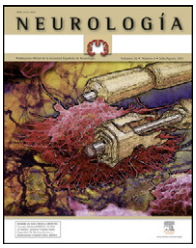
Neurología. 2011;26(8):474—480

NEUROLOGÍA

www.elsevier.es/neurologia



ORIGINAL ARTICLE

Differential functioning of mini-mental test items according to disease☆

G. Prieto, A.R. Delgado*, M.V. Perea, V. Ladera

Departamento de Psicología Básica, Psicobiología y Metodología, Universidad de Salamanca, Salamanca, Spain

Received 27 July 2010; accepted 7 January 2011

KEYWORDS
Screening test;
Mini-mental State Examination;
Testing;
DIF;
Parkinson's disease;
Alzheimer's type dementia

Abstract
Introduction: Comparing the height of males and females would be impossible if the measuring device did not have the same properties for both populations. In a similar way, the cognitive level of diverse groups of patients should not be compared if the test has different measurement properties for these groups. Lack of Differential Item Functioning (DIF) is a condition for measurement invariance between populations.
Material and methods: The most internationally used screening test for dementia, the MMSE (or Mini-mental State Examination), has been analysed using an advanced psychometric technique, the Rasch Model. The objective was to determine the invariance of mini-mental measurements from diverse groups: Parkinson's disease patients, Alzheimer's type dementia and normal subjects. The hypothesis was that the scores would not show DIF against any of these groups. The total sample was composed of 400 subjects.
Results: Significant differences between groups were found. However, the quantitative comparison only makes sense if no evidence against measurement invariance was found: given the kind of items showing DIF against Parkinson's disease patients, the MMSE seems to underestimate the cognitive level of these patients.
Conclusions: Despite the extended use of this test, 11 items out of 30 show DIF and consequently score comparisons between groups are not justified.

PALABRAS CLAVE
Prueba de cribado;
Test Mini-mental;
Psicometría;
Funcionamiento diferencial de los ítems;

Funcionamiento diferencial de los ítems del test Mini-mental en función de la patología

Resumen
Introducción: Sería imposible comparar la estatura de los varones y las mujeres si el metro no tuviese las mismas propiedades en ambas poblaciones. De forma similar, no se debería comparar el deterioro cognitivo de sujetos con distintas patologías si el test empleado no tuviese las mismas propiedades métricas en los grupos analizados. La ausencia de funcionamiento diferencial de los ítems (DIF) es una condición de la invarianza métrica entre poblaciones.

☆ Please cite this article as: Prieto G, et al. Funcionamiento diferencial de los ítems del test Mini-mental en función de la patología. Neurología. 2011;26:474—80.

* Corresponding author.
E-mail address: adelgado@usal.es (A.R. Delgado).

2173-5808/$ - see front matter

Enfermedad de Parkinson;
Demencia tipo Alzheimer

Material y métodos: Este artículo analiza el test de cribado de la demencia más utilizado internacionalmente, el Mini-mental State Examination (MMSE), mediante un modelo psicométrico avanzado, el modelo de Rasch, con el objetivo de poner a prueba la invarianza de las medidas obtenidas en distintos grupos: pacientes con enfermedad de Parkinson, pacientes con demencia tipo Alzheimer y sujetos normales. Para ello, se ha contrastado la hipótesis de que el MMSE no muestra DIF contra ninguno de estos grupos en una muestra total de 400 sujetos.
Resultados: Los resultados del análisis indican que existen diferencias significativas entre los grupos; sin embargo, la comparación cuantitativa sólo tiene sentido si no existe evidencia en contra de la invarianza métrica: dado el tipo de ítems que muestran DIF contra el grupo de pacientes con enfermedad de Parkinson, cabe pensar que el MMSE podría estar sobreestimando el nivel de deterioro cognitivo de estos pacientes.
Conclusiones: Pese a lo extendido del uso del MMSE, el funcionamiento de 11 de sus 30 ítems no es igual para los distintos grupos por lo que la comparación de las puntuaciones no estaría justificada.

Introduction

The Mini-Mental State Examination (MMSE) is the most common internationally used screening test for dementia. The 30 test items are classified into 6 cognitive domains: temporal orientation, spatial orientation, registration of information, attention and calculation, memory and language. The original version[1] has been adapted to different countries and translated into different languages, including Spanish,[2] Portuguese,[3] Japanese,[4] Chinese,[5] Turkish,[6] Italian[7] and Hebrew.[8]

Screening tests are used in different populations. An essential condition of the metric equivalence of the scores in a test of different groups is the absence of differential item functioning (DIF). An item presents a DIF associated to membership of a group when subjects with the same value in the variable are being measured, but from different groups, have a different probability of solving the item correctly. The first procedures to detect DIF were applied to scores obtained through classical theory tests (CTT).[9] Although CTT has been the main psychometric technique used in analysing test scores, its limitations have led to other alternative techniques being proposed, of which the most parsimonious is the Rasch[10] model. This, given good adjustment, allows measuring people and items jointly in the same latent variable with interval properties. It is important to note that this model does not require representative samples. The reason is that one of its properties — specific objectivity — guarantees that, given sufficient adjustment of the data to the model, the item parameters are independent from the subject sample and the subject parameters, from the item sample. It is necessary to have a sufficient number of subjects at all levels of the latent variable. The Rasch model, statistically included in the item response theory (IRT), allows us to easily test the invariance of the measurements that a test provides in different groups, as well as having the other properties that make its use particularly recommendable.[11,12] The Rasch model is especially easy to apply in its most basic version, when it deals with analysing dichotomous data such as the responses to the 30 items in the MMSE; that is, when there is only one correct answer and the data can be binary-coded (1/0). In these cases, we can model the probability that the subject n will give a correct response to item i, p_{ni}, with the following formula:

$$p_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp[\beta_n - \delta_i]}$$

β_n, subject level n in the latent variable; β_n, value of item n in the latent variable.

The adjustment of the data obtained in applying the test to the Rasch model is calculated through statistics based on the residues Infit and Outfit, whose distribution is similar to χ^2. With regard to DIF, its existence is tested by calculating the standardised difference between the item difficulty parameter estimators in each group, controlling possible differences between groups in the latent variable and using the Bonferroni correction.[13] We usually recommend removing items with DIF due to their lack of general validity, especially when test scores will be used to take important decisions.[14] However, depending on the measurement objectives, they could also serve as a source of hypotheses about the cause of the observed difference.

The aim of this paper was to analyse MMSE through the Rasch model so as to test the invariance of the measurements obtained by the MMSE in different groups: patients with Parkinson's disease, patients with an Alzheimer-type dementia and normal subjects. To do so, we subjected to contrast the hypothesis that MMSE does not show DIF against some of the groups.

Subjects and methods

The MMSE[2] was carried out through individual interviews with 400 Spanish subjects living in the *Castilla y León* region, using Spanish as their native language; 121 had Parkinson's disease (PD group), 110 Alzheimer-type dementia (ATD group) and 169 were normal subjects (N group). The range of levels in the construct obtained by these three groups is very wide and guarantees that the estimation error did not vary significantly through the latent variable. On the other hand, we must remember that the property of specific objectivity

allows us to obtain estimators in non-representative sample parameters. The PD group was composed of 59 females and 62 males diagnosed with PD without dementia. The PD was diagnosed by a neurologist on the basis of akinesia associated with one of the other 2 main signs (tremor, rigidity) and response capacity to treatment with levodopa. All the patients complied with the criteria in the UK Parkinson's Disease Society Brain Bank for the diagnosis of idiopathic PD.[15]

The ATD group was composed of 62 females and 48 males diagnosed with ATD, according to the NINCDS-ADRDA[16] criteria, confirmed by specialists. The severity level of dementia according to clinical dementia rating (CDR)[17] was slight in 59 cases, moderate in 42 and severe in 9. Subjects with a known history or with a suspicion of transient cerebral ischemia, alcoholism, head injury or diseases such as cancer, thyroid dysfunction and depression were excluded. As for the comparison group (N), this comprised 90 females and 79 males with ages varying from 15 to 65 years old and with no type of neurological, psychopathological and/or neuropsychological disorders in their clinical history. There were 77 subjects with a perfect score ($X = 30$), who were excluded from the initial sample, as the model could not estimate their parameters and they did not provide information for the analysis. Of these, 75 were normal subjects and 2 were diagnosed with PD.

Results

To start with, the adjustment of the data to the model seemed sufficient: only 1 item and 10% of subjects are severely out of step; from these, only 1 belonged to the ATD group, 5 to the PD group and the rest were normal subjects. The internal consistency of Cronbach's alpha coefficient was 0.93 and the reliability (estimated following the Rasch model) was 0.86, an acceptable level. The reliability of the items was very high at 0.99, in terms of the model.

The parameter estimates of item difficulty can be seen in Table 1, next to the error and misfit indicators Infit and Outfit. We must point out that the only item severely out of step, number 25 (order 1), had a discrimination of 0.17. This separated it from the rest of the items, whose discrimination varied between 0.37 for the items 28 (clock) and 29 (repetition-phrase) and 0.77 from item 10 (flat). It is also noteworthy that the Outfit indicator detected dependence between several items, for example, those that formed part of the calculation function, through exceedingly low values that signal a deterministic pattern in the responses. As for the order of item difficulty, this can easily be seen in Fig. 1, where subjects and items are scaled together along the latent variable (zero in the scale conventionally corresponds to the mean difficulty of the items). At the top of this chart, we find the last 3 calculation items followed by the 3 memory ones, which are the hardest. At the bottom, starting with the easiest, we can see the so-called visual-type items ''clock'' and ''pen'', a registration item ''pencil'', a spatial orientation item ''city'', and the two remaining registration items. As expected, the items related to the so-called visual types point to the lower extreme of the variable, while the memory and calculation items point to the upper part.

The means comparison between the 3 groups in the Rasch scores indicates there are great and significant differences between the 3 groups, $F(2.397) = 654.41$, $P > .001$, whose means are, from highest to lowest, as follows: 3.56 (N), 2.20 (PD) and −0.37 (ATD).

The joint scaling of the subjects and items (Fig. 1) and the comparisons between the groups based on this scale were taken from the total sample, which implies starting from the supposition that measurement is invariable in the 3 groups. If it were not so, then neither the Rasch measurements nor the sum of the points normally used (and that is the sufficient statistic from which subject and item scores were estimated), would properly reflect the aptitude of the groups. This led to our aim to test the invariance of the measurements obtained through MMSE in patients with Parkinson's disease, patients with Alzheimer-type type dementia and normal subjects.

The results of the DIF analysis indicate that there are various items that work differentially in one or more groups. Fig. 2 shows the indicators of difficulty estimated separately for each group, controlling the differences between groups in the latent variable. Let us take the item of spatial orientation ''flat'', an item with a DIF against the ATD group: it is not that there is a difference between groups in the latent variable, which there is, it is that the quantitative comparison can only be made legitimately if you have previously rejected the DIF hypothesis between groups. That is, the item must have the same difficulty for people at the same level in the construct, whatever group they belong to. If on the contrary, the difficulty of this item for the ATD group is greater than for the other 2 groups (as seen in Fig. 2), the quantitative comparison between them would be compromised.

Table 2 summarises the results of the DIF analysis, starting with the 5 comparisons that prejudice the ATD group in 4 items, of which 3 are of spatial orientation and only 1 in calculation (in this, the comparison is with the N group and not the PD one). Below we can see the items with DIF against the PD group, 5 items and a total of 6 comparisons: (1) calculation item, the repetition of a sentence and the drawing are harder for patients with PD than those with ATD, even when they have the same level of latent variable; (2) calculation items and the drawing are harder for patients with PD than for normal subjects with the same level of latent variable. Finally, 6 comparisons focused on 4 items show DIF compared with subjects in the normal group, although 1 is severely out of step and has a low discrimination index. All the differences indicated are statistically significant after the Bonferroni correction, which corrects the level of *(* to take into account the number of comparisons undertaken, which were 90 (30 items $\times$ 3 groups) in this study. In total, 17 of the 90 comparisons shed statistically significant results, which affect 11 items. There is also a clear pattern in the type of items that prejudice each of the groups with a pathology: the spatial orientation items prejudice the ATD group more, while the calculation items and drawing ones prejudice the PD group more.

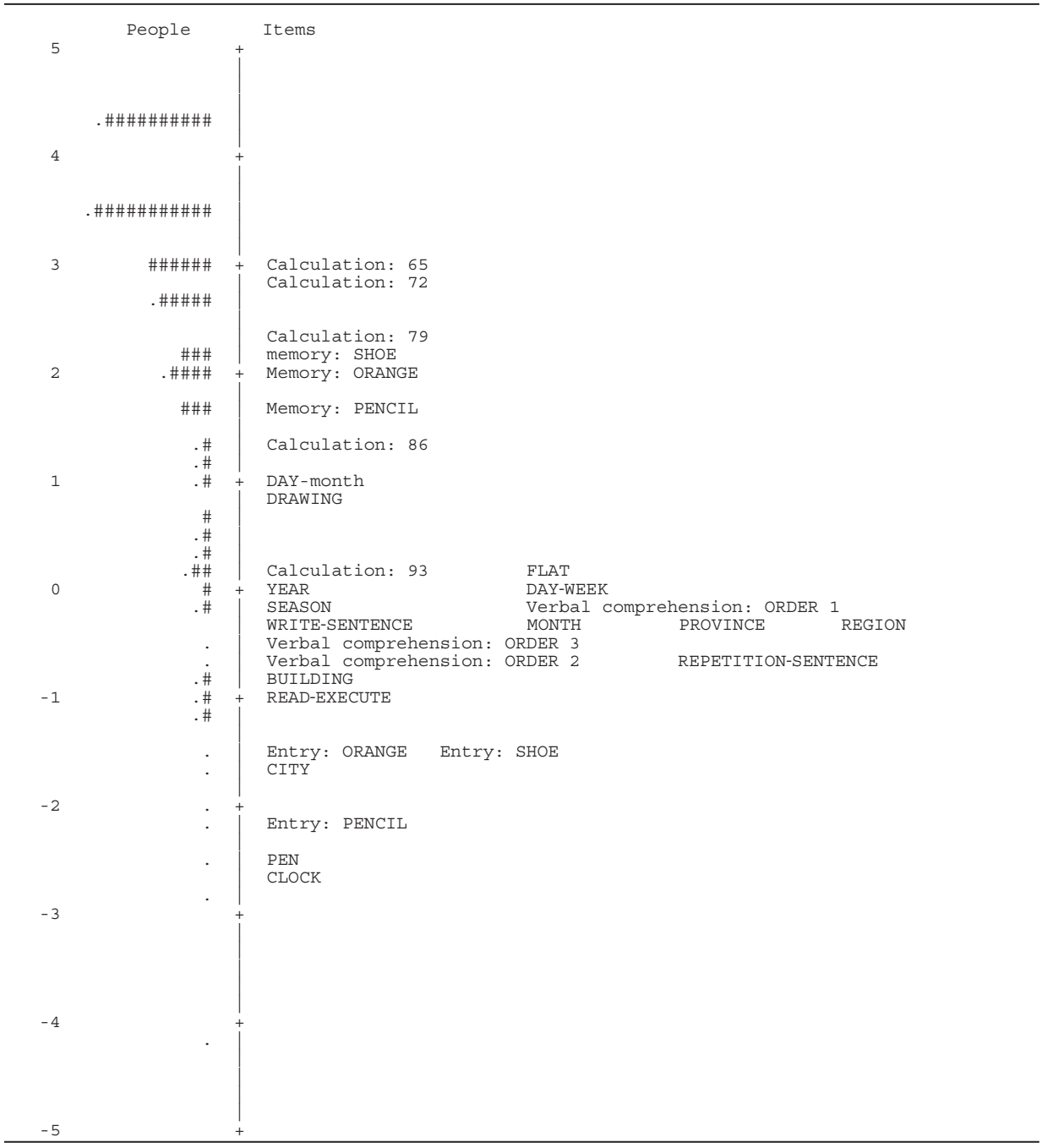
Table 1 Difficulty, estimation error and adjustment indicators of the items.

Item	Difficulty	Error	Infit	Outfit	Label
1	−0.07	0.16	0.85	0.61	Temporal orientation: year
2	−0.18	0.16	1.01	1.24	Temporal orientation: season
3	0.95	0.14	0.83	1.08	Temporal orientation: day−month
4	−0.02	0.16	0.85	0.78	Temporal orientation: day−week
5	−0.34	0.17	0.91	0.51	Temporal orientation: month
6	−0.31	0.17	0.90	1.47	Spatial orientation: region
7	−0.34	0.17	0.74	0.93	Spatial orientation: province
8	−1.65	0.21	0.89	0.35	Spatial orientation: city
9	−0.82	0.18	0.85	0.34	Spatial orientation: building
10	0.17	0.15	0.54	0.37	Spatial orientation: flat
11	−2.18	0.25	1.02	0.29	Entry: pencil
12	−1.52	0.21	0.69	0.23	Entry: orange
13	−1.52	0.21	0.69	0.24	Entry: shoe
14	0.10	0.16	1.26	1.59	Calculation: 93
15	1.31	0.14	1.15	1.06	Calculation: 86
16	2.40	0.13	0.72	0.64	Calculation: 79
17	2.86	0.13	0.56	0.41	Calculation: 72
18	2.99	0.13	0.61	0.45	Calculation: 65
19	1.60	0.13	1.35	1.60	Memory: pencil
20	2.05	0.13	1.31	1.44	Memory: orange
21	2.13	0.13	1.06	1.01	Memory: shoe
22	−2.44	0.27	0.75	0.17	Visual denomination: pen
23	−2.75	0.29	0.91	0.33	Visual denomination: clock
24	−0.66	0.17	1.45	1.89	Phrase repetition
25	−0.10	0.16	1.82	5.10	Verbal comprehension: order1
26	−0.63	0.17	1.10	1.27	Verbal comprehension: order 2
27	−0.54	0.17	1.24	2.00	Verbal comprehension: order 3
28	−1.02	0.19	1.02	0.93	Written comprehension: read-perform
29	−0.31	0.17	0.98	0.68	Writing-sentence
30	0.83	0.14	1.32	1.25	Drawing

Table 2 Differential item functioning: comparison, contrast, item, label and group prejudiced.

Comparison	*t*	Item	Label	Difficult
PD-ATD	−4.26	6	Spatial orientation: region	ATD
PD-ATD	−4.09	7	Spatial orientation: province	ATD
PD-ATD	−6.01	10	Spatial orientation: flat	ATD
N-ATD	−3.62	10	Spatial orientation: flat	ATD
N-ATD	−3.69	17	Calculation: 72	ATD
PD-ATD	6.25	15	Calculation: 86	PD
PD-ATD	4.46	24	Phrase repetition	PD
PD-ATD	5.98	30	Drawing	PD
N-PD	−4.06	17	Calculation: 72	PD
N-PD	−3.72	18	Calculation: 65	PD
N-PD	−4.76	30	Drawing	PD
N-ATD	3.94	14	Calculation: 93	N
N-ATD	3.87	19	Memory: pencil	N
N-ATD	8.56	25	Verbal comprehension: order 1	N
N-PD	4.76	6	Spatial orientation: region	N
N-PD	6.09	19	Memory: pencil	N
N-PD	7.29	25	Verbal comprehension: order 1	N

$P < .00055$ in all comparisons (α/no. of comparisons).



Nota: Each "#" represents 6 subjects; each "." from 1 to 5 subjects.

Figure 1 Map of the variable: people and items jointly scaled. Each ''#'' represents 6 subjects; each ''.'', from 1 to 5 subjects.

Discussion

Compared to CTT, all IRT models allow us to quantify the level of the items and the people in the same measurement. The results of this study show that MMSE items vary greatly as indicators of cognitive deterioration, with the items that are the hardest being calculation and memory, and the easiest registration and visual denomination. That is, an incorrect response to the latter indicates a greater degree of deterioration than in the first ones. The people studied likewise presented a high variability, with there being large significant differences between the normal, PD and ATD subjects. Although the diagnostic interest of this data seems unquestionable, we must take into account not only the differences between the items, but between the groups of people as well, which could be distorted by differential item functioning. Screening tests are used to classify people in ''pathological'' categories according to the score comparison of the person with empirically determined cut-off points. In the MMSE case, the classification procedure

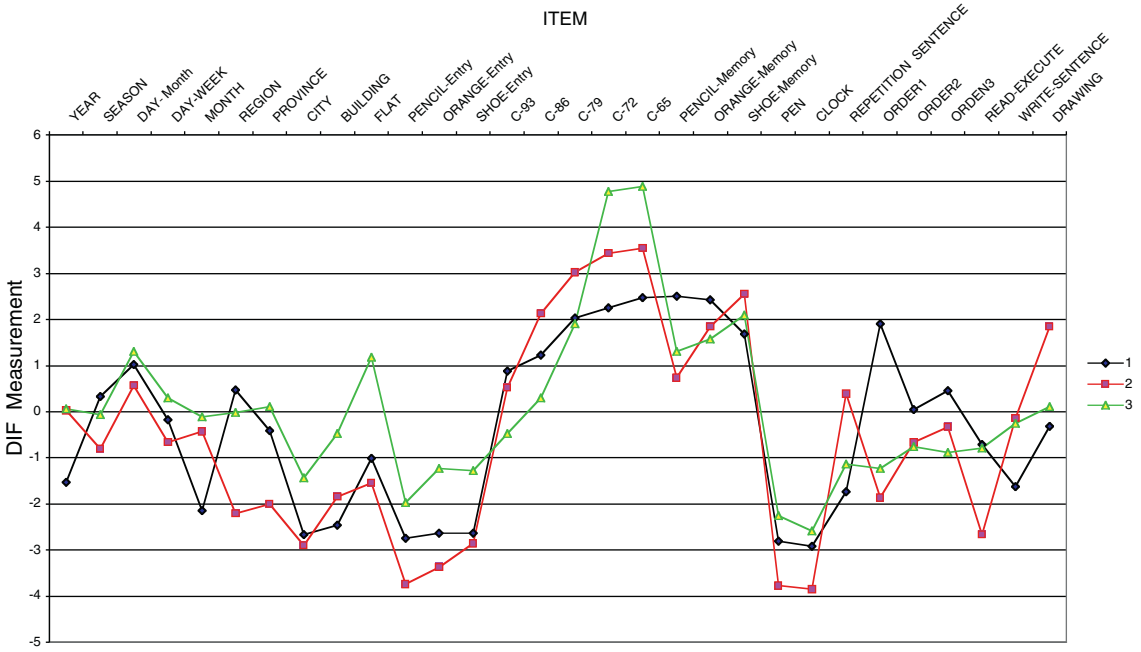


Figure 2 Item difficulty estimators for each group monitoring the level of the latent variable. The values 1, 2, and 3 correspond to N, PD and ATD, respectively.

would only have a generalised validity if the people with the same score had the same level of cognitive deterioration independently of to which of the groups affected by the different pathologies they belong. That is, the MMSE should work in a similar way (have the same metric properties) in the different groups of subjects. If not, the measurements of the different groups would not be comparable. This would be the case of comparing of the height of males and females, which would be impossible if the measuring tape did not have the same properties (it would function differentially) for both populations.

One of the most common procedures to analyse metric invariance of an instrument among populations is that used in this study: that of differential item functioning (DIF) between normal patients and those with ATD and PD. The results indicate that 11 of the 30 MMSE items present some type of DIF, which means that a certain score in the test may not be the indicator of the same level of cognitive deterioration in subjects with ATD and PD. Quantitative comparison is only appropriate if there is no evidence of metric invariance. Given the type of items that show DIF when comparing the PD group (repetition of a sentence, verbal expression of calculations and drawing), we must think of factors that limit the execution of these specific items, which do not indicate just cognitive deterioration. A plausible alternative explanation could be apathy, which has come to be recognised as one of the most relevant symptoms of differential diagnosis.[18] If this is so, MMSE could be over-estimating the cognitive deterioration level of these patients, which means it would not be the best instrument for the screening of dementia in patients with Parkinson's disease or in people where it is suspected. Despite what has been written about MMSE, the functioning of 11 of its 30 items is not the same for the different groups, which is why the comparison of the scores in the different groups would not be justified.

Although using other tests based on CTT for the screening of dementia in patients with PD has been proposed,[19] this solution makes the comparison between scores even more difficult. The most satisfactory psychometric solution is to use a test with a generalised validity, where there are no items that can prejudice the subjects of a group for reasons other than that of a latent variable. The Rasch model is the proper tool to build and validate a test with these characteristics.

Funding

This project has been funded by the Junta de Castilla y León (Ref.: SA057A08).

Conflicts of interest

The authors have no conflicts of interest to declare.

References

1. Folstein MF, Folstein SE, McHugh PR. ''Mini-mental state''. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12:189—98.
2. Lobo A, Ezquerra J, Gómez F, Sala F, Seva A. El mini-examen cognoscitivo. Un test sencillo, práctico, para detectar alteraciones intelectivas en pacientes médicos. Actas Luso Esp Neurol Psiquiatr Cienc Afines. 1976;7:189—202.
3. Guerreiro M, Silva AP, Botelho M, Leitão O, Castro-Caldas A, Garcia C. Adaptação à população portuguesa da tradução do Mini Mental State Examination (MMSE). Rev Port Neurol. 1994;1:9.

4. Ishizaki J, Meguro K, Ambo H, Shimada M, Yamaguchi S, Hayasaka C, et al. A normative, community based study of Mini-Mental State in elderly adults: the effect of age and educational level. J Gerontol B: Psychol Sci Soc Sci. 1998;53: 359—63.
5. Sahadevan S, Lim P, Tan J, Chan S. Diagnosis performance of two mental status tests in the older Chinese: influence of education and age on cutoff values. Int J Geriatr Psychiatry. 2000;15:234—41.
6. Gungen C, Ertan T, Eker E, Yasar R, Engin F. Reliability and validity of the standardized Mini-Mental State Examination in the diagnosis of mild dementia in Turkish population. Turk Psikiyatri Derg. 2002;13:273—81.
7. Noale M, Limongi F, Minicuci N. Identification of factorial structure of MMSE based on elderly cognitive destiny: the Italian Longitudinal Study on Aging. Dement Geriatr Cogn Disord. 2006;21:233—41.
8. Werner P, Heinik J, Mendel A, Reicher B, Bleich A. Examining the reliability and validity of the hebrew version of the Mini-Mental State Examination. Aging. 1999;11:329—34.
9. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719—48.
10. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.
11. Prieto G, Delgado AR, Perea MV, Ladera V. Scoring Neuropsychological Tests Using the Rasch Model: an illustrative example with the rey-osterrieth complex figure. Clin Psychol. 2009;24:45—56.
12. Cadavid N, Delgado AR, Prieto G. Construcción de una escala de depresión con el modelo de Rasch. Psicothema. 2007;19:515—21.
13. Linacre JM. A user's guide to WINSTEPS & MINISTEPS: Rasch Model computer programs. Chicago: Winsteps.com; 2006.
14. Prieto G, Delgado AR. Fiabilidad y validez. Pap Psicol. 2010;31:67—74.
15. Gibbs WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry. 1988;51:745—52.
16. Mchann G, Drachan D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRD Work group under the auspice of Departamente of Health and Human Services Task Force on Alzheimer's disease. Neurology. 1984;34:939—44.
17. Hughes CP, Berg L, Danzinger WL. A new clinical scale for the staging of dementia. Br J Psychiatry. 1988;43:2412—3.
18. García-Ramos R, Villanueva C, Del Val J, Matías-Guíu J. Apatía en la enfermedad de Parkinson. Neurología. 2010;25:40—50.
19. Parrao-Díaz T, Chaná-Cuevas P, Juri-Claverías C, Kunstmann C, Tapia-Núñez J. Evaluación del deterioro cognitivo en una población de pacientes con enfermedad de Parkinson mediante el test minimental Parkinson. Rev Neurol. 2005;40:339—44.