta manera, la puntuación del test se dividió en tres niveles utilizando como criterio para la categorización la media obtenida en la muestra del MEC más y menos una desviación típica.

Finalmente, y teniendo en cuenta que la utilización de las técnicas derivadas de la TRI exigen como paso previo la comprobación del ajuste de los datos a alguno de los modelos, se siguieron los siguientes pasos:

1.– Evaluación del supuesto de unidimensionalidad mediante el test del autovalor.

2.– Análisis de los residuales obtenidos al ajustar los datos a los modelos logísticos de 1, 2 y 3 parámetros.
3.– Una vez seleccionado el modelo al que se ajustaban los datos, aplicación del método de Linn y Harnisch (1981) para el estudio del DIF mediante el programa GENESTE.

El método de Linn y Harnisch está basado en los residuales estandarizados y consiste en lo siguiente: una vez evaluado el ajuste de los datos del grupo que va a servir de grupo focal o de comparación, en nuestro caso la muestra del territorio MEC, a alguno de los modelos, se calcula la curva característica de cada uno de los ítems en este grupo. A continuación se calculan los residuos estandarizados de cada uno de los grupos de referencia (las distintas Comunidades Autónomas) con respecto a la curva característica del grupo de comparación o grupo focal (MEC).

Resultados

En la tabla 1 se recogen los resultados de los primeros análisis descriptivos que se llevaron a cabo. Estos resultados muestran que las subáreas en las que el porcentaje de aciertos ha sido mayor fueron las de Ciencias Naturales y Ciencias de la Tierra y del Espacio, destacando, fundamentalmente, las Comunidades Catalana y Gallega sobre el resto de las Comunidades.

Tabla 1
Porcentaje medio de aciertos en cada área
C.NA= Ciencias Naturales; FISI= Física; QUIM= Química; CTYE= Ciencias de la Tierra y del Espacio; F.CIE= Fundamentos de la Ciencia

Com.	C.NA	FISI.	QUIM.	CTYE	F. CIE
And.	69,6	56,8	56,9	66,1	56,4
Can.	66,9	59,1	55,2	65,5	53,6
Cat.	71,1	61,2	60,0	74,1	58,7
Gal.	70,6	60,0	57,8	69,1	54,8
P. V.	63,2	53,2	53,3	59,3	50,3
Val.	62,8	53,0	49,9	61,2	46,8
MEC.	66,7	57,9	54,1	65,7	52,6
MEC.	67,3	57,3	55,3	65,8	53,3

Tabla 2
Estadísticos obtenidos tanto en la muestra general como en cada una de las Comunidades Autónomas

	And.	Can.	Cat.	Gal.	P. Vas.	Val.	Mec.	Tot.
Nº íts.	60	60	60	60	60	60	60	60
Nº suj.	309	159	277	130	158	199	524	1.756
Media	37,5	36,7	39,3	38,0	34,1	33,8	36,0	36,4
Varian.	62,4	76,5	56,3	78,8	68,7	63,8	83,6	77,9
D.T.	7,9	8,6	7,5	8,9	8,3	8,0	9,1	8,8
Apunt.	–.38	–.09	–.28	–.51	.15	–.12	–.22	–.33
Kurtos	–.21	–.62	–.40	–.42	–.41	–.55	–.42	–.43
Vm.	16	14	18	14	14	14	11	11
V.M.	56	57	55	55	54	53	56	57
Mdna.	38	37	39	40	34	33	36	37
Alfa	.82	.85	.81	.86	.83	.81	.86	.85
E.T.	3,37	3,36	3,32	3,32	3,47	3,47	3,37	3,39
M P.	.63	.61	.66	.63	.57	.55	.60	.61
MIt/T.	.29	.32	.28	.33	.29	.29	.33	.31
MBise	.40	.43	.38	.45	.38	.38	.44	.42

En la tabla 2 se recogen los estadísticos obtenidos mediante el programa ITEM.

Tomando como grupo focal o de comparación la muestra del territorio MEC, los resultados obtenidos pusieron de manifiesto que en las Comunidades de Andalucía, Cataluña y Galicia la media obtenida en la prueba de Ciencias era estadísticamente superior a la media del MEC. En la Comunidad Valenciana y en el País Vasco, también se encontraron diferencias significativas pero, en este caso, las medias obtenidas en estas Comunidades eran inferiores a la media obtenida en el grupo del territorio MEC; finalmente, no se encontraron diferencias significativas entre las medias de la Comunidad Canaria y territorio MEC.

Ante estos resultados, y antes de llegar a ninguna conclusión respecto al rendimiento diferencial por Comunidades, se procedió, como ya se comentó anteriormente, a la evaluación del DIF mediante tres procedimientos: aproximación factorial, procedimiento Mantel-Haenszel y técnicas derivadas de la TRI.

Aproximación factorial

Tabla 3
Porcentaje de varianza explicada por cada factor en las distintas Comunidades Autónomas

Fac.	And.	Can.	Cat.	Gal.	P.V.	Val.	Mec.
1	22,5	23,0	21,3	23,5	20,4	20,3	33,7
2	6,61	6,87	7,31	6,85	6,85	6,52	6,42
3	5,76	6,63	6,14	6,06	6,40	6,28	6,05
4	5,49	6,10	5,83	5,97	5,62	6,04	
5	4,87	5,24	5,48	5,38	5,47	5,23	
6	4,60	5,19	5,28	4,94	5,31	5,15	
7	4,55	4,70	4,76	4,63	5,06	4,93	
8		4,57		4,55	4,89	4,61	
9		4,05		4,02	4,26		
10				3,78	4,25		
11				3,53			
12				3,34			

Tabla 4
Porcentaje de varianza explicada por el primer factor en el espacio de los factores

Fac.	And.	Can.	Cat.	Gal.	P.V.	Val.	Mec.
1	41,32	36,89	37,97	30,70	29,80	34,37	73,00

En la tabla 5 se incluyen los índices de congruencia que, como puede observarse, están próximos a la unidad.

Tabla 5
Indices de congruencia de Burt y Tucker

Andalucía - MEC	0,97
Canarias - MEC	0,95
Cataluña - MEC	0,94
Galicia - MEC	0,91
P. Vasco - MEC	0,96
Valencia - MEC	0,96

La estructura factorial de la prueba de Ciencias en las distintas Comunidades parece apuntar a la existencia de un factor dominante, teniendo en cuenta que el porcentaje de varianza explicada por el primer factor es, en todas ellas, significativamente más alto que el del resto de los factores. Un análisis de los ítems que saturan en este primer factor, mostró que son los mismos en todas las Comunidades. Esto, junto al elevado valor de los coeficientes de congruencia obtenidos, podría ser tomado como un indicador de la inexistencia de DIF.

Procedimiento Mantel-Haenszel

Siguiendo la recomendación de Holland y Thayer (1986) los valores obtenidos al aplicar el estadístico se transformaron a una escala logarítmica, «M-H Delta», de manera que la escala resultante fuera simétrica y el punto 0 correspondiera a la hipótesis nula.

«M-H Delta» indica la diferencia, para un determinado nivel de habilidad, en la dificultad del ítem entre los dos grupos comparados. Los valores de Delta negativos corresponden a aquellos ítems que han resultado más fáciles, por término medio, en el grupo de referencia. En general se acepta que un valor absoluto de M-H Delta mayor que la unidad indica posible existencia de DIF pues representa una diferencia de un 10% entre los grupos comparados y un valor mayor de 1,5 parece denotar una clara presencia de DIF.

La tabla 6 incluye para cada Comunidad, el número de ítems en los que se presume la existencia de DIF.

Tabla 6
Número de ítems en cada Comunidad en los que se puede apreciar DIF mediante el procedimiento de Mantel-Haenszel

	And.	Can.	Cat.	Gal.	P.V.	Val.
Items	4	4	3	6	6	3

Aproximación desde la TRI

En este contexto se considera que un ítem presenta DIF si la forma de su curva característica depende del grupo en el que se haya calculado (Lord, 1977, 1980; Mellenbergh, 1983), de ahí que los métodos desarrollados dentro de este marco estén basados en la comparación de las CCs estimadas separadamente en los distintos grupos.

Dado que al utilizar la aproximación factorial se había puesto de manifiesto la existencia de un factor dominante y el mismo en las distintas muestras, se ha asumido que la estructura factorial de la prueba en las distintas muestras es aproximadamente unidimensional, aunque somos conscientes de las limitaciones que tiene el uso de la proporción de varianza total explicada por el primer factor como indicador de la unidimensionalidad (ver Ferrando, 1996).

Una vez asumida la unidimensionalidad, supuesto obligatorio para los modelos logísticos de 1, 2 y 3 parámetros, en lugar de ir verificando los otros supuestos uno a uno (igualdad del parámetro *a* y existencia o no de aciertos por azar, parámetro *c*), se procedió al análisis de los residuos estandarizados obtenidos con cada uno de ellos. En caso de un buen ajuste, la media de los residuos estandarizados debería estar próxima a cero y la desviación típica en torno a la unidad. A medida que los residuos se alejan de cero en valor absoluto, peor será el ajuste del modelo. Si se utiliza un nivel de confianza del 95%, el análisis de residuales requiere calcular el porcentaje de residuos que caen fuera del intervalo formado por la media más y menos 2 desviaciones típicas aproximadamente, en caso de un buen ajuste este porcentaje no debería ser superior al 5%.

Los resultados obtenidos al realizar el análisis de residuales de los tres modelos logísticos en las distintas Comunidades Autónomas se ofrecen en la tabla 7. Como puede apreciarse, el modelo que mejor se ajusta es el modelo logístico de 2 parámetros. En este modelo, hay más de un 95% de residuos estandarizados incluidos en el intervalo de referencia. Por eso, ha sido este modelo el elegido para llevar a cabo el estudio del DIF.

Tabla 7
Proporción de residuos estandarizados incluidos en el intervalo formado por +/–2 desviaciones típicas

Model.	And.	Can.	Cat.	Gal.	P.V.	Val.	Mec.
Rasch	.399	.413	.377	.395	.401	.385	.409
2PL	.965	.970	.969	.980	.981	.978	.964
3PL	.858	.859	.845	.866	.790	.843	.835

Una vez seleccionado el modelo se calcularon los parámetros de cada uno de los ítems en el grupo de comparación o focal (grupo MEC) y, a continuación, se calcularon los residuos estandarizados obtenidos al comparar cada ítem de cada uno de los grupos de referencia (las distintas Comunidades Autónomas) con respecto a la curva característica que obtuvo en la muestra del MEC.

En la tabla 8 se incluyen aquellos ítems en los que el porcentaje de residuos estandarizados que caían fuera del intervalo formado por la media más menos 1,96 desviaciones típicas era mayor del 5%.

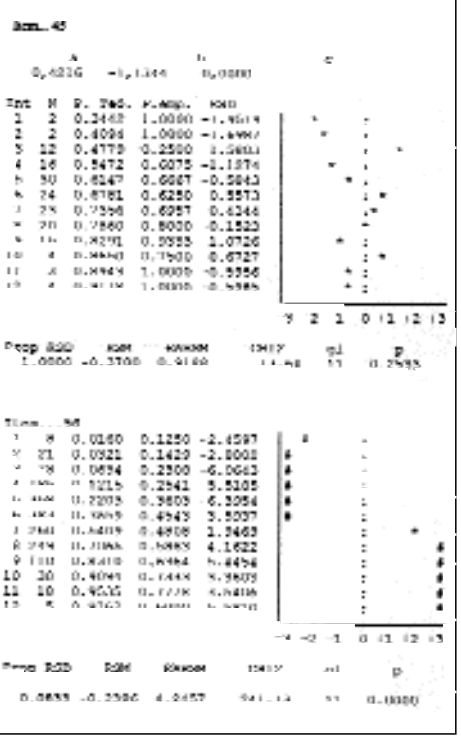
Tabla 8
Número de ítems fuera del intervalo y porcentaje de ítems que se ajustan

And.	Can.	Cat.	Gal.	P.V.	Val.
21 90%	7 96,54%	26 88,41%	16 93,82%	12 94,45%	17 92,12%

Como puede observarse, también mediante la aproximación de la TRI se han detectado ítems que muestran un comportamiento diferencial en las distintas Comunidades.

En la tabla 9 se incluye un ejemplo de los resultados obtenidos al aplicar el procedimiento de Linn y Harnisch con los ítems 45 y 56, utilizando como grupo de referencia la muestra del País Vasco. El módulo de análisis de residuales del programa GENESTE (San Luis y col., 1995), permite la visualización gráfica de los residuos obtenidos, de manera que es muy fácil observar para qué nivel/es de la variable de habilidad los sujetos del grupo focal obtienen valores más altos que los del grupo de referencia y viceversa. En el ítem 45 se observa que hay ajuste entre los resultados obtenidos, mientras que en el ítem 56 se observa un claro desajuste sobre todo para los niveles altos y bajos de la escala de habilidad.

Tabla 9
Ejemplo del análisis de residuales mediante el procedimiento de Linn y Harnisch



Los valores próximos al cero indican que las frecuencias empíricas (las correspondientes al grupo de referencia) son iguales a las pronosticadas por el modelo. Los signos (*) a la izquierda del cero indican que el ítem ha resultado más fácil, en esos niveles de rasgo, para los sujetos de la Comunidad de referencia, mientras que los signos (*) a la derecha del cero indican una mayor dificultad para estos sujetos. No obstante el signo (*) indica que los residuos se encuentran dentro del intervalo de confianza y, por lo tanto, no denotan desajuste. Por el contrario, el signo (#) representa aquellos residuos estandarizados que caen fuera del intervalo de confianza y, por lo tanto, indican desajuste.

Conclusiones

Los resultados obtenidos ponen de manifiesto que, independientemente de que pueda haber o no diferencias reales en el rendimiento en Ciencias de los niños y niñas de 13 años de las distintas Comunidades, antes de poder hacer las comparaciones pertinentes y sacar alguna conclusión al respecto es necesario depurar la prueba puesto que, a través de los análisis realizados, se han detectado ítems susceptibles de mostrar DIF. Por eso, a partir de los resultados iniciales, en los que se obtuvieron diferencias significativas entre las medias obtenidas en la prueba de Ciencias por las distintas Comunidades respecto a la muestra de territorio MEC, no se debe llegar a ninguna conclusión concreta sin antes haber depurado la prueba eliminando aquellos ítems que muestran un claro comportamiento diferencial en los grupos a comparar.

Por otra parte, creemos necesario seguir perfeccionando los procedimientos para el cálculo del DIF, dadas las diferencias encontradas, en cuanto a los ítems detectados, en función del procedimiento utilizado.

Finalmente, y partiendo de la necesidad de depurar la prueba antes de sacar conclusiones acerca de las diferencias encontradas en las distintas Comunidades en cuanto al rendimiento en Ciencias, creemos que sería interesante analizar en profundidad las causas de este comportamiento diferencial.

Referencias

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement,* 29, 1, 67-91.

Adams, R. J. y Rowe, K. J. (1988). Item bias. En J. P. Reeves (Ed.), *Educational Research, Methodology and Measurement An International Handbook.* Oxford: Pergamon Press.

Angoff, W. H. (1982): Use of difficulty and discrimination indices for detecting item bias. En R. A. Berk (Ed.). *Handbook of methods for detecting test bias,* Baltimore, MD: The Johns Hopkins University.

Barbero, I. Prieto, P. (1995). Efectos de la violación de los supuestos del modelo de Rasch sobre la robustez de las estimaciones. *Psicothema,* 7, 2, 419-426.

Bertrand, B. y King, F. (1989). Part I Sampling. En *A World of Differences. An International Assessment of Mathematics and Science. Technical Report.* Centre de Reserche et de Developpement en Mesure et Education. CREDME Université Laval Québec-Canada.

Cohen, A. S. y Kim, S. H. (1993). A Comparison of Lord's x2 and Raju's Area Measures in Detection of DIF. *APM,* 17, 1, 39-52.

Ferrando, P. J. (1996). Evaluación de la unidimensionalidad de los ítems mediante Análisis factorial. *Psicothema,* 8, 2, 397-410.

Holland, P. W.; Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *Technical Report 86-89.* Educational Testing Service Princeton NJ.

Holland, P. W. y Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (Eds.) *Test validity,* Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.

Ironson, G. H. y Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *JEM, 16,* 209-22.

Ironson, G. H. (1983). Using item response theory to measure bias. En R. K. Hambleton (Ed.). *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Lapointe, A. E., Mead, N. A. y Phillips, G. W. (1989). *A World of Differences. An International Assessment of Mathematics and Science.* Educational Testing Service.

Linn, R. L.; Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

Lord, F. M. (1977). *Practical Applications of Item Characteristic Curve Theory.* Princenton, N. J.: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, Nueva Jersey, LEA.

Mazor, K. M., Clauser, B. E. y Hambleton, R. K. (1994). Identification of nonuriform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psichological measurement,* 54, 2, 284-291.

Mellenberg, G. J. (1989). Item bias and Item Response *Theory. International* Journal of Educational Research, 13, 2, 127-143.

Navas, M. J. (1993). *Aplicación de la teoría de respuesta al ítem al campo de la medida. Creación de un banco de ítems para evaluar la capacidad matemática.* Tesis doctoral (sin publicar).

Navas, M. J. (1994). Utilización del análisis factorial y medidas del área como métodos en la detección de sesgo. *Psicothema,* 6, 3, 493-501.

Prieto, P. y Barbero, I. (1996). Detección del funcionamiento diferencial de los ítems mediante análisis de residuales: una aplicación de la TRI. *Psicothema,* 8, 1, 179-187.

Prieto, P. y Barbero, I. (). Identification of nonuniform DIF: A comparison of Mantel-Haens-

zel and IRT analysis procedures. *Educational and psychological Measurement* (en prensa).

San Luis, C.; Prieto, P.; Sánchez-Bruno, A. y Barbero, I. (1994). GENESTE: un programa de control para TRI. *Psicológica*, 16, 297-304.

Shepard, L.; Camilli, G.; Averill, M. (1981). Comparison of procedures for detecting test items bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.

Shepard; Camilli; Williams (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, 22, 2, 77-105.

Thissen, D.; Steinberg, L. y Wainer, H. (1988). Use of Item Response Theory in the Study of Group Differences in Trace Lines. En H. Wainer y Braun (Eds.) *Test Validity.* Hillsdale: Lawrence Erlbaum Associates.

Aceptado el 8 de noviembre de 1996