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

QUALITY OF THE CLINICAL PSYCHOLOGY INTERNSHIP (PIR) EXAMINATION ITEMS

Rafael Moreno[1], Rafael J. Martínez[1] and José Muñiz[2]
[1]*University of Seville.* [2]*University of Oviedo*

This article explores the quality of the test items used in Spain for the selection of candidates for the four-year Clinical Psychology Internship (Psicólogo Interno Residente, PIR). Completion of this internship is necessary for obtaining the Specialist in Clinical Psychology qualification. 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 the literature provides for the systematic construction of items and tests. The most noteworthy results of the exploration carried out can be summarized as follows. Despite a general compliance with several of the guidelines, there is inadequate specification of the content and skills to be assessed. Furthermore, over sixteen per cent of the items present formal or content errors that hinder the exposition of the domain of interest, and this suggests insufficient editorial review of the test prior to its administration. In addition, around twelve per cent of the items lead respondents to the correct response either directly, or indirectly by allowing them to discard one or more of the alternatives. In view of the above, there is clearly a need for greater rigour in the construction of test items for selecting future Clinical Psychologists in Spain.
Key words: *Item construction, PIR Tests, Guidelines, Spain.*

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 items, Exámenes PIR, Directrices, España.*

The validity of assessment tests basically depends on their construction and use, which is why substantial efforts have been made, at an international level, for the improvement of both aspects (Muñiz & Bartram, 2007; Muñiz, Fernández-Hermida, Fonseca, Campillo, & Peña, 2011; Muñiz & Hambleton, 2000; Muñiz, Prieto, Almeida, & Bartram, 1999). Focusing on their construction, a systematic approach in accordance with clear and efficient rules is an essential condition for obtaining satisfactory and defensible results – particularly if those results are socially or economically significant for the population assessed, such effects being multiplied in line with the size of the population in question.

In the Spanish context, one of the assessments with such characteristics is the annual exam for the selection of candidates for the Clinical Psychology Internship (*Psicólogo Interno Residente*, PIR) throughout the network of public hospitals and health centres – a course that leads to qualification as a Specialist in Clinical Psychology. Given the significance of the results of this exam, it would seem important to examine the quality of such exams and their items. The only material available for this purpose consists of the exams themselves, since the authorities responsible do not publish either the individual results or the psychometric properties of the tests. Analysis of the

Correspondence: 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

correctness and defects of these exams and their items can contribute relevant information.

Muñiz and García Mendoza (2002) carried out a first exploration of the PIR exams, identifying some errors. The present work sets out to continue this line of work, examining the extent to which the items and tests are in line with the guidelines provided in the literature for their systematic construction. We analyze an intentional sample of exams which have already been applied and published by Spain's Ministry of Health, Social Policy and Equality, using as an analytical criterion a recent version of the guidelines on the construction of items and tests, summarized in Table 1 (Moreno, Martínez, & Muñiz, 2006).

METHOD

Material for analysis

Although the PIR exams have been organized at a national level since 1993, for the purposes of the present study we consider as the population of such exams those applied between 2001 and 2008, since these are the ones which, at the time of carrying out our research, had been published on the website of the Ministry of Health, Social Policy and Equality (2011), the body responsible for their administration. This website also publishes the answers considered as correct for each item. In our study we selected the PIR exams from 2005 and 2008, as an intermediate example and the latest edition, respectively,

TABLE 1
GUIDELINES FOR MULTIPLE-CHOICE ITEMS (MORENO, MARTÍNEZ, & MUÑIZ, 2006)

A. On basic principles

1. To ensure the validity of the items and exams, the objective and domain of the assessment should be defined in as much detail as possible.
2. It is also advisable to specify the context in which the items will be used, which includes the population addressed and the circumstances in which they will be applied.

B. On the expression of the domain and the context in each item and exam

3. The objective, domain and context of interest should be the definitive criteria for item construction. Each item should cover a significant aspect or unit of the domain, and form together with the other items a relevant examination.
4. Each item should clearly show the content in question. Both the syntax and the semantics should fit with those of the domain and context of reference, without adding unnecessary difficulties.
5. Once the items have been constructed, it should be ensured that the whole set of them fits with the domain and context of reference, especially with regard to the number of items and their distribution throughout the exam.

C. On the response options

C.1. Aspects that should facilitate the expression of the domain of interest and not add unnecessary difficulties
6. Each option should be the briefest possible continuation of or response to the question statement.
7. Item construction is more efficient when there is just one correct option, and it is totally, not partially, correct. Otherwise, the applicable criteria should be clarified.
8. The spatial distribution of the options should facilitate perception of the item's content.
9. The content of each option should be independent of the rest. Therefore, the options "All of the above" and "None of the above" should be used with caution.
10. The options for each item should be appropriately ordered, so that examinees do not have to undertake this task prior to answering the question.
C.2. Aspects that should prevent direct inducement to the correct answer or undue facilitation of the exclusion of one or more of the alternatives
11. The options should be plausible for examinees who do not know the correct answer, permitting those who do know it to identify it and discard the rest. The use of content and terms close to the correct option and of common errors from examinees are appropriate means of achieving this.
12. Item designers should avoid giving clues to the correctness or incorrectness of one or more options. They should avoid the use of terms that can unduly provide information about what is stated in the question.
13. Caution should be taken to avoid item characteristics which, without being clear inducements as to the correctness or otherwise of an option, differentiate it from the rest, leading examinees to wonder whether the difference might be significant. Different length and type of content of an option are common design errors in this regard.
14. The number of options included should lend plausibility to all the options for the examinee who does not know the right answer. Three is generally adequate, though larger numbers can also be appropriate.
15. Care should be taken to ensure that the set of items as such does not contain any inappropriate clues or inducements. It is therefore advisable to review the exam design more than once in the light of the guidelines prior to its application and/or publication.

for the population considered, with 260 multiple-choice items and 5 response options for each item.

Procedure

All the items of the two selected PIR exams (2005 and 2008) were scrutinized in relation to the criteria in Table 1. With a view to refining the criteria through which to apply each one of the guidelines, we carried out a pilot study with a convenience sub-sample of items from the two exams. Two coders then analyzed, independently, the fit or lack of fit to the guidelines for all the items in the sub-sample, carrying out a correspondence test with a random sub-sample. In cases of disagreement, a guideline was deemed as failing to be met when one of the coders continued with this judgement after a joint review of the item in question. We set out to make an exhaustive review, indicating lack of fit for each guideline, though, on the other hand, we point out only a few examples of other important aspects, as additional information, including scarcely significant but improvable deviations.

RESULTS

Agreement in the categorizations by the two independent coders for each of the guidelines was calculated via the percentage of agreements and the Kappa index with correction of agreements due to chance. For this purpose we selected a simple random sample of 47 items (Measurement error = 7.8%, with a 95% confidence level; estimated proportion of agreements = 0.9). Table 2 shows the results of these analyses, together with a summary of the percentages of lack of fit for the various guidelines examined.

As regards agreement or correspondence in the coding, estimated mean percentage of agreement was 97.5%, with values ranging from a minimum of 87.2% to a maximum of 100%, this latter value being found in 7 of the 13 categorizations. Kappa index values also reflect a high degree of agreement, with values of over 0.75, but with the exception of the categorization of *Guideline 4*, referring to the syntactic and semantic correctness of the items, in which case an extremely low index was obtained (k = .044), related to the lack of agreement over the only registered case of failure to meet the criteria.

The results of the analysis are shown below (Table 2) for each of the guidelines. The objective of the exams – *Guideline 1* – is set out in the official call for applications (see, for example, *Orden SAS/2448/2010*, published in the *Boletín Oficial del Estado*, 2010); they are *selective* exams, so that successful examinees can choose among the positions offered (at different types of institution, or in different locations). On the other hand, the knowledge domain to be assessed is not made explicit. It is only mentioned in *Orden 14882* (*Boletín Oficial del Estado*, 27 June, 1989), referring exclusively to the MIR (Medicine) and FIR (Pharmacy) exams, which stipulates that the exams "*shall cover the content of the knowledge areas included in the respective degree courses*". This emphasis on the knowledge domain mastered in the degree course has been generalized to the internship exams introduced later, as is the case of Psychology, the level of those taking the exam being evaluated as the candidate's previous academic record plus his or her exam score. The website of the National Association of Clinical Psychologist and Interns (*Asociación Nacional de Psicólogos Clínicos y Residentes*, ANPIR, 2005) confirms that "*The content of the PIR exam will correspond to all the courses/units (both mandatory and optional) of the Psychology degree syllabus*", and specifies that the PIR exams cover the different academic areas of Psychology, though with more weight being given to clinical-related content, this being understood to include subjects such as Psychopathology, Therapies, Assessment, Psychodiagnosis, Personality and Differential Psychology. What is not indicated are the types of abilities – memory-based, reasoning-related, or other – required to be assessed within the domain of degree-course content.

As far as the context in which the exams take place is concerned – Guideline 2 –, the call for applications sets out in adequate fashion the relevant circumstances and details, such as starting times, duration, location, confidentiality conditions, exam booklets, answer sheets and how they should be filled out, and rules about entering and leaving the exam hall.

The calls for application do not mention whether the objective, knowledge domain and relevant context were used explicitly as criteria for the construction or choice of each item included in the exams – Guideline 3 –; hence, it was not possible to assess whether the items and exams in the sample studied are in line with these criteria – Guideline 5. It is true that the items analyzed are units of the psychology domain to which the objective refers, and also that they are suitable for the context in which the exams will take place; however, given our lack of knowledge about whether, prior to the item construction, a particular distribution of the relevant content and

abilities was specified, we were unable to assess the exams' degree of fit with these two criteria. As an alternative form of analysis, we describe the distributions of both aspects in the samples considered.

With regard to content, in both exams we find the same categories, the majority of the items being grouped in accordance with them, though the last 10 items in each exam are of diverse content (probably because they are reserve items). As can be seen in Table 3, content related to clinical aspects accounts for 51.3% in 2005 and 76.6% in 2008, with Psychopathology being the most common in the sample exams, the remainder of categories presenting percentages of over 10, with the exception of Psychodiagnosis in 2005, and Personality and Differential Psychology in both samples studied. Non-clinical content presents percentages lower than 10% of the items, with the exception of Basic Processes and Developmental and Educational Psychology, both in 2005, which account for 13.1% and 11.1%, respectively – higher proportions, indeed, than were found for some clinical aspects.

TABLE 2
PERCENTAGES OF ERRORS OR NON-COMPLIANCE WITH GUIDELINES ON ITEM CONSTRUCTION IN THE PIR EXAMS OF 2005 AND 2008 AND AGREEMENT INDICES FOR CATEGORIZATION

	2005		2008				
Guideline	**Errors**	**%**	**Errors**	**%**	**% Agreement**	**k**	ME
A. On basic principles							
1. Domain							
Content	a		a		87.2%	.855	.108
Abilities	a		a		97.9%	.879	.235
2. Context	0	0.0%	0	0.0%	100.0%	1.00	
B. On the expression of the domain and the context in each item and exam							
3. Significant aspect or unit	a		a				
4. Syntax and semantics	1	0.4%	0	0.0%	91.5%	.044	.979
5. Fit	a		a				
C. On the response options							
C.1. Aspects that should facilitate the expression of the domain of interest and not add unnecessary difficulties							
6. Continuation	47	18.1%	42	16.1%	97.9%	.911	.173
7. One correct answer only	5	1.9%	5	1.9%	100.0%	1.00	
8. Appropriately spaced	0	0.0%	0	0.0%	100.0%	1.00	
9. Independence	b		b				
10. In order	7	2.6%	2	0.7%	100.0%		
C.2. Aspects that should prevent direct inducement to the correct answer or undue facilitation of the exclusion of one or more of the alternatives							
11. Plausible	0	0.0%	3	1.1%	100.0%	1.00	
12. No clues	14	5.3%	0	0.3%	97.9%	.877	.238
13. Homogeneity	19	6.9%	24	9.2%	95.7%	.776	.303
14. Appropriate number	0	0.0%	0	0.0%	100.0%	1.00	
15. No inducement	0	0.0%	2	0.8%	100.0%	1.00	

[a] Not assessable, given the lack of explicitness about the objective for each item and for the exam as a whole, in terms of content and abilities.

[b] Failures to comply with this guideline are categorized in guidelines 6, 7 and 12.

K: Kappa coefficient

ME: Measurement error

As regards abilities, the vast majority of the items require memory-based identification (96.2% and 90.4% in each exam, respectively), so that examinees are required to recall information referred to in the question and link it correctly with one of the response options offered; for example, definition of a technical term or vice versa (such as 05/4 and 05/20, respectively), description of a concept (05/36), connection between two ideas or notions (05/179, 05/215), author of some idea or vice versa (05/1 and 05/21, respectively), or purpose of some instrument or vice versa (05/197 and 05/209) (the references in brackets refer to exam year and item, respectively; if these are followed by one or more numbers, this indicates the options. For example, 05/9/1-2 refers to Options 1 and 2 of item 9 from the 2005 exam). Another ability that is important to be tested

TABLE 3
DISTRIBUTION OF CONTENT OF THE EXAM ITEMS

	2005			2008		
Areas	**Items**	n	%	**Items**	n	%
1. Psychopathology[a]	13[b], 61, 66-72, 79, 101-118, 120, 214-235, 239, 241, 243, 254, 260	56	21.6%	1-17, 19-72, 124, 129, 208	74	28.5%
2. Therapies and treatments[a]	64, 73-74, 76-77, 142-164, 186-190, 193-195, 244-246, 251-253, 255	43	16.6%	18, 85-87, 91-93, 95-96, 98, 101-123, 125-128, 130-153	61	23.5%
3. Psychodiagnosis and Behavioural Assessment[a]	78, 80, 196-213, 236	21	8.1%	73-79, 81-83,	40	15.4%
4. Personality and Differential Psychology[a]	23-24, 31, 36-40, 191-192, 238, 240, 242	13	5.0%	177-206 84, 154-176	24	9.2%
5. Basic Processes and History	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. Psychometrics, Statistics, Methods	81-100	20	7.7%	226-234, 253, 259-260	12	4.6%
7. Social and Organizational Psychology	22, 121-141	22	8.4%	243-251, 254,	11	4.2%
9. Developmental and Educational Psychology	3, 62-63, 65, 75,				*258*	
10. Psychobiology and Psychophysiology	165-185, 257-259 19, 41-60, 237	29 22	11.1% 8.4%	80, 235-242, 255 207, 209-216, 257	10 10	3.8% 3.8%
Total		**260**			**260**	

a Content related to the clinical area.
b As regards number of items, the hyphens in this table represent intervals between questions.
n: Number of items

is that of reasoning so as to respond correctly to the item. In our view, this requires the examinee to make some comparison of a covariation between terms so as to respond correctly. In the 2005 exam we found no items of this type, while in the 2008 version there were the following eleven items: 16, 19, 22, 25, 39, 50, 54, 69, 78, 109, and 160, illustrated below through the presentation of one example.

08/54. How can we distinguish a dementia condition from an amnesiac syndrome in a patient who continually complains about his or her memory?:

1. By the person's age.
2. By the presence of retrograde amnesia.
3. By the conservation of working memory.
4. By the presence of global cognitive decline that progresses as the disorder advances.
5. By the presence of anterograde amnesia.

In any case, it should be noted that responding to these items could be memory-based (rather than reasoning-based) if the content involved can be found in the psychological material studied by all or some of the examinees, so that there is no need for them to consider the covariation. This occurs, for example, in items 05/32 and 05/45. Therefore, the 11 (4.2%) items mentioned would be the maximum number requiring reasoning as well as memory

Other items require the application of knowledge to the resolution of practical cases or examples of some psychological concept or characteristic. We have identified the following items of this type: 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 and 05/228, and 08/28, 08/31, 08/89, 08/94, 08/96, 08/98, 08/107, 08/126, 08/249 and 08/250 – that is, 14 and 10 items in the two sample exams, 3.8% and 5.3%, respectively. In sum, the large proportion of items requiring memory ability could be considered a bias if we understand that the profile of the qualified Psychologist to be assessed should include not only this type of ability, but also at least the other two considered here.

According to our analysis, the items do not seriously hinder the understanding of the text due to problems of syntax, semantics or clarity of the expressions used – *Guideline 4*. One exception is item 05/180, (0.4% of the total), which was indeed probably removed, since it appears without a response in the published version. Nevertheless, there are questions that contain elements or errors that may distract examinees from their task, and which could have been corrected if the exams, prior to their publication, had been more carefully checked. There are some obvious instances of incorrect (or at least improvable) punctuation – such as in items 05/5, 05/100, 05/161, 05/173, 05/180, 05/189 and 05/190; furthermore, we found similar types of error, such as in 08/122 and 08/252 – missing accent – and in 08/122 – incorrect accents that change the meaning of the word (for example, "*aún*" ("still") instead of "*aun*" ("even") in items 05/7 and 05/216). There are also some unclear expressions (e.g., "*se realizan*" in 05/36), typos – "*de*" instead of "*se*" in 05/76 –, different formats in references to quotations – sometimes in italics, such as in 05/104 and 05/112, and others in normal font, such as in 05/218 and 05/221 –, use of terms from other languages without the use of italics (05/250, 08/44), or anglicisms (e.g., "*similaridades*" in 08/166). Finally, readers are addressed with different forms in the same exam, some items using the "usted" (polite form) – 05/100, 05/224 – and others the "tú" (informal form) – 05/193, 05/194.

As regards the response options, we examined first of all the appropriateness of the syntactic continuity or concordance between the options and the question, how far they avoided unnecessary repetition of the terms used in the questions, and their brevity – *Guideline 6*. We found that some items involve a question whose answer is among the response options, while others begin a sentence or phrase for completion from the options. In either case, the majority of the items meet the requirements of this guideline, even if there are exceptions.

The items listed below include one or more options that are syntactically incoherent with the statement they are supposed to complete: 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 and 05/255/2-4-5 in the first exam analyzed, and 08/181/5, 08/195/1-2-3-4-5, 08/203/3-4, 08/244/5 and 08/248/4 in the second; respectively 10 and 5 items, 3.8% and 1.9% of the total in each exam.

We found a considerable number of items in which there was unnecessary repetition of content in the options, instead of merely stating it in the question. In the 2005 exam, this occurs in the cases of items 9, 16, 27, 80, 109, 115, 121, 132, 139, 142, 159, 177, 178, 185, 188, 189, 196, 203, 228, 234, 236, 240, 246, 249 and 257, and in 2008, in those of items 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 and 239; that is, 25 and 29 items in the first and second exams, 9.6% and 11.1%, respectively, that fail to meet the requirements of this Guideline 6. This obliges examinees to read more, taking up valuable time which they could otherwise use for concentrating on obtaining the best possible result in the exam as a whole.

Another problem with the exams analyzed – though a less common one – involves items in which at least one of the options is excessively long, with more than double the number of words than the question statement itself. These are items 26, 30, 35, 40, 79, 125, 126, 147, 210, 220, 248 and 250 in the 2005 exam, and 33, 53, 127, 166, 227, 241 and 251 in the 2008 version, that is, 12 and 7 items (4.6% and 2.7%), respectively. In any case, some of these options are quite clear enough, calling into question the need for a criterion about their lack of brevity. In spite of this, though, we do consider that such excessive wordiness can force examinees to read text that could often be made much simpler. Moreover, it is also a fact that some of the briefer question statements are too succinct, and could be made clearer with a few more words. Attention to this aspect could help the construction of more appropriate items. In sum, taking into account the three aspects mentioned, we found totals of 47 and 42 items in the first and second exams, respectively, that were out of line with Guideline 6 (18.1% and 16.1%).

Information on whether there is a single correct answer per item –*Guideline 7* – is not made sufficiently explicit in the call for applications. And in the instructions given out with the exam itself, such information is provided only in indirect fashion through the use of the singular: "*Check that the option you mark on the answer sheet corresponds to the exam question number*". This implied criterion is confirmed when one observes on the marking sheets of the exams analyzed that each assessed item appears with a single correct answer – though of course the examinees do not have access to such information at the time of the exam.

The analysis of such answers based, in the cases of some items, on the opinion of experts in the respective content, reveals that, indeed, the majority of the items have one, and only one, correct answer. However, the following items have more than one correct answer: 05/9, 05/106, 05/180, 08/68, 08/189, 08/214 and 08/244, and in fact all of these were removed. In addition, another two items which were not removed, 08/178 and 08/182, could present the same problem in the opinion of some of the experts consulted. For their part, items 05/77 and 05/194 were removed, probably because none of their options were clearly correct. In sum, at least some 5 items (1.9%) in each exam failed to meet the criteria of this Guideline 7.

The layout of the options – *Guideline 8* – is vertical for all the items, which appear in two columns, in clear and sufficiently large letters, and appropriately spaced throughout the exam as a whole and on each page. At least this is the case in the version that appears on the website – though it is not clear whether the format was the same in the version received by the examinees. An aspect that could cause slight difficulty for reading is the division of some items between two different pages or columns. In the exams analyzed, this occurs on 20 of the 24 pages from 2005 and 22 of the 26 from 2008.

Lack of independence between the response options – *Guideline 9* – may occur because the content of one option is part of another option, or because the two are similar. Since overlaps are more easily appreciated by experts in the item content, we consulted them, though only for those items about which we had doubts; there is a possibility, therefore, that we have failed to detect some other cases of failure to follow this guideline. We might mention, in any case, the items already referred to in relation to Guideline 7 with more than one right answer, as well as 05/39, 08/188 and 08/229, with overlap between incorrect options, identified as inappropriate in Guideline 12. Another form of violating the rule of independence of the different options is the answer "None of the above" or "All of the above" (Martínez, Moreno, Martín, & Trigo, 2009). Although these do not appear as such in the exams analyzed, some items include options that could be considered versions of "None of the above": those with content that denies the truth of what is written in the question statement, and thus in the rest of the options. In the list of items that failed to meet Guideline 6 due to including options that do not follow syntactically from the statements, the following do so in the way we have just described: 05/83/5, 05/95/5, 05/214/4, 05/215/5, 05/255/5, 08/143/5 and 08/181/5. In sum, the analysis of non-accordance with this guideline overlaps with those carried out on Guidelines 6, 7 and 12, so that it is inappropriate to duplicate the listing of such errors; moreover, it may be wise to reconsider the pertinence of the current version of this Guideline 9.

As regards the order of the options for each item – *Guideline 10* –, we do not know whether in the different versions mentioned in Instruction 1 of the 2008 exam it is always the same, so that the following results correspond only to the published version. The options for items 05/83, 05/92, 05/94, 05/96, 08/86 and 08/201 are unnecessarily disordered, and this brings with it some difficulty or additional work that is counterproductive and irrelevant to the content in question. This difficulty is all the greater when the content of the options is varied, so that many examinees may well have to put the content of the options in order before responding. This occurs in item 05/47 (below), as it does in items 05/185 and 05/238.

05/47. The lemnisci that appear in a transverse section of the mid-brain at the level of the superior colliculi are:

1. Medial, spinal and trigeminal.
2. Lateral, medial and trigeminal.
3. Medial, lateral, trigeminal and spinal.
4. Lateral, spinal and trigeminal.
5. Trigeminal and spinal.

In sum, 7 items from the 2005 exam and 2 from that of 2008 fail to comply with this guideline (2.6% and 0.7%). Furthermore, there are other items which, while strictly speaking complying with the guideline, are difficult to read. Their options present two or three different contents – "proportion" and "reduction" in the example item shown below – which could be used to group them and make the item easier to read.

72. Body Mass Index is:

1. The proportion between one's height and the square of one's weight.
2. The reduction in one's weight in the last six weeks.
3. The proportion between one's height and one's weight.
4. The reduction in body fat as a function of one's weight.
5. The proportion between one's weight and the square of one's height.

The vast majority of the items present plausible sets of options, forming an appropriate framework for the correct one – *Guideline 11*. Most of them do so using content belonging to the same thematic field, in which therefore the recommended use of common errors in examinees is more likely. We found just 3 items (1.1%) – 08/32/4, 08/160/2 and 08/231/4 – with options that were so easily discardable as to be implausible.

Whilst in the guideline mentioned above plausibility is considered in terms of technical language, we should also include as inappropriate the use of "ordinary" language that can provide clues for examinees who have insufficient knowledge for choosing or discarding certain options – *Guideline 12*. In any case, the boundary between the two types of language is sometimes blurred, especially when the technical language involved is easy to understand or in common use. The correct option for the following item, number 4, illustrates this.

05/4. ¿Which term refers to the phenomenon of learning two languages simultaneously from birth (during the initial phases of language acquisition)?:

1. Additive bilingualism.
2. Subtractive bilingualism.
3. Second-language acquisition.
4. Native bilingualism.
5. Linguistic attrition.

We found various instances of failure to comply with this guideline. Items 05/6/5, 05/10/1, 05/14/1, 05/20/2 and 05/170/3 steer respondents towards the right answer on including in the initial statement itself the term used in the correct answer (or one that is quite similar). Items 05/89/3, 05/160/3, 05/255/2-4 and 08/229/1-2-5 provide clues through the accordance (or its absence) in grammatical gender or number between the statement and the correct option. Other items allow the exclusion of some options as incorrect: 05/6/2-3 and 05/39/4-5 because they include two options with similar content – so that they can be discarded, since each item has just one right answer; 05/22/3, 05/170/1 and 05/188/5 because they include one or more options with content that is incompatible with the question statement; and 05/90/1-2 because it uses the terms "always" and "never", which rarely tend to be correct. In sum, we detected 14 instances of failure to meet the criteria of this guideline (5.3%) in the 2005 exam and one (0.3%) in the 2008 exam. Such failures mean that the number of relevant options is reduced in the items in question, distorting the principle of randomness involving one correct option and four distractors – and this puts at a disadvantage the best-prepared candidates and unduly benefits those who do not know the right answer.

On analyzing the homogeneity of the options in terms of their content and length – *Guideline 13* – we identified few items that failed to comply with the criteria. Some give inappropriate hints at the correct option because the content is different from that of the other answers – 05/181/4, 05/251/2 – or the option is more detailed – 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 and 08/240/2. These shortcomings account for a total of 8 items in 2005 and 7 in 2008 (3.1% and 2.7%).

Other items give undue prominence to one of the incorrect options: Items 05/83/05, 05/94/5, 05/147/5, 05/210/2, 05/214/2, 05/215/1 and 08/173/3 do so with regard to length – which we have noted as being the case when the option in question has at least a third more words than one or more of the rest; others do so by using content that is different from the rest, which introduces an unnecessary distortive element that could be avoided by bringing the option in line with the rest, or vice versa. This would be the case of items 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 and 08/260/4, a total of 10 and 17 items, respectively, that draw attention to the incorrect option in the two exams analyzed (3.8% and 6.5%), which gives grand totals of 18 and 24 (6.9% and 9.2%) items that fail to comply with Guideline 13 in one way or another.

As regards the number of options in each item – *Guideline 14* – all those included in the two exams analyzed have five. This means it must be viable and relevant to present this quantity of plausible alternatives in all the content involved, which is not always easy – particularly when the population of content to be covered is so broad as to include all the fields of Psychology. This may explain the items pointed out in Guideline 12 as having trivial or repeated options, and those which in Guideline 13 present an incorrect option that is different from the rest. In addition, there are the items mentioned in Guideline 9, which use in some of their options the formula "None of the above". In sum, there are a large number of items presenting problems of greater or lesser importance that justify the recommendation in the literature to reduce the number of options used (Abad, Olea, & Ponsoda, 2001; Bruno & Dirkzwager, 1995; Delgado & Prieto, 1998; Haladyna, Downing, & Rodriguez, 2002; Rogers & Harley, 1999); this, of course, goes against the initial preference for a large number of alternatives as a way of reducing the influence of random responses, but special care should be taken in in any case to ensure the plausibility and homogeneity of all the options.

With regard to the avoidance of leading examinees to the right answer via information provided in other questions from the same exam – *Guideline 15* –, we found just two cases of failure to comply with this guideline, both of them in 2008: the right answer to item 194 is literally written in the question statement for item 155, while items 115 and 116 interchange the content of the statement and the correct response – option 4 in both cases.

CONCLUSIONS

The study carried out provided us with a description of the selected samples of PIR exams and items. First of all, it can be said that the objective of the exams and the context in which they are applied are clearly set out in the call for applications, whilst the way in which the knowledge domain to be assessed is indicated is insufficiently explicit and also only partial: insufficiently explicit because one has to go back to a regulation from 1989 to find clear information stating that the content to be assessed corresponds to the Psychology degree course, which also implies a certain vagueness in the delimitation of the domain (compounded by the lack of specification of the distribution of the content tested); and partial because there is no mention of the types of abilities candidates will be required to demonstrate.

Given this situation, the assessment of a hypothetical concordance of the exams with criteria that are not made explicit has to be filed under "pending". Describing, as an alternative source of information, the two exams analyzed, half and three-quarters of the content of their items, respectively, relates to clinical aspects. As regards the abilities assessed, the large proportion (a majority) of items based on memory can be considered a bias or lack of representativeness in the exams with respect to the variety of abilities potentially acquired in the Psychology degree course. It may be that this memory-based bias is generated largely by the strategy followed in order to respond clearly to possible complaints or appeals from dissatisfied examinees, so that the experts who set the exams feel obliged to construct items whose answer can be found in some textbook or other source of reference. Since this requirement is easier to fulfil, in principle, with purely memory-based items than with those that involve cognitive strategies such as inference, synthesis or application, the numbers of such cognitive strategies may be limited for this reason. However, well-constructed items involving more than the pure exercise of memory that appear in recent exams for both PIR and MIR (Medical

Internship) serve to demonstrate the possibility of correcting the bias we have mentioned.

As regards the items, all of them present five response options with just one that is correct; their spatial distribution makes them easy to read; and for the most part they are adequately expressed from a syntactic and semantic point of view. However, in the exams analyzed there was a significant number of items with small typographical errors, spelling mistakes and faulty expression which, though in the majority of cases did not unduly hinder the examinee's understanding – so that we do not compute them as compliance failures in Table 2 – , do suggest insufficient checking of the exams prior to their application, and this could easily be remedied. Furthermore, there are items in which some of the options' syntax does not agree with that of the question statement. In addition, there are two more aspects that could make things unnecessarily difficult for examinees: repetition of the same content in all the options, and a lack of order in them that obliges examinees to take on a task – organizing them so that they can be better understood – which is irrelevant to the item's objective. In sum, we counted 22.6% and 18.7% of items involving aspects that burden examinees with difficulties other than showing their mastery of the domain in question, and hence fail to comply with one or more of the guidelines in sections B and C.1. of Table 1.

It should also be mentioned that although the majority of the items present sets of options with homogeneous content, we found a range of problems that could distort this condition and affect the plausibility of all the options, the consequence being direct inducement to the correct answer or undue facilitation of the exclusion of one or more of the alternatives. In addition, there are items that inappropriately highlight some options, through their content or through their form of expression, breaking the homogeneity and giving rise to unnecessary doubts in the respondent. In sum, we found between 12.2% and 11.4% of items that presented at least one of these problems, covered by the guidelines in section C.2. of Table 1, and this represents a far from negligible threat to the validity of the exam results.

On the whole, the present study shows that the construction of PIR items and exams analyzed could be improved across a range of aspects, and this would produce results that were more valid, and indeed fairer in relation to both the aptitudes of the examinees and the efforts of the exam designers. The guidelines proposed in the literature make this possible, and there is no reason to not take greater advantage of them.

Finally, it should be borne in mind that on lacking access to the actual responses from examinees, our analysis has revolved around formal and content-based aspects of the items, though we would ideally have been able to complement this approach with a psychometric analysis of the responses. In this regard, it is pertinent to speculate on whether for the construction and analysis of the PIR exams their designers continue to employ classical psychometric technology, or whether they are incorporating the powerful psychometric techniques developed in recent years (Abad, Olea, Ponsoda, & García, 2011; Bartram & Hambleton, 2006; Downing & Haladyna, 2006; Drasgow, Luecht, & Bennett, 2006; Muñiz et al., 2005; Schmeiser & Welch, 2006; Wilson, 2005). To cite just one example, it would be relevant to know whether there is any type of control on the equivalence of exams from different years as regards their degree of difficulty – a simple task if one uses the psychometric models of Item Response Theory and a substantial item bank. If this is not the case, it would be surprising that the authority responsible for the PIR exams, the Ministry of Health, Social Policy and Equality, were using an outdated approach – from a psychometric point of view –, compared to the sophistication of the Educational Diagnostic Assessments carried out by the Ministry of Education and the Education Councils of Spain's various Autonomous Regions (Fernández & Muñiz, 2011). And it is not a case of using advanced psychometric technology just for the sake of it, but rather of taking advantage of the new approaches so as to assess examinees' responses in fairer and more rigorous fashion, as universally recommended in ethical and deontological guidelines.

NOTE

We should like to thank the following staff from the University of Seville for their clarifications with regard to the content of certain items: Francisco Javier Cano, Estrella Díaz, Miguel Ángel Garrido, Montserrat Gómez de Terreros, Mª Dolores Lanzarote, Manuel Portavella and Juan Francisco Rodríguez. Many thanks also to Concepción Fernández Rodríguez from the University of Oviedo, whose suggestions were extremely useful. And of course, any errors in the present work are attributable to its authors.

REFERENCES

Abad, F. J., Olea, J., & 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., & García, C. (2011). *Medición en ciencias sociales y de la salud* [*Measurement in social and health sciences*]. Madrid: Síntesis.

Asociación Nacional de Psicólogos Clínicos y Residentes, ANPIR. (2005). *Cómo preparar el PIR* [*How to prepare for the PIR*] http://www.anpir.org/modules/news/article.php?storyid=9 (retrieved 15th February 2011).

Bartram, D., & 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., & 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., & Prieto, G. (1998). Further evidence favouring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14* (3), 197-201.

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

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

Fernández, R., & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas [Design of booklets for the assessment of basic abilities]. *Aula Abierta, 39*(2), 3-34.

Haladyna, T. M., Downing, S. M., & 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., & 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* [*Specialist Healthcare Training*] .http://sis.msps.es/fse/PaginasDinamicas/Consulta_Cuadernos/ConsultaCuadernosDin.aspx?MenuId=CE-00&SubMenuId=CE-01&cDocum=32

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

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

Muñiz, J., Fernández Hermida, J. R., Fonseca, E., Campillo, A., & Peña, E. (2011). Evaluación de tests editados en España [Assessment of tests published in Spain]. *Papeles del Psicólogo, 32* (2), 113-128.

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

Muñiz, J., & García-Mendoza, A. (2002). La construcción de ítems de elección múltiple [The construction of multiple-choice items]. *Metodología de las Ciencias del Comportamiento, Especial*, 416-422.

Muñiz, J., & Hambleton, R. K. (2000). Adaptación de los tests de unas culturas a otras [Adapting tests from one culture to another]. *Metodología de las Ciencias del Comportamiento, 2*, 129-149.

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

Rogers, W. T., & 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., & Welch, C. (2006). Test development. In 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.