Papeles del Psicólogo, 2011. Vol. 32(3), pp. 254-264
http://www.cop.es/papeles

CALIDAD DE LOS ÍTEMS DE LOS EXÁMENES PIR

Rafael Moreno[1], Rafael J. Martínez[1] y José Muñiz[2]
[1]*Universidad de Sevilla.* [2]*Universidad de Oviedo*

Se explora la calidad de los ítems de las pruebas utilizadas en España para la selección de los candidatos a ocupar durante cuatro años las plazas que permiten obtener la titulación oficial de Especialista en Psicología Clínica (Psicólogo Interno Residente, PIR). Puesto que los resultados individuales de cada candidato no son públicos, lo que se analiza es una muestra intencional de las pruebas aplicadas en los últimos años, evaluando su ajuste a las directrices que la literatura ofrece para la construcción sistemática de ítems y pruebas. Los resultados más destacables de la exploración realizada son los siguientes. Junto al cumplimiento adecuado de varias directrices, se observa una insuficiente especificación de los contenidos y competencias objetos de la evaluación. Asimismo, más de un dieciséis por ciento de los ítems contienen errores formales o de contenido que dificultan la exposición del dominio de interés, lo cual pone de manifiesto una insuficiente revisión de las pruebas antes de su aplicación. Y en torno a un doce por ciento de los ítems inducen de manera directa la respuesta correcta, o indirectamente al permitir la exclusión de una o más de las alternativas. Todo ello muestra la conveniencia de construir de forma más rigurosa los ítems de las pruebas utilizadas para seleccionar a los Psicólogos Internos Residentes.
Palabras claves: *Construcción de ítems, Exámenes PIR, Directrices, España.*

This article explores the quality of the test items used in Spain for the selection of candidates for the four years contracts of Resident Intern Psychologist (PIR). That residency training program is necessary in order to obtain the official title of Specialist in Clinical Psychology. Since the individual responses to the test are not made public, we analyzed an intentional sample of the test items used in recent years, assessing their compliance with the guidelines that literature provides for the systematic construction of items and tests. The most outstanding results of the exploration carried out can be summarized as follow. Besides a general compliance with various guidelines, it was found an inadequate specification of contents and skills to be assessed. Over sixteen percent of the items present formal or content errors that make difficult the expresion of the items domain, which shows an insufficient editorial review of the test prior to its administration. In addition, around twelve percent of the items directly induce the correct response, or indirectly by allowing the exclusion of one or more of the alternatives. This shows the need to develop a more rigorous test items for selecting future Clinical Psychologists in Spain.
Key words: *Items construction, PIR Tests, Guidelines, Spain.*

La validez de las pruebas de evaluación descansa de modo fundamental en su construcción y uso, razón por la que se vienen dedicando importantes esfuerzos internacionales a la mejora de ambos aspectos (Muñiz y Bartram, 2007; Muñiz, Fernández-Hermida, Fonseca, Campillo y Peña, 2011; Muñiz y Hambleton, 2000; Muñiz, Prieto, Almeida y Bartram, 1999). Centrándonos en la construcción, realizarla de modo sistemático acorde con reglas claras y eficientes es condición importante para unos resultados satisfactorios y defendibles; más aún si esos resultados tienen trascendencia social o económica para la población evaluada, multiplicándose tales efectos con el tamaño de ésta.

Correspondencia: Rafael Moreno. *Departamento de Psicología Experimental. Facultad de Psicología. Universidad de Sevilla. Calle Camilo José Cela s/n. 41018 Sevilla. España.*
E-mail: rmoreno@us.es

En el ámbito español, una de las evaluaciones con tales características es el examen anual para la selección de candidatos a ocupar plazas de Psicólogo Interno y Residente (PIR) en la red de centros hospitalarios y de salud del Estado, para obtener la especialidad de Psicología Clínica. Dada la trascendencia de sus resultados, interesa conocer la calidad de tales pruebas y sus ítems. El único material disponible para ello son las propias pruebas utilizadas, al no publicar el organismo responsable los resultados individuales, ni las propiedades psicométricas de tales pruebas. El análisis de la corrección y defectos de éstas y sus ítems puede aportar una información relevante.

Muñiz y García Mendoza (2002) realizaron una primera exploración de los ítems de pruebas PIR, poniendo de manifiesto la presencia de algunos errores. El presente trabajo pretende continuar dicha tarea, explorando el

ajuste de ítems y pruebas a directrices que la literatura ofrece para su construcción sistemática. Se analiza una muestra intencional de las pruebas ya aplicadas y publicadas por el Ministerio de Sanidad, Política Social e Igualdad, utilizando como criterio de análisis una versión reciente de directrices sobre construcción de ítems y pruebas, resumida en la Tabla 1 (Moreno, Martínez y Muñiz, 2006).

MÉTODO

Material a analizar

Aunque las pruebas PIR se convocan a nivel nacional desde 1993, para este estudio consideramos como población de tales pruebas las aplicadas en las convocatorias de 2001 a 2008 por ser hasta el momento del inicio de este trabajo las publicadas en la web del Ministerio de Sanidad, Política Social e Igualdad (2011), organismo responsable de dichas pruebas. En dicha web aparecen también las respuestas consideradas correctas para cada ítem. Para este estudio se eligieron como muestras las pruebas de 2005 y 2008, respectivamente intermedia y última de las constitutivas de la población considerada, cada una con 260 ítems de elección múltiple con cinco alternativas de respuesta.

Procedimiento

Todos los ítems de las dos pruebas PIR seleccionadas (2005 y 2008) fueron sometidos a los criterios de la Ta-

TABLA 1
DIRECTRICES PARA ÍTEMS DE ELECCIÓN MÚLTIPLE (MORENO, MARTÍNEZ Y MUÑIZ, 2006)

A. Sobre fundamentos

1. Para facilitar la validez de los ítems y pruebas, el objetivo y el dominio de la evaluación deben definirse del modo más detallado posible.
2. Es conveniente especificar también el contexto en el que se van a utilizar los ítems, lo que incluye la población a la que irán dirigidos y las circunstancias en que se aplicarán.

B. Sobre la expresión del dominio y el contexto en cada ítem y prueba

3. El objetivo, dominio y contexto de interés deberán ser los criterios determinantes en la construcción. Cada ítem debe recoger una unidad significativa de tal referente y formar con los demás una prueba relevante.
4. Cada ítem debe mostrar claramente el contenido pretendido. Tanto la sintaxis como la semántica deben estar ajustadas a las del dominio y contexto de referencia, sin añadir dificultades innecesarias.
5. Construidos los ítems ha de procurarse el ajuste de su conjunto con el dominio y el contexto de referencia, especialmente a través del número de ítems y su distribución en la prueba.

C. Sobre las opciones de respuesta

C.1. Aspectos que deben facilitar la expresión del dominio de interés y no añadir dificultades
6. Cada opción ha de ser continuación o respuesta lo más breve posible del enunciado.
7. Suele resultar más eficiente la construcción cuando la opción correcta es sólo una y no lo sea parcialmente. En otro caso, deben aclararse los criterios introducidos.
8. La disposición espacial de las opciones debe facilitar la percepción del contenido del ítem.
9. El contenido de cada opción debería ser autónomo del resto. Por ello conviene usar con cautela las opciones "Todas las anteriores" y "Ninguna de las anteriores".
10. Las opciones de cada ítem deberían aparecer ordenadas, sin exigir esa tarea previa.
C.2. Aspectos que deben impedir la inducción indebida de una respuesta inadecuada
11. Las opciones deberían ser plausibles para el sujeto que no conoce la respuesta correcta, permitiendo al que la conozca identificarla y desechar el resto. Contenidos y términos próximos a la opción correcta y errores comunes de los sujetos son medios adecuados para lograrlo.
12. Deben evitarse indicios sobre la corrección o incorrección de una o más de las opciones. Evítese usar términos que puedan aportar información indebida sobre lo planteado en el enunciado.
13. Deben evitarse características que, sin ser indicios claros sobre la corrección o no de una opción, la destaquen del resto y por ello puedan plantear la duda de que tal diferencia sea indicador significativo. La longitud y el contenido diferencial de una opción son errores usuales.
14. El número de opciones a incluir debe permitir la plausibilidad de todas las opciones para el sujeto que no conozca la correcta. Tres suele ser adecuado, aunque un número mayor también podría serlo.
15. Debe cuidarse que el conjunto de ítems como tal no contenga tampoco ninguna clave inductora indebida. Por ello, conviene revisar más de una vez todo lo hecho a partir de las directrices previas.

bla 1. Para depurar los criterios con los que aplicar cada una de esas directrices se realizó un estudio piloto con una submuestra accidental de ítems de las dos pruebas. Posteriormente dos codificadores analizaron independientemente el ajuste o no de todos los ítems de las dos muestras a tales directrices, realizando una prueba de concordancia con una submuestra aleatoria. En los casos de desacuerdo, se ha señalado como incumplimiento de una directriz cuando uno de los codificadores así lo seguía considerando tras una revisión conjunta del ítem en cuestión. Se ha tratado de ser exhaustivos al señalar incumplimientos de cada directriz, mostrando en cambio sólo algunos ejemplos de aspectos señalados como información adicional, incluyendo desviaciones poco significativas pero mejorables.

RESULTADOS

La concordancia en las categorizaciones de los dos codificadores independientes para cada una de las directrices se estimó mediante el cálculo del porcentaje de acuerdos y del índice *kappa* con corrección de los acuerdos debidos al azar. Para ello se seleccionó una muestra aleatoria simple de 47 ítems (*E.M.* = 7.8%, con un *N.C.* = 95%; y proporción de acuerdos estimados de 0.9). En la tabla 2 pueden observarse los resultados de estos análisis, junto con el resumen de los porcentajes de incumplimientos de las diferentes directrices examinadas.

Respecto a la concordancia en la codificación se estima un porcentaje de acuerdo medio del 97.5%, con valores que oscilan entre un mínimo del 87.2% hasta un máximo de 100%, ocurriendo esto último en 7 de las 13 categorizaciones realizadas. La estimación del valor del índice *kappa* refleja también un alto grado de concordancia con valores superiores a 0.75, con la excepción de la categorización de la directriz 4 referida a la corrección sintáctica y semántica de los ítems, en cuyo caso se obtuvo un índice extremadamente bajo (k = .044), ligado en todo caso a la falta de acuerdo sobre el único caso de incumplimiento registrado.

Los resultados del análisis realizado son expuestos a continuación con arreglo a cada una de las directrices. El objetivo de las pruebas –*Directriz 1*- queda expuesto en las convocatorias oficiales (véase por ejemplo la Orden SAS/2448/2010 publicada en Boletín Oficial del Estado, 2010); son pruebas *selectivas*, para que los examinandos que conforme a ellas ocupen los puestos que permiten tal derecho puedan elegir entre las plazas ofertadas. En cambio, el dominio a evaluar no es explicitado. Sólo se menciona en la Orden 14882 de 27 de junio de1989 (Boletín Oficial del Estado, 1989), referida exclusivamente a las pruebas MIR –Medicina- y FIR –Farmacia-, al señalar que las pruebas *"versarán sobre el contenido de las áreas de enseñanza comprendidas en las licenciaturas respectivas"*. Ese dominio de materias de la licenciatura ha sido generalizado de hecho a las restantes pruebas incorporadas con posterioridad, como es el caso de Psicología que nos ocupa, quedando valorado el nivel de doctorado en términos de expediente académico a añadir a la puntuación de la prueba. La web de la Asociación Nacional de Psicólogos Clínicos y Residentes, ANPIR (2005) confirma que *"Los contenidos del examen PIR versan sobre todas las asignaturas (obligatorias y optativas) del plan de la carrera de Psicología"*, y especifica que las pruebas PIR se refieren a las distintas áreas académicas de Psicología, aunque teniendo mayor peso los contenidos referidos a los aspectos clínicos, entendiendo como tales los de Psicopatología, Terapias, Evaluación, Psicodiagnóstico, Personalidad y Psicología diferencial. Lo que no queda señalado es el tipo de competencias –memorísticas, de razonamiento o de otro tipo- que desean evaluarse dentro del dominio de contenidos de la licenciatura.

Sobre el contexto en el que se desarrollan las pruebas –*Directriz 2*-, puede considerarse que las citadas convocatorias lo delimitan adecuadamente al señalar circunstancias y detalles como horas de inicio, duración, lugar del examen, condiciones de confidencialidad de los ejercicios, cuaderno de examen, hoja de respuesta, modo de cumplimentarla y avisos para la entrada y salida del aula donde se realizan las pruebas.

En las convocatorias no queda señalado si el objetivo, dominio y contexto de interés fueron utilizados explícitamente como los referentes para la construcción o elección de cada ítem a incluir en las pruebas –*Directriz 3*-, y por ello no resulta posible evaluar si los ítems y pruebas de la muestra estudiada se ajustan a esos referentes –*Directriz 5*-. Es cierto que todos los ítems analizados son unidades del dominio psicológico al que se refiere el objetivo, y también que son adecuados al contexto planteado para la prueba; pero al no conocerse si previamente a la construcción de ítems se especificó una determinada distribución de contenidos y de competencias de interés, no ha sido posible evaluar el ajuste de las pruebas a ambos criterios. Como análisis alternativo describimos las distribuciones de ambos aspectos en las muestras consideradas.

En cuanto a contenidos, en ambas pruebas se encuentran las mismas categorías, apareciendo la mayoría de los ítems agrupados por ellas, aunque los 10 últimos ítems de cada prueba son de contenidos diversos, probablemente de reserva. Como se observa en la tabla 3, los contenidos relacionados con la clínica suman 51.3% en 2005 y 76.6% en 2008, siendo Psicopatología el más frecuente en los exámenes muestra, y presentando el resto de este tipo porcentajes superiores al 10 con la excepción de Psicodiagnóstico en 2005, y Personalidad y Psicología Diferencial en ambas muestras estudiadas. Los contenidos no clínicos presentan por su parte porcentajes inferiores al 10% de los ítems, con la excepción de los de Procesos Básicos y Psicología Evolutiva y de la Educación, ambos en 2005, que ascienden al 13.1% y al 11.1% respectivamente, superiores por tanto a algunos de los clínicos.

TABLA 2
PORCENTAJES DE ERRORES O INCUMPLIMIENTO DE DIRECTRICES SOBRE CONSTRUCCIÓN DE ÍTEMS EN LAS PRUEBAS PIR DE 2005 Y 2008, E ÍNDICES DE ACUERDO DE LA CATEGORIZACIÓN

	2005		2008				
Directriz	**Errores**	**%**	**Errores**	**%**	**% Acuerdo**	**k**	***EM***
A. Sobre fundamentos							
1. Dominio							
Contenido	a		a		87.2%	.855	.108
Competencias	a		a		97.9%	.879	.235
2. Contexto	0	0.0%	0	0.0%	100.0%	1.00	
B. Sobre la expresión del dominio y el contexto en cada ítem y prueba							
3. Unidad significativa	a		a				
4. Sintaxis y semántica	1	0.4%	0	0.0%	91.5%	.044	.979
5. Ajustado	a		a				
C. Sobre las opciones de respuesta							
C.1. Aspectos que deben facilitar la expresión del dominio de interés y no añadir dificultades							
6. Continuación	47	18.1%	42	16.1%	97.9%	.911	.173
7. Sólo una correcta	5	1.9%	5	1.9%	100.0%	1.00	
8. Bien espaciadas	0	0.0%	0	0.0%	100.0%	1.00	
9. Autónomas	b		b				
10. Ordenadas	7	2.6%	2	0.7%	100.0%		
C.2. Aspectos que deben impedir la inducción indebida de una respuesta inadecuada							
11. Plausibles	0	0.0%	3	1.1%	100.0%	1.00	
12. Sin indicios	14	5.3%	0	0.3%	97.9%	.877	.238
13. Homogéneas	19	6.9%	24	9.2%	95.7%	.776	.303
14. Número adecuado	0	0.0%	0	0.0%	100.0%	1.00	
15. Sin inducción	0	0.0%	2	0.8%	100.0%	1.00	

[a] No evaluables, al no estar explícito el objetivo perseguido con cada ítem y con el conjunto de la prueba, en términos de contenidos y competencias.

[b] Los incumplimientos de esta directriz se han categorizado dentro de los de las directrices 6, 7 y 12.

K: coeficiente Kappa

EM: Error de medida

En cuanto a competencias, la inmensa mayoría de los ítems exigen identificaciones memorísticas (96.2% y 90.4% en una y otra prueba), basadas en el recuerdo de alguna información dada en el enunciado a asociar a alguna otra que aparece en las opciones; por ejemplo definición de un término técnico o viceversa (como 05/4 y 05/20 respectivamente), característica de algún concepto (05/36), conexión entre dos ideas o nociones (05/179, 05/215), autoría de alguna idea o viceversa (05/1 y 05/21 respectivamente) o finalidad de algún instrumento o viceversa (05/197 y 05/209) (Esta anotación indica convocatoria e ítem respectivamente; si además va seguida por uno o más números, éstos indican opciones. Por ejemplo, la anotación 05/9/1-2 señala las opciones 1 y 2 del ítem 9 de la prueba de 2005). Otra competencia encontrada sería la de razonar para poder responder correctamente al ítem. En nuestra opinión, ello exige que el examinando realice alguna comparación de una covariación entre términos para responder correctamente. En la prueba de 2005 no en-

TABLA 3
DISTRIBUCIÓN DE CONTENIDOS DE LOS ÍTEMS DE LAS PRUEBAS

	2005			2008		
Temáticas	**Ítems**	***n***	**%**	**Ítems**	***n***	**%**
1. Psicopatología[a]	13[b], 61, 66-72, 79, 101-118, 120, 214-235, 239, 241, 243, 254, 260 64, 73-74, 76-77,	56	21.6%	1-17, 19-72, 124, 129, 208 18, 85-87, 91-93,	74	28.5%
2. Terapias y tratamientos[a]	142-164, 186-190, 193-195, 244-246, 251-253, 255	43	16.6%	95-96, 98, 101-123, 125-128, 130-153	61	23.5%
3. Psicodiagnóstico y Evaluación conductual[a]	78, 80, 196-213, 236	21	8.1%	73-79, 81-83, 177-206	40	15.4%
4. Personalidad y Ps. Diferencial[a]	23-24, 31, 36-40, 191-192, 238, 240, 242	13	5.0%	84, 154-176	24	9.2%
5. Procesos Básicos e Historia	1-2, 4-12, 14-18, 20-21, 25-30, 32-35, 119, 247-250, 256	34	13.1%	88-90, 94, 97, 99-100, 217-225, 252, 256	18	6.9%
6. Psicometría, Estadística, Métodos	81-100	20	7.7%	226-234, 253, 259-260	12	4.6%
7. Psicología Social y de las Organizaciones	22, 121-141	22	8.4%	243-251, 254,	11	4.2%
9. Psicología Evolutiva y de la Educación	3, 62-63, 65, 75,				258	
10. Psicobiología y Psicofisiología	165-185, 257-259	29	11.1%	80, 235-242, 255	10	3.8%
	19, 41-60, 237	22	8.4%	207, 209-216, 257	10	3.8%
Total		**260**			**260**	

[a] Contenidos relacionados con el área clínica.
[b] En referencia a los números de ítems, los guiones en esta tabla representan intervalos de preguntas.
n: número de ítems

contramos ningún ítem de este tipo, mientras que en la de 2008 habría los once siguientes: 16, 19, 22, 25, 39, 50, 54, 69, 78, 109, 160, ilustrados a continuación por uno de ellos.

08/54. ¿Cómo podemos diferenciar un cuadro demencial de un síndrome amnésico en un paciente con continuas quejas sobre su memoria?:

1. Por la edad del sujeto.
2. Por la presencia de amnesia retrógrada.
3. Por la conservación de la memoria operativa.
4. Por la presencia de deterioro cognitivo global que progresa a medida que avanza el trastorno.
5. Por la presencia de amnesia anterógrada.

En todo caso hay que advertir que estos ítems podrían ser memorísticos y no de razonamiento si los contenidos que plantean estuvieran expresados en los textos psicológicos estudiados por todos o algunos de los examinandos, no exigiendo entonces a éstos realizar la covariación contenida en el ítem, como por ejemplo ocurre en los ítems 05/32 y 05/45. Por ello, los 11 (4.2%) mencionados serían en todo caso el número máximo de ítems que exigirían razonamiento además de memoria.

Otros ítems exigen aplicar conocimientos a la resolución de casos prácticos o ejemplos de alguna noción o característica psicológica. Hemos identificado los siguientes ítems de este tipo: 05/05, 05/93, 05/94, 05/95, 05/96, 05/124, 05/145, 05/153, 05/156, 05/161, 05/216, 05/217, 05/225 y 05/228, y 08/28, 08/31, 08/89, 08/94, 08/96, 08/98, 08/107, 08/126, 08/249, 08/250, es decir 14 y 10 ítems en las dos pruebas de muestra, 3.8% y 5.3% respectivamente. En definitiva, la gran proporción de ítems de competencias memorísticas podría considerarse un sesgo si entendemos que el perfil de la Titulación de Psicología, que se pretende evaluar, no sólo incluye ese tipo de competencias y sí al menos las otras dos aquí consideradas.

De acuerdo a nuestro análisis, los ítems no impiden ni obstaculizan seriamente la comprensión del texto por problemas de sintaxis, semántica o claridad de sus expresiones –*Directriz 4*-. Una excepción la constituye el ítem 05/180, (0.4% del total) que probablemente fue anulado, dado que aparece sin respuesta en la plantilla publicada. No obstante, existen preguntas que contienen elementos o fallos que pueden distraer al lector de su tarea, y que hubieran resultado corregibles en una revisión cuidadosa de las pruebas analizadas. Hay signos de puntuación claramente incorrectos o al menos mejorables -como en los ítems 05/5, 05/100, 05/161, 05/173, 05/180, 05/189 y 05/190, más los 08/122 y 08/252-, falta de tilde –como en 08/122-, tildes indebidas que cambian el significado de la palabra -como el "aún" en lugar de "aun" de los ítems 05/7 y 05/216-, referentes no claros de expresiones –como el de "se realizan" en 05/36-, errores tipográficos –"de" por "se" en 05/76-, diferentes formatos en referencias a citas de texto –a veces en cursivas como en 05/104 o 05/112 y otras sin ellas como en 05/218 o 05/221-, uso de términos de otras lenguas sin uso de las cursivas -05/250, 08/44-, anglicismo – "similaridades" en 08/166-, así como diferencias al dirigirse al lector en una misma prueba, al que algunos ítems tratan de usted -05/100, 05/224- y otros tutean -05/193, 05/194-.

Respecto a las opciones de respuesta consideramos en primer lugar la debida continuación o concordancia sintáctica de las opciones respecto al enunciado, sin repetición innecesaria de términos del enunciado y con brevedad –*Directriz 6*-. Encontramos que unos ítems plantean una pregunta cuyas posibles respuestas son las opciones, y otros comienzan una frase a completar con las opciones. En ambos casos, la mayor parte de los ítems cumplen con los requisitos de esta directriz, aunque hay excepciones.

Los ítems señalados a continuación tienen una o más opciones que no concuerdan sintácticamente con el enunciado al que deberían completar: 05/74/2-5, 05/83/5, 05/94/5, 05/95/1-5, 05/179/1-2-5, 05/204/1-2-4-5, 05/214/4, 05/215/3-5, 05/219/5 y 05/255/2-4-5 en la primera prueba analizada, y 08/181/5, 08/195/1-2-3-4-5, 08/203/3-4, 08/244/5 y 08/248/4 en la segunda; respectivamente 10 y 5 ítems, 3.8% y 1.9% del total de cada prueba.

Es mayor el número de ítems que repiten innecesariamente un mismo contenido en todas sus opciones, en lugar de escribirlo una sola vez en el enunciado. En la prueba de 2005, lo hacen los ítems 9, 16, 27, 80, 109, 115, 121, 132, 139, 142, 159, 177, 178, 185, 188, 189, 196, 203, 228, 234, 236, 240, 246, 249 y 257, y en la de 2008 los ítems 3, 11, 13, 14, 45, 55, 60, 61, 80, 83, 88, 89, 91, 95, 99, 102, 105, 118, 125, 130, 131, 148, 150, 155, 192, 220, 231, 238 y 239; es decir, 25 y 29 ítems en la primera y segunda prueba, 9.6% y 11.1% respectivamente de este incumplimiento de la directriz 6, que obliga a leer de más a los examinandos, restándoles innecesariamente tiempo para la tarea central de responder lo mejor posible al conjunto de la prueba.

Menor es el número de ítems en los que al menos una de las opciones tiene una longitud excesiva –considerada por nosotros como más del doble de palabras que el enunciado-. Son los 26, 30, 35, 40, 79, 125, 126, 147, 210, 220, 248 y 250 en la prueba de 2005, y 33, 53, 127, 166, 227, 241, 251 en la de 2008; es decir 12 y 7 ítems (4.6% y 2.7%). En todo caso algunos de ellos resultan muy claros, haciendo dudar de la conveniencia del criterio fijado para medir la no brevedad de las opciones. A pesar de ello, entendemos que en general el incumplimiento de este componente de la directriz lleva al examinando a leer textos que podrían simplificarse en muchos de los casos. Eso sin mencionar que una buena parte de los enunciados muy breves resultan inadecuadamente sucintos, pudiendo mejorarse con una redacción más completa. La atención a este aspecto puede ayudar a construir ítems más adecuados. En resumen, sumando los tres componentes señalados encontramos en cada prueba un total de 47 y 42 ítems que incumplen la presente directriz 6 (18.1% y 16.1%).

La información sobre si hay una única respuesta correcta por ítem –*Directriz 7*- no queda suficientemente explicitada en la convocatoria de las pruebas. Y en las instrucciones que se entregan en el momento de la prueba, la información sobre una sola respuesta correcta hay que entenderla indirectamente a través del singular utilizado en la hoja de instrucciones: *"Compruebe que la respuesta que va a señalar en la Hoja de Respuestas corresponde al número de pregunta del cuestionario"*. Ese criterio sobreentendido queda confirmado cuando se observa en la Hoja de Respuestas de las pruebas analizadas que cada ítem valorado aparece con una sola respuesta correcta, información que el examinando no tiene obviamente en el momento de la prueba.

El análisis de dichas respuestas, apoyado en el caso de algunos ítems con la opinión de expertos en los respectivos contenidos, permite señalar que efectivamente la mayoría de los ítems tienen una y sólo una opción correcta. Sin embargo, los siguientes ítems tienen más de una opción correcta: 05/9, 05/106, 05/180, 08/68, 08/189, 08/214, 08/244, y de hecho todos ellos fueron anulados. Adicionalmente, otros dos ítems que no fueron anulados, 08/178 y 08/182, podrían presentar el mismo problema en opinión de algunos de los expertos consultados. Por su parte, los ítems 05/77 y 05/194 fueron anulados probablemente por no ser ninguna de sus opciones claramente correctas. En resumen, al menos en torno a 5 ítems (1.9%) en cada prueba no se ajustaron a esta directriz 7.

La disposición espacial de las opciones –*Directriz 8*– es vertical en todos los ítems, apareciendo éstos en dos columnas, con letras adecuadas en tamaño y claridad, y espaciados en el conjunto de la prueba y de cada página. Al menos así es en la versión que aparece en la web, desconociendo si se trata del mismo formato que recibieron los examinandos. Una ligera dificultad para la lectura es la partición de algunos ítems en dos páginas o en dos columnas diferentes. En la versión analizada, ello ocurre en 20 de las 24 páginas que contienen los de la prueba de 2005, y 22 de las 26 en la de 2008.

La falta de autonomía entre las opciones de respuestas –*Directriz 9*- puede producirse porque el contenido de una opción es parte del de otra opción, o porque ambas sean similares. Como los solapamientos son más fáciles de apreciar por expertos en el contenido de los ítems, los hemos consultado aunque sólo para aquellos ítems en los que teníamos dudas; por ello cabe la posibilidad de que no hayamos detectado algunos otros que incumplan esta directriz. Pueden mencionarse en todo caso los ítems ya señalados en la directriz 7 con más de una respuesta correcta, y los 05/39, 08/188 y 08/229 con solapamiento entre opciones incorrectas señalados como indicios indebidos en la directriz 12. Otro modo de incumplir la autonomía de las distintas opciones son las opciones "Ninguna de las anteriores" o "Todas las anteriores" (Martínez, Moreno, Martín y Trigo, 2009). Aunque como tales no aparecen en las pruebas analizadas, algunos ítems contienen opciones que podrían considerarse una versión de "Ninguna de la anteriores": aquéllas con un contenido que niega que sea cierto lo planteado en el enunciado, y por tanto en el resto de opciones. En el listado de ítems que incumplían la directriz 6 por tener opciones que no son continuación sintáctica de sus enunciados, los siguientes lo hacían en el modo que acabamos de decir: 05/83/5, 05/95/5, 05/214/4, 05/215/5, 05/255/5, 08/143/5 y 08/181/5. En conjunto, el análisis de los incumplimientos de esta directriz se solapa con los realizados sobre las directrices 6, 7 y 12, por lo que no resulta conveniente duplicar el recuento de estos errores, y además debe hacer reconsiderar la pertinencia de la versión actual de esta directriz 9.

Respecto al orden de las opciones de cada ítem –*Directriz 10*- no sabemos si en las diferentes versiones que menciona la instrucción 1 de la prueba de 2008 es siempre el mismo, por lo que los siguientes resultados corres-

ponden sólo a la versión publicada. Las opciones de los ítems 05/83, 05/92, 05/94, 05/96, 08/86 y 08/201, aparecen innecesariamente desordenadas, lo que añade una cierta dificultad o tarea adicional contraproducente e irrelevante para el contenido que preguntan. Dificultad que es mayor cuando las opciones presentan dos o más contenidos, obligando probablemente a muchos examinandos a ordenar el contenido de las opciones antes de poder responder; es lo que ocurre en el ítem 05/47 que aparece a continuación, de modo semejante a los 05/185 y 05/238.

05/47. Los lemniscos que aparecen en un corte transversal del mesencéfalo a nivel de los colículos superiores son:

1. Medial, espinal y trigeminal.
2. Lateral, medial y trigeminal.
3. Medial, lateral, trigeminal y espinal.
4. Lateral, espinal y trigeminal.
5. Trigeminal y espinal.

En resumen, 7 ítems en la prueba del 2005, y 2 en la de 2008 incumplen esta directriz (2.6% y 0.7%). Además, existen otros ítems que aunque sin llegar a incumplir la directriz dificultan la lectura. Sus opciones presentan dos o tres contenidos –"proporción" y "disminución" en el ítem de ejemplo que aparece a continuación- que podrían servir para agruparlas y facilitar su lectura.

72. El Índice de Masa Corporal es:

1. La proporción entre la altura y el cuadrado del peso.
2. La disminución del peso desde los últimos seis meses.
3. La proporción entre la altura y el peso
4. La disminución de la grasa corporal en función de la altura.
5. La proporción entre el peso y el cuadrado de la altura.

La gran mayoría de los ítems presentan conjuntos plausibles de opciones, enmarcando adecuadamente a la correcta –*Directriz 11*-. La mayoría lo hacen utilizando contenidos pertenecientes a un mismo campo temático, en el que por tanto es más probable la utilización recomendada de errores comunes del examinando. Tan sólo hemos encontrado 3 ítems (1.1%) -los 08/32/4, 08/160/2 y 08/231/4- con alguna opción no plausible por fácilmente desechable.

Mientras que en la directriz recién comentada la plausibilidad se considera en términos de lenguaje técnico, hay que incluir también como indicios indebidos los términos del lenguaje ordinario que, por su simple apariencia, pueden informar o dar pistas al examinando que no tenga conocimientos suficientes para elegir o descartar determinadas opciones –*Directriz 12*-. En todo caso, el límite entre ambos lenguajes es a veces difuso, especialmente cuando el técnico es fácil o de uso común. La opción correcta del siguiente ítem, la 4, ilustra lo anterior.

05/4. ¿Cómo se denomina el fenómeno de aprender dos lenguas de forma simultánea desde el nacimiento (durante las fases iniciales de la adquisición del lenguaje)?:

1. Bilingüismo aditivo.
2. Bilingüismo sustractivo.
3. Adquisición de segunda lengua.
4. Bilingüismo nativo.
5. Atrición lingüística.

Hemos encontrado diversos incumplimientos de esta directriz. Los ítems 05/6/5, 05/10/1, 05/14/1, 05/20/2 y 05/170/3 orientan hacia la respuesta correcta al incluir en el enunciado el mismo o similar término distintivo de la opción correcta. Los 05/89/3, 05/160/3, 05/255/2-4 y 08/229/1-2-5 orientan por la concordancia o falta de ella en género o número gramatical con el enunciado. Otros ítems permiten excluir algunas opciones como incorrectas: los 05/6/2-3 y 05/39/4-5 por tener dos opciones de contenido semejante, que por tanto pueden ser desechadas al tener cada ítem una sola correcta; los 05/22/3, 05/170/1, 05/188/5 por contener alguna opción de contenido incompatible con lo planteado en el enunciado; y el 05/90/1-2 por usar los términos "siempre" y "nunca" que raramente suelen ser correctos. En resumen, hemos detectado 14 incumplimientos de esta directriz (5.3%) en la prueba de 2005 y uno (0.3%) en la de 2008. Tales incumplimientos hacen que el número de opciones relevantes se reduzca en los ítems señalados, distorsionando la corrección al azar planeada en principio para una opción correcta y cuatro distractoras, lo cual perjudica a los candidatos mejor preparados al beneficiar indebidamente a los que no conocen la respuesta correcta de esos ítems.

Al analizar la homogeneidad de las opciones en cuanto a su contenido y longitud –*Directriz 13*- se detecta pocos ítems que la incumplen. Algunos destacan indebidamente la opción correcta con un contenido diferente al resto -05/181/4, 05/251/2- o más detallado que el de las alternativas -05/5/5, 05/20/2, 05/34/2, 05/60/5, 05/79/1, 05/248/1, 08/53/1, 08/54/4, 08/83/3, 08/94/1, 08/140/3, 08/218/2 y 08/240/2-. En total 8 ítems en 2005 y 7 en 2008 (3.1% y 2.7%).

Otros ítems destacan indebidamente una de las opciones incorrectas: Los 05/83/05, 05/94/5, 05/147/5, 05/210/2, 05/214/2, 05/215/1, 08/173/3 lo hacen en la longitud, –lo que se ha anotado cuando la opción

destacada supera en al menos un tercio a la anterior en el número de palabras-. Y otros lo hacen utilizando un contenido diferente al resto, lo que introduce un elemento distorsionador innecesario que puede ser evitado igualando dicha opción al resto o viceversa; son los ítems 05/86/4, 05/123/1, 05/237/2, 05/258/5, 08/20/3, 08/21/2, 08/28/1, 08/39/5, 08/67/4, 08/81/2, 08/86/3, 08/102/5, 08/144/2, 08/145/5, 08/146/3, 08/149/3, 08/167/5, 08/200/1, 08/229/5 y 08/260/4. En total son 10 y 17 los ítems que destacan la opción incorrecta en las dos pruebas analizadas (3.8% y 6.5%), y por tanto un total 18 y 24 (6.9% y 9.2%) los ítems que incumplen la directriz 13 de una u otra forma.

En cuanto al número de opciones de cada ítem –*Directriz 14*- todos los incluidos en las dos pruebas analizadas tienen cinco. Ello exige que sea factible y relevante presentar ese número de alternativas plausibles en todos los contenidos preguntados, algo no siempre fácil y más cuando la población de contenidos a preguntar es tan amplia como la de todas las materias de Psicología. Puede ser la razón para los ítems señalados en la directriz 12 con opciones triviales o repetidas, y los que en la directriz 13 presentan una opción incorrecta diferente al resto. A ellos podrían añadirse los ítems señalados en la directriz 9, que usan en algunas de sus alternativas una versión de la opción "Ninguna de las anteriores". En resumen, un número elevado de ítems con problemas de mayor o menor importancia que avala la recomendación de la literatura de reducir el número de las alternativas a utilizar (Abad, Olea y Ponsoda, 2001; Bruno y Dirkzwager, 1995; Delgado y Prieto, 1998; Haladyna, Downing y Rodriguez, 2002; Rogers y Harley, 1999), aunque vaya en contra de la preferencia inicial por un amplio número de alternativas como modo de reducir la influencia de las respuestas dadas al azar, o que al menos recuerda el especial cuidado a poner para lograr la plausibilidad y homogeneidad de todas las opciones.

En lo que se refiere a la evitación de claves inductoras de la respuesta correcta de una pregunta mediante información aportada en otras preguntas de la prueba –*Directriz 15*-, tan solo hemos encontrado dos casos de incumplimiento de esta directriz, ambas en la prueba de 2008: la respuesta correcta al ítem 194 está literalmente escrita en el enunciado del 155, y los ítems 115 y 116 intercambian los contenidos del enunciado y de la respuesta correcta, la 4 en ambos.

CONCLUSIONES

La exploración realizada permite obtener una descripción de las muestras seleccionadas de pruebas e ítems de las convocatorias PIR. En primer lugar, el objetivo de las pruebas y el contexto en el que aplicarlas quedan claramente expuestos en la convocatoria, mientras que el dominio a evaluar con las pruebas es señalado de modo poco explícito y parcial: poco explícito porque hay que remontarse a una orden de 1989 para encontrar señalado que el contenido a evaluar es el de la licenciatura correspondiente, lo que implica además cierta vaguedad en los límites del dominio, a lo que se añade la no especificación de la distribución de los contenidos que interesa considerar; y parcial porque nada se dice respecto al tipo de competencias a evaluar.

Ante esta situación la evaluación del posible ajuste de las pruebas a un referente no explicitado ha de quedar en suspenso. Describiendo como vía alternativa de información las dos pruebas analizadas, una y otra contienen respectivamente la mitad y tres cuartas partes de los ítems relacionados con los aspectos clínicos. En cuanto a las competencias evaluadas, la muy mayoritaria proporción de ítems de tipo memorístico podría considerarse un sesgo o fallo de representatividad de las pruebas respecto a la variedad de competencias que pueden lograrse en los estudios de licenciatura, y actualmente del grado. Tal vez este sesgo memorístico esté generado en gran medida por la estrategia seguida para poder responder de forma clara a posibles reclamaciones, consistente en exigir a los expertos que construyen los ítems que la solución aparezca explícita en un texto de referencia. Al ser esta exigencia más fácil de cumplir en principio con ítems puramente memorísticos que con otros que requieran estrategias cognitivas como inferir, sintetizar o aplicar, el número de éstos podría estar quedando limitado por tal razón. Sin embargo, los ítems no puramente memorísticos bien construidos que ya aparecen en las pruebas PIR, y también en pruebas MIR (Médicos Internos y Residentes) recientes, son muestras de que es posible corregir el sesgo actual que comentamos.

En lo que se refiere a los ítems, todos presentan cinco opciones de respuesta con una sola correcta, están distribuidos espacialmente de manera que resulta fácil su lectura, y en su mayoría están expresados adecuadamente desde el punto de vista sintáctico y semántico. Sin embargo, en las pruebas analizadas se aprecia un número no desdeñable de ítems con pequeños errores tipográficos, ortográficos y de expresión que, aunque pueden ser

salvados sin dificultad por el lector en la mayoría de los casos –por lo que no los computamos como incumplimientos en la tabla 2-, muestran una insuficiente tarea de revisión de las pruebas previas a la aplicación y que por tanto podrían ser corregidos sin mayor dificultad. Igual ocurre con algunos ítems en los que algunas de las opciones no concuerdan en sintaxis con el enunciado. También aparecen otros dos aspectos que pueden introducir innecesarias dificultades al lector: la repetición de un mismo contenido en todas las opciones y una falta de orden en ellas que obligaría a una tarea irrelevante para el objetivo del ítem, como es ordenarlas u organizarlas para una mejor comprensión de las mismas. En resumen, contabilizamos un 22.6% y 18.7% de ítems que presentan aspectos que añaden dificultades a la expresión del dominio de interés y por tanto incumplen una o más directrices de los apartados B y C.1. de la Tabla 1.

Por otra parte, aunque una mayoría de los ítems presentan conjuntos de opciones de contenidos homogéneos, se encuentran diversos problemas que pueden distorsionar esa condición y romper la debida plausibilidad de todas las opciones, teniendo la consecuencia de inducir la respuesta correcta de modo directo o facilitar la exclusión indebida de una o más de las alternativas. A ello habría que añadir los ítems que destacan indebidamente alguna de sus opciones bien por contenido o por expresión, rompiendo la homogeneidad debida y pudiendo crear dudas innecesarias a la persona que ha de responder. En resumen, encontramos entre un 12.2% y un 11.4% de ítems que presentan al menos uno de estos problemas, referidos a las directrices agrupadas en el apartado C.2. de la Tabla 1, lo que supone una amenaza no desdeñable para la validez de los resultados que tal prueba aporte.

En suma, el presente estudio muestra que la construcción de los ítems y pruebas PIR analizadas puede ser mejorada en diversos aspectos, lo que redundaría en resultados más válidos y por tanto más justos con las aptitudes de los examinados y con el esfuerzo de los constructores. Las directrices planteadas en la literatura así lo permiten y no hay razón para no aprovecharlas en mayor medida.

Recordar finalmente que al no disponer de las respuestas empíricas de los aspirantes a las pruebas, el análisis de los ítems que hemos realizado se ha centrado en los aspectos formales y de contenido de los ítems, siendo lo ideal en todo caso haber podido complementar este enfoque con el análisis psicométrico de las respuestas. En este sentido, cabe preguntarse si para la construcción y análisis de las pruebas PIR se sigue utilizando la tecnología psicométrica clásica, o se están incorporando los potentes desarrollos psicométricos de los últimos años (Abad, Olea, Ponsoda y García, 2011; Bartram y Hambleton, 2006; Downing y Haladyna, 2006; Drasgow, Luecht y Bennett, 2006; Muñiz et al., 2005; Schmeiser y Welch, 2006; Wilson, 2005). Por citar un solo ejemplo, cabe plantearse si existe algún control sobre la equiparación de la dificultad de las pruebas de unos años y otros, algo sencillo de realizar si se utilizasen los modelos psicométricos de Teoría de Respuesta a los ítems y un amplio Banco de ítems. Si no fuera así, chocaría el planteamiento no actualizado desde el punto de vista psicométrico del Ministerio de Sanidad Política Social e Igualdad, responsable de las pruebas PIR, en comparación con las sofisticadas Evaluaciones Diagnósticas Educativas llevadas a cabo por el Ministerio de Educación y las Consejerías de Educación de las distintas Comunidades Autónomas (Fernández y Muñiz, 2011). Y no se trata de usar tecnología psicométrica avanzada por el mero hecho de usarla, sino de aprovechar los adelantos de los que se disponen para evaluar de forma más justa y rigurosa las respuestas de las personas, tal como contemplan las más elementales normas éticas y deontológicas.

NOTA

Queremos agradecer las aclaraciones que sobre el contenido de determinados ítems nos aportaron los siguientes profesores de la Universidad de Sevilla: Francisco Javier Cano, Estrella Díaz, Miguel Ángel Garrido, Montserrat Gómez de Terreros, Mª Dolores Lanzarote, Manuel Portavella y Juan Francisco Rodríguez. Muchas gracias también a Concepción Fernández Rodríguez, Profesora de la Universidad de Oviedo, cuyas sugerencias nos fueron de gran utilidad. Por supuesto, los fallos del trabajo sólo son imputables a sus autores.

REFERENCIAS

Abad, F. J., Olea, J. y Ponsoda, V. (2001). Analysis of the optimum number of alternatives from the Item Response Theory. *Psicothema, 13* (1), 152-158.

Abad, F. J., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Asociación Nacional de Psicólogos Clínicos y Residentes, ANPIR. (2005). *Cómo preparar el PIR*. http://www.anpir.org/modules/news/article.php?storyid=9 (descargado el 15 de Febrero de 2011).

Bartram, D. y Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet.* Chichester, UK: Wiley and Sons.

Boletín Oficial del Estado (1989). Orden 14882. *BOE número 153,* de 28 de junio de 1989, pp. 20164 a 20167.

Boletín Oficial del Estado (2010). Orden SAS/2448/2010. *BOE número* 230, Sec. II. B, de 22 de septiembre de 2010, pp. 80254-80449.

Bruno, J. E. y Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55* (6), 959-966.

Delgado, A. R. y Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14* (3), 197-201.

Downing, S. M. y Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M. y Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Fernández, R. y Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula Abierta, 39(2),* 3-34.

Haladyna, T. M., Downing, S. M., y Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15* (3), 309-334.

Martínez, R., Moreno, R., Martín, I. y Trigo, E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21 (2),* 326-330.

Ministerio de Sanidad, Política Social e Igualdad. (2011). *Formación Sanitaria Especializada.*http://sis.msps.es/fse/PaginasDinamicas/Consulta_Cuadernos/ConsultaCuadernosDin.aspx?MenuId=CE-00&SubMenuId=CE-01&cDocum=32

Moreno, R., Martínez, R. y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2,* 65-72.

Muñiz, J. y Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12,* 206-219.

Muñiz, J., Fernández Hermida, J. R., Fonseca, E., Campillo, A., y Peña, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo, 32* (2), 113-128.

Muñiz, J., Fidalgo, A. M., García-Cueto, E., Martínez, R. y Moreno, R. (2005). *Análisis de los ítems.* Madrid: La Muralla.

Muñiz, J. y García-Mendoza, A. (2002). La construcción de ítems de elección múltiple. *Metodología de las Ciencias del Comportamiento, Especial,* 416-422.

Muñiz, J. y Hambleton, R. K. (2000). Adaptación de los tests de unas culturas a otras. *Metodología de las Ciencias del Comportamiento, 2,* 129-149.

Muñiz, J, Prieto, G., Almeida, L. y Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15,* 151-157.

Rogers, W. T. y Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59* (2), 234-247.

Schmeiser, C. B. y Welch, C. (2006). Test development. En R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 307-353). Westport, CT: American Council on Education/Praeger.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.