Actualidades en Psicología, 32(124), 2018, 65-74
http://revistas.ucr.ac.cr/index.php/actualidades

ISSN 2215-3535
DOI: https://doi.org/10.15517/ap.v32i124.31837

Contrasting Two Verbal Fluency Scoring Systems Using the Rasch Rating Scale Model

Comparación de Dos Sistemas de Puntuación de la Fluidez Verbal Mediante el Modelo de Escalas de Calificación

Gerardo Prieto[1]
Ana R. Delgado[2]
M.Victoria Perea[3]
Ricardo García[4]
Valentina Ladera[5]

University of Salamanca, Spain

Abstract. Objective: Two scoring systems for a verbal fluency test were compared using the Rasch Rating Scale Model. Method: The analysis was carried out on 289 participants, 92 of whom had had a Parkinson's disease diagnosis. Scores were calculated with two different category systems: a conventional procedure and a percentile-based one. Results: The percentile-based Rasch scores produce adequate categories and reliable measures, while the correlation with the Mini Mental State Examination evinces concurrent validity. After statistically controlling for age, percentile-based Rasch measures discriminated between both groups, demonstrating predictive validity. Conclusions: The analysis of the two procedures allows for the recommendation of the use of percentile-based categories.

Keywords. Neuropsychological assessment, parkinson's disease, verbal fluency, Rasch Rating Scale Model.

Resumen. Objetivo: Comparar dos sistemas de puntuación para un test de fluidez verbal con el Modelo de Escalas de Calificación. Método: Se analizaron datos de 289 participantes, de los cuales 92 habían sido diagnosticados con Parkinson. Las puntuaciones se calcularon con dos sistemas de categorización: un procedimiento convencional y otro basado en percentiles. Resultados: Las puntuaciones Rasch procedentes de percentiles dan lugar a categorías adecuadas y medidas fiables; la correlación con las puntuaciones del test Minimental es evidencia de validez concurrente. Tras controlar estadísticamente el efecto de la edad, las medidas Rasch procedentes de percentiles discriminan entre ambos grupos, lo que evidencia validez predictiva. Conclusiones: El análisis de los dos procedimientos permite recomendar el uso de las categorías basadas en percentiles.

Palabras clave. Enfermedad de Parkinson, evaluación neuropsicológica, fluidez verbal, Modelo de Escalas de Calificación.

[1]Gerardo Prieto. University of Salamanca, Spain. Dirección Postal: Avda. de la Merced 109-131, 37005 Salamanca, Spain. E-mail: gprieto@usal.es
[2]Ana R. Delgado. University of Salamanca, Spain. E-mail: adelgado@usal.es
[3]M.Victoria Perea. University of Salamanca, Spain. E-mail: vperea@usal.es
[4]Ricardo García. University of Salamanca, Spain. E-mail: rigar@usal.es
[5]Valentina Ladera. University of Salamanca, Spain. E-mail: ladera@usal.es

Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional.

Introduction

Verbal fluency (VF) ability is usually measured as the number of words generated under stimulus constraints such as category or first letter (Lezak, Howieson, Bigler, & Tranel, 2012). It implies multiple cognitive processes related to the activation of different brain areas (Troyer & Moscovitch, 1997), including lexical selection, phonemic coding, working memory, and executive control (Paulesu et al., 1997). VF tasks are used to assess verbal production speed, ability to initiate behaviors in response to a novel task (Bryan & Luszcz, 2000), denomination ability, response speed, mental organization, search strategy, and some aspects of short- and long-term memory, Light, Parker, & Levin, 1997). Spreen and Strauss (1998) consider VF tasks to be estimators of initiation capability, sustained attention, processing speed, and the ability to suppress inadequate responses. Deficits in VF are frequently found in diseases such as Parkinson's (Azuma, Cruz, Bayles, Tomoeda, & Montgomery, 2003; Dubois, et al., 2007; Henry, & Crawford, 2004; Jankovic, 2008) as well as in mild cognitive impairment (Rinehardt et al., 2014).

The commonest VF tasks are semantic VF (in which the participant is asked to evoke words of a certain category, e.g., animal, fruit, clothes) and phonemic VF (in which the participant is asked to evoke words starting with a letter, e.g., P, S, F) (Bryan, & Luszcz, 2000). Action VF is the ability to evoke words for action. It is also considered to be an executive functioning measure in clinical populations (Burgess, Alderman, Evans, Emslie, & Wilson, 1998; Piatt, Fields, Paolo, Koller, & Tröster, 1999). In the clinical field, VF tasks are used to detect cognitive decline (Holtzer, Goldin, & Donovick, 2009; Radanovic et al, 2009), and to tell apart normal aging from mild cognitive impairment (Bertola et al., 2014). An exhaustive review of VF tasks and their assessment utility in diverse populations can be found in Lezak, Howieson, Bigler and Tranel (2012).

Not requiring any materials, VF tasks are easy to apply in any cultural context, and so it is usual to find them as part of many neuropsychological assessment protocols such as those for language or executive functions. For instance, the Frontal Assessment Battery (FAB) includes a VF task to measure mental flexibility (Dubois, Slachevsky, Litvan, & Pillon, 2000). However, the scoring of VF tests has not received the attention that it deserves. Even though the psychometrical properties of VF scores have been hardly studied, parametric statistical methods are typically used on these scores, taking interval status for granted.Counts are sometimes arbitrarily categorized, as is the case of the FAB VF item (0-2 words= 0; 3-5 words = 1; 6-9 words = 2; > 9 words = 3).

The Rasch approach to measurement can be used to contrast the quality of scoring systems (Delgado, 2007; Prieto & Delgado, 2003; Prieto, Delgado, Perea, & Ladera, 2010). From a methodological perspective, the advantages of applying the Rasch family of models are already well known (Freitas, Prieto, Simões, & Santana, 2014). Of special interest is the fact that the measured attribute can be represented on a single dimension, an interval-scaled variable where people and items are jointly located. However, these models are still underused in the neuropsychological assessment field. Thus our objective was the empirical contrast of the functionality of two quantitative scoring systems for a VF test composed of three "items" (semantic, phonemic and action) by means of the

Rating Scale Model, an extension of the Rasch Model for polytomous items (RRSM; Andrich, 1978), whose formulation is:

$$\ln (Pnik / Pni(k-1)) = Bn - Di - Fk$$

Pnik: probability that person's n answer to item i is category k;

Pni(k-1): probability that the answer to item i or response is k-1;

Bn: ability or attribute of person n;

Di: location of item i;

Fk: transition point (step) between k and k-1.

Methods

Sample

A secondary analysis of the VF scores of 289 participants (142 female; age range: 45-95; education: 2-20 years) was carried out. Of these, 92 had been diagnosed with Parkinson's disease (P), while the remaining 197 subjects came from a community sample and served as comparison group (C). Informed consent was required. All procedures were performed in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Instruments

Semantic, phonemic and action fluency tasks were regarded as "items" composing a VF test. In the semantic task participants were asked to evoke as many animal names as they could in one minute. In the phonemic task, participants were asked to evoke as many words starting with the letter P as they could in one minute. In the action VF task, participants were asked to evoke as many verbs as they could in one minute. Combining the tasks is justified given both their common content and large score inter-correlations (*r* semantic-phonemic = .49; *r* semantic-action = .60; *r* phonemic-action = .71).

Procedure

Scores were calculated with two different category systems: the arbitrary one used by the FAB VF item (0-2 words = 0; 3-5 words = 1; 6-9 words = 2; > 9 words = 3), and a percentile-based procedure. A percentile rank is the percentage of the data that is below a concrete score. By using percentile rank ranges we have calculated the number of words corresponding to each category, as can be seen in table 1.

Data Analysis

Both sets of scores were then separately calibrated by means of the RRSM. As to person measures, maximum and minimum scores were imputed given that RRSM does not allow estimating extreme scores. Data analysis was performed with Winsteps 3.92.0 (Linacre, 2016), and the adequacy of the response categories analyzed with the following criteria: (a) sufficient frequency and regular distribution of the categories; (b) the average measures according to category increase monotonically in the rating scale; (c) no category misfit; (d) the transition points go up monotonically (Linacre, 2002).

Table 1

Percentile Range, Word Number Range and Percentile-based Category

Percentile Range	Semantic	Phonemic	Action	Category
0-9	0 -10	0-5	0-5	0
10-24	11-12	6-8	6-7	1
25-49	13-15	9-11	8-10	2
50-74	16-18	12-13	11-13	3
75-89	19-20	14-17	14-16	4
≥90	≥ 21	≥ 18	≥ 17	5

The model fit was evaluated with *Outfit*, based on the chi-square statistic, and *Infit*, based on the same statistic but with each observation weighted by its statistical information. *Infit/Outfit* values over 2 indicate severe misfit (Linacre, 2016). Principal component analysis of residuals was used to assess unidimensionality. According to Reckase (1979), the percent of variance explained should be over 20% and there should not be a second dominant factor.

After selecting the more adequate scoring system, Differential Item Functioning (DIF) for gender and for group (P and C) was tested so as to refute the hypothesis that VF scores show differential validity in these groups (Wolfe & Smith, 2007). Correlation coefficients between Rasch-modeled scores, demographic variables, and the MMSE were calculated. The difference in means between P and C groups was statistically contrasted controlling for the effect of the associated demographic variables.

Results

It can be seen from table 2 that the arbitrary response categories (FAB) were not functional according to Linacre criteria (2002). The second column shows that the observed frequency for the category 0 is not enough (it should be at least 10) to properly estimate the thresholds. The sum of the observed frequencies for the categories is the number of items by the number of subjects. The category score distribution is very asymmetrical:

Table 2

Arbitrary (FAB) Category System Statistics.

Category	Observed[a]	Average[b]	*Infit*	*Outfit*	*Thresholdc*
0	2	-2.15	0.83	0.87	-
1	41	-0.38	0.87	0.89	-4.13
2	137	2.84	1.01	1.00	-0.05
3	687	7.23	1.02	1.04	4.18

[a]Observed category frequency= count of observations in category.

[b]Average measure = sum (Bn - Di) / count of observations in category

most of the observed frequencies (79%) are clustered in the category 3, artificially reducing the variability and thus the reliability of the person scores (*Model Person Separation Reliability = .33; Cronbach's alpha= .56).*

Table 3

Percentile-based Category System Statistics

Category	Observed[a]	Average[b]	*Infit*	*Outfit*	*Threshold*
0	60	-2.38	.94	.94	-4.03
1	115	-1.47	1.11	1.12	-2.29
2	219	-.52	.90	.88	-.80
3	233	.53	1.00	1.00	.77
4	134	1.77	.88	.89	2.30
5	106	2.41	1.11	1.10	4.06

[a]Observed category frequency= count of observations in category.

[b]Average measure = sum (Bn - Di) / count of observations in category

Conversely, the percentile-based response categories are clearly functional, as can be seen from table 3.

Score reliability was much better than with the previous system (*Model Person Separation Reliability = .82; Cronbach's alpha =.79*). Thus, the remaining analyses were carried out on the scores calculated with this percentile-based response category system that, once modeled with the RRSM, will be called measures.

The unidimensionality assumption was fulfilled: the variance explained by the main dimension was very large (64.5%); the eigenvalue of the residual variance first component was 1.74. It can be seen from Table 4 that the remaining fit statistics were also good.

Table 4

Item Statistics

Item	*D*	*SE*	*Infit*	*Outfit*
Semantic	-.11	.08	1.25	1.25
Phonemic	.11	.08	.99	.99
Action	.00	.08	.72	.72

Differential Item Functioning (DIF) occurs when an item has a different probability of being passed by persons of a certain group after controlling for the measured attribute. To test for DIF in the Rasch approach, the standardized difference between group parameter locations is calculated after adjusting for group differences and a Bonferroni correction of the significance level is then carried out (Linacre, 2016). Neither gender-related nor group item DIF was found.

As to person measures, 15 out of 289 were imputed given that RRSM does not allow to estimate the perfect (12 maximum and 3 minimum) scores. The frequency of severe misfit (Infit and/or Outfit > 2) was 38 (13.1%). VF measures had a mean value of 0.30 (*SD*= 1.97), i.e., slightly over the mean item difficulty, which is conventionally located at the scale zero. The unit of the interval variable constructed by means of the RRSM is the logit.

VF measures significantly correlated with age (r = -.21, p < .001) and education years (r = .52, p < .001), but not with gender (r = .02, p =.69). For P and C groups, the mean (*SD*) was .27 (2.17) and .32 (1.87), respectively, which is a non-significant difference, t (287) = .18, p =.87. Statistically controlling for education by means of ANCOVA, the difference between P and C remains non-significant, F (1, 286)= 1.62, p = .20. However, when the effect of age is controlled, the difference between P and C becomes significant, F (1 , 286)= 6.35, p = .01. This is evidence for predictive validity.

Finally, the correlation of VF measures with the MMSE scores is r = .57, p < .001, evidencing concurrent validity.

Discussion

Two scoring systems have been evaluated with the RRSM corroborating that the arbitrary category system was not functioning adequately. Percentile-based response categories were clearly functional, and the resulting scores showed good fit and generalized validity for both genders as well as for P and C groups. As usual, VF measures significantly correlated with age and education years, but not with gender. Predictive validity was also supported given the mean differences between P and C scores (after controlling for age), which evidences diagnostic utility. Concurrent validity was also supported, given the large positive correlation of VF Rasch-modeled measures with the MMSE scores.

Even though the correct performance in the various VF tasks requires shared cognitive processes (Troyer, & Moscovitch, 1997) including sustained attention, searching strategy maintenance, lexical selection, inhibition ability, working memory and articulation, there are also some differences.

Semantic VF is related to verbal memory and storing (specially linked to the temporal lobe: Birn et al., 2010; Hodges, & Patterson, 2007) while phonemic VF is less dependent on memory and more related to initiation and shifting abilities (linked to the frontal lobes: Troster et al., 1998; Troyer, Moscovitch, Winocur, Alexander, & Stuss, 1998; Troyer, Moscovitch, Winocur, Leach, & Freedman, 1998). Action VF requires working memory, frontal executive processing, initiation ability, sustained attention and searching strategy maintenance (Perea, Ladera, & Rodríguez, 2005).

In this study, percentiles are given for each of the VF scores, apart from considering the whole test score. In practice this is very useful for clinicians, given the above exposed differences in cognitive processing. VF patterns are used to tell apart deficits associated to the frontal lobe from those associated to the temporal lobe. Frontal lobe injuries lead to

low phonemic (Baldo, Shimamura, Delis, Kramer, & Kaplan, 2001; Hodges et al., 1999) and action VF (Damasio, & Tranel, 1993) while temporal lobe injuries give place to deficits in semantic VF (Baldo, Schwartz, Wilkins, & Dronkers, 2006; Hodges et al. 1999) with relatively well preserved verb-evoking ability (Damasio, & Tranel, 1993). Our data allow the location of an individual VF task performance helping to tell apart anterior injuries (frontal) from the posterior (temporal) ones.

Finally, it is relevant to note that, even though most neuropsychological test scores are ordinal-level at best, parametric statistical methods are usually found in the reporting of data analysis. The RRSM logistic transformation has served to construct an interval-level variable, which is desirable from both a scientific perspective and a diagnostic one (e.g., measuring change in patient status is allowed). Comparison of a patient with the remaining participants is implicit in the percentile-based category system, which facilitates personalized interpretation. Finally, as usual in the Rasch approach, unexpected response patterns can give place to new clinical and/or scientific hypotheses (Prieto, Delgado, Perea, & Ladera, 2010).

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi: 10.1007/BF02293814

Azuma, T., Cruz, R. F., Bayles, K. A., Tomoeda, C. K., & Montgomery E. B. (2003). A longitudinal study of neuropsychological change in individuals with Parkinson's disease. *International Journal of Geriatric Psychiatry, 18,* 1043-1049. doi: 10.1002/gps.1022

Baldo, J. V., Schwartz, S., Wilkins, D., & Dronkers, N. F. (2006). Role of frontal versus temporal cortex in verbal fluency as revealed by voxel-based lesion symptom mapping. *Journal of the International Neuropsychological Society, 12,* 896-900. doi: : 10.10170S1355617706061078

Baldo, J. V., Shimamura, A. P., Delis, D. C., Kramer, J., & Kaplan, E. (2001). Verbal and design fluency in patients with frontal lobe lesions. *Journal of the International Neuropsychological Society, 7,* 586-596. doi: 10.1017/S1355617701755063

Bertola, L., Mota, N. B., Copelli, M., Rivero, T., Diniz, B. S., Romano-Silva, M. A., . . . Malloy-Diniz, L. F. (2014). Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in Aging Neuroscience, 6,* 185. doi: 10.3389/fnagi.2014.00185

Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B. Bandettini, P. A., & Martin, A. (2010). Neural systems supporting lexical search guided by letter and semantic category cues: A self-paced overt response fMRI study of verbal

fluency. *NeuroImage, 49,* 1099-1107. doi: 10.1016/j.neuroimage.2009.07.036

Bryan, J., & Luszcz, M. (2000). Measurement of executive function: considerations for detecting adult age differences. *Journal of Clinical and Experimental Neuropsychology, 22,* 40-55. doi: : 10.1076/1380-3395(200002)22:1;1-8;FT040

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society, 4*, 547-558.

Damasio, A., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Science of the United States of America, 90,* 4957-4960.

Delgado, A.R. (2007). Using the Rasch Model to quantify the causal effect of instructions. *Behavior Research Methods, 39,* 570-573.

Dubois, B., Burn, D., Goetz, C., Aarsland, D., Brown, R .G., Broe, G. A... Emre, M. (2007). Diagnostic procedures for Parkinson's disease dementia: Recommendations from the movement disorder society task force. *Movement Disorders, 22*, 2314-2324. doi: 10.1002/mds.21844

Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. (2000). The FAB: a Frontal Assessment Battery at bedside. *Neurology, 55*, 1621-1626. doi: 10.1212/WNL.55.11.1621

Freitas, S., Prieto G., Simões, M.R., & Santana, I. (2014). Psychometric properties of the Montreal Cognitive Assessment (MoCA): an analysis using the Rasch model. *The Clinical Neuropsychologist, 28,* 65-83. doi: 10.1080/13854046.2013.870231

Henry, J. D., & Crawford J. R. (2004). Verbal fluency deficits in Parkinson's disease: a meta-analysis. *Journal of International Neuropsychological Society, 10,* 608-622. doi: 10.1017/S1355617704104141

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *Lancet Neurology, 6*, 1004-1014. doi: 10.1016/S1474-4422(07)70266-1

Hodges, J. R., Patterson, K., Ward, R., Garrard, P., Bak, T., Perry, R., & Gregory, C. (1999). The differentiation of semantic dementia and frontal lobe dementia (temporal and frontal variants of frontotemporal dementia) from early Alzheimer's disease: a comparative neuropsychological study. *Neuropsychology, 13,* 31-40.

Holtzer, R., Goldin, Y., & Donovick, P. J. (2009). Extending the Administration Time of the Letter Fluency Test Increases Sensitivity to Cognitive Status in Aging. *Experimental Aging Research, 35,* 317-326. doi 10.1080/03610730902922119

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery &Psychiatry, 79*, 368-376. doi: 10.1136/jnnp.2007.131045

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment (5th ed.*). New York: Oxford University Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J. M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2016. Retrieved from http://www.winsteps.com/

Paulesu, E., Goldacre, B., Scifo, P., Cappa, S. F., Gilardi, M.C., Castiglioni, I., Perani, D., & Fazio, F. (1997). Functional heterogeneity of left inferior cortex as revealed by fMRI. *Neuroreport, 8,* 2011-2016.

Perea, M.V., Ladera, V., & Rodríguez, M. A. (2005). Fluencia de acciones en personas mayores. [Action fluency tasks in Spanish subjects over 65 years old]. *Psicothema, 17,* 263-266.

Piatt, A., Fields, J., Paolo, A., Koller, W., & Troster, A. (1999). Lexical, semantic, and action verbal fluency in Parkinson's disease with and without dementia. *Journal of Clinical and Experimental Neuropsychology, 21,* 435-443. doi: 10.1076/jcen.21.4.435.885

Prieto, G., & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. [Rasch-modelling a test]. *Psicothema, 15,* 94-100.

Prieto, G., Delgado, A. R., Perea, M.V., & Ladera, V. (2010). Scoring neuropsychological tests using the Rasch model: An illustrative example with the Rey-Osterrieth Complex Figure. The Clinical Neuropsychologist, 24, 45-56. doi:10.1080/13854040903074645

Radanovic, M., Diniz, B. S., Mirandez, R. M., Novaretti, T. S., Flacks, M. K., Yassuda, M. S., & Forlenza, O. V. (2009). Verbal fluency in the detection of mild cognitive impairment and Alzheimer's disease among Brazilian Portuguese speakers: the influence of education. *International Psychogeriatrics, 21,* 1081-1087. doi:10.1017/S1041610209990639

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.doi: 10.2307/1164671

Rinehardt, E., Eichstaedt, K., Schinka, J. A., Loewenstein, D. A., Mattingly, M., Fils, J., ...Schoenberg, M. R. (2014). Verbal fluency patterns in mild cognitive impairment and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 38,* 1-9. doi:10.1159/000355558

Ruff, R. M., Ligth, R. H., Parker, S. B., & Levin, H.S. (1997). The psychological construct of word fluency. *Brain and Language, 57,* 349-405. doi:10.1006/brln.1997.1755

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: administration, norms, and commentary.* Oxford: Oxford University Press.

Troster, A. I., Fields, J. A., Testa, J. A., Paul, R. H., Blanco, C. R., Hames, K. A...Beatty, W.W. (1998). Cortical and subcortical influences on clustering and switching in

the performance of verbal fluency tasks. *Neuropsychologia, 36,* 295-304.

Troyer A. K, & Moscovitch, M. (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology, 11,* 138-146.

Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M.P, & Stuss, D. (1998). Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia, 36,* 499-504.

Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society, 4,* 137-43.

Wolfe, E. W., & Smith, E.V. (2007). Instrument development tools and activities for measure validation using Rasch models: part II-validation activities. *Journal of Applied Measurement, 8,* 204-234.

Recibido: 20 de Diciembre 2017
Aceptado: 17 de Abril 2018