Optimizing rating scales of the big five socialization scale in athletes by Rasch model*

Optimizando las cinco categorías de calificación para una escala de socialización de atletas bajo el modelo Rasch

Received: 18 June 2015 | Accepted: 14 July 2016

Daniel Bartholomeu[**]
Fundação Instituto de Ensino para Osasco, Brasil
José Maria Montiel
Fundação Instituto de Ensino para Osasco, Brasil
Afonso Antonio Machado
Universidade Estadual Paulista, Brasil

ABSTRACT
Personality is one of the main psychological aspects that influence athletes' performance in competition and one of the most studied subjects in sport psychology. We aimed to optimize item scales in a socialization test based on the big-five model, assessed by means of adjectives and administered to a sample of 225 athletes of both genders, with 56.9% male. Age ranged from 14 to 45 years with a mean of 20 (SD = 5.21). Participants attended to basketball (11%), football (21.8%), handball (17.3%), jiu-jitsu (10.2%), tennis (5.60) and volleyball (16 %) sport modalities. The results indicated that the four point scale was the best item scale structure regarding validity evidences. Hence, this structure could be adopted in this scale aiming to better socialization assessment in athletes.

Keywords
psychological assessment, human performance, personality.

RESUMEN
La personalidad es uno de los principales aspectos psicológicos que influyen en el rendimiento de los atletas en la competencia y uno de los temas más estudiados en la psicología del deporte. El objetivo de este estudio consistió en optimizar los ítems de las escalas en una prueba de socialización basada en el modelo de los Cinco Grandes, estableciendo asociaciones a partir de adjetivos y aplicando una muestra de 225 atletas de ambos sexos, donde el 56.9 % eran hombres. El rango de edad oscilaba entre 14 y 45 años, con una media de 20 (DE = 5.21). Los participantes se distribuyeron en las siguientes modalidades: baloncesto (11 %), fútbol (21.8 %), balonmano (17.3 %), jiu-jitsu (10.2 %), tenis (5.6 %) y voleibol (16 %). Los resultados indicaron que la escala de cuatro puntos ofrece los mejores resultados sobre la estructura de la escala. Por lo tanto, esta escala se podría adoptar para una mejor evaluación de la socialización en los atletas.

Palabras clave
evaluación psicológica, rendimiento humano, personalidad.

**Strictu Sensu Post Graduate Program in Educational Psychology. FIEO-University Center - UNIFIEO/SP, Brasil.

To cite this article: Bartholomeu, D., Montiel, J. M., & Machado, A. A. (2016). Optimization rating scales for an athletes' socialization scale by Rasch model: Scales of socialization. *Universitas Psychologica, 15* (4). http://dx.doi.org/10.11144/Javeriana.upsy15-4.orsa

Introduction

Personality is one of the most important psychological aspects that affects athletes' performance in competition and one of the most studied subject since the emergence of Sport Psychology after the World War II. From 1950 to 1970 there was a tendency to measure this variable in athletes mainly to compare their results with groups of non-athletes, or between different sports (Cratty, 1984). This manuscript focuses on the Five-Factor Model of personality (FFM). This model does not provide an a priori theoretical explanation of the five factors of personality but derives then based on the factor analysis of numerous instruments such as the 16 PF (Cattell & Cattell, 1995), MMPI (Butcher et al., 1992), Murray Needs, among others that provide similar solutions to the FFM, regardless of the underlying theory (Hutz et al., 1998).

Numerous authors, such as Goldberg (1982), studied the factorial solutions found between personality tests and contributed to the current understanding of the factors to explain personality. Research during the past 10 years has demonstrated the solidity of such factors. Thus, this model has been considered by many authors as the best alternative for the description of personality. It is suggested that these are basic dimensions of personality desirable to ascertain in any people with whom one will interact (Hutz et al., 1998; McCrae, Costa, & Piedmont, 1993). Whereas the FFM has its origins in the analysis of language used to describe people, using traits descriptors (adjectives) can help identify these personality factors. In the opinion of Goldberg (1982), if a personality feature is evident, a word could be enough to describe it.

In Brazil, the study of Hutz et al. (1998) investigated the adequacy of a list of adjectives used to describe personality on a big five factor model. Factor analysis demonstrated the existence of five factors as expected. The first factor was "Socialization" ("agreableness"), followed by "Extraversion", "Scrupulosity", "Neuroticism", and "Openness to Experience", with Cronbach's alpha coefficients ranging between 0.78 and 0.88. The total variance explained by factors was approximately 44% and multivariate analyzes of variance (MANOVA) indicated significant gender differences in all factors except neuroticism.

Despite this study, no researches were made with athletes with this scale as well as no test on the items rating scale. In Brazilian athletes, this fact assumes greater importance since no personality test has specific psychometric properties to this context, especially in the FFM. The present research focuses on Socialization items once we used Rasch model that supposes unidimensionality to run. Regarding polytomous scales meanings, specifically in the case of socialization, each consecutive number in the scale represents a higher amount of socialization and the increase in scale points indicates a higher incidence of this aspect. As a person moves in the continuum of these variables, each successive point becomes the most likely response (assuming that subjects were able to distinguish between all levels of the scale).

Among the various models of item response theory available, we analyzed the item fit to Rasch model. This system considers only the item difficulty parameter and people ability as a function to determine the probability to score an item. This model is the most popular within the item response theory due to its greater mathematical simplicity (Muñiz, 1990). Also, it assumes data additivity, defined as measurement units (logits) that have the same size in the continuous (interval data), if the data fit to the model. Thus, these parameters are estimated and used to determine the response patterns expected for each item. The adjustment is derived from a comparison of these with the observed patterns that provide validity evidences for the test. In

turn, the standard errors associated with item calibration and people ability estimate are used in the reliability estimates in this model. These errors can be used to describe the confidence interval in which the true item difficulty and people ability are found (Wright & Stone, 1988).

Rasch model can be used in optimizing the number of scale categories in a test without the need to administer different versions of the same scale. Some ways of observing the number of response categories appropriate to the items of a test can be proposed (Bartholomeu, Montiel, & Machado, 2013). The thresholds parameters can be also observed to determine which categories are not effective in measuring the variable under investigation. Ordered thresholds imply that as a person moves along the continuum of socialization, each category becomes the most likely answer. The disorder occurs in the responses for the same reasons previously mentioned and can be best seen through the probability curves graph.

Finally, an outfit analysis can detect the use of random categories (Linacre, 1997, 1999). Some possible solutions to these mismatches may be combining adjacent categories, change the location of adjacent categories, or treating the missing responses as if the name is not appropriate or does not share the same trait that the other (Linacre, 1997, 1999). It is worth noting that the diagnosis should be made taking into account these three criteria, besides visual inspection of the probability plot of categories agreement. In this context, considering the lack of studies with personality tests in Brazilian athletes and that no research, were found analyzing item category optimization in personality tests to athletes, this study aimed to optimize item rating scales in a socialization factor assessed by means of adjectives.

Method

Participants

Two hundred and twenty-five athletes of both genders, with 56.9% male and ages ranging from 14 to 45 years, mean of 20 (SD = 5.21), were studied. Regarding the modalities, these were basketball (11%), football (21.8%), handball (17.3%), jiu-jitsu (10.2%), tennis (5.60) and volleyball (16%). The survey was conducted in several clubs in the state of São Paulo, Brazil. The educational level was also varied, ranging from elementary school (4.7%) to the doctoral level (0.6%), and most subjects (62.8%) had incomplete university level. 58.7% of the participants worked and practiced sports, although most have had a routine of intense training, mostly three to five days a week (78%) as well as two to three hours a day (76.2%). Regarding competition, 90.7% reported that they had attended in competitions.

Instruments

Big Five Adjectives (Hutz et al., 1998)

This test presents a list of 64 adjectives and participants should indicate the agreement with that item as a good descriptor of its personality in a five-point Likert scale. The answers are arranged from 1, Strongly Disagree to 5, Strongly Agree. In relation to the factors attained by the instrument, Factor I assesses Socialization and adjectives are affable, sociable, docile, nice, generous, romantic, gentle, kind, understandable, friendly, cold, kind, passionate, friendly, sentimental, and delicate, making a total of 16 items in this subscale. Factor II concerns the Extraversion and its items are shy, extroverted, communicative, resourceful, introverted, embarrassed, quiet, inhibited, and shut, totaling 10 items. In turn, Factor III has information on Realization (Scrupulosity) and adjectives that characterize it are honored, responsible, dedicated, hardworking, studious, honesty, disorganized, efficient, careful, methodical, organized, meticulous, devoted, and pervaded, forming a total of 14 items for this subscale. Factor IV concerns the Neuroticism and comprises adjectives such as pessimistic, happy, bored, affirmative, selfish,

unhappy, depressed, insecure, obnoxious, lonely, anxious, and sad, totaling 12 items on this scale.

Finally, Factor V informs about Opening and adjectives are curious, funny, creative, philosophical, courageous, energetic, adventurous, audacious, imaginative, intellectual, artistic, and impulsive composing the total of 12 items. The result for each scale was obtained by summing the scores given to each item divided by the total number of items in each corresponding subscale. Some psychometric properties of the instrument were taken by Hutz et al. (1998). Factor analysis provided a total variance explained by these five factors of 43.91%. Beside this, the Cronbach's alpha values for the dimensions are 0.88, 0.88, 0.84, 0.89, 0.78, respectively, and can be considered satisfactory.

Procedure

Data collection was collective and made after participants' acceptance by completing an authorization term of free and informed consent. The project followed all ethical principles for research with human subjects and was approved by the Research Ethics Committee of Anhanguera University under the number 810/2011. The evaluation was held at the Club in a meeting room with chairs and appropriate conditions for application of the instruments. We distributed an answer sheet of the test for the participants. The application of the instrument was part of an intervention project that aimed collecting information to be used later for determining the techniques to be employed and communicated to the participants that the data would be kept confidential and would also later used in a research. Following the objectives of the study, the analyzes were performed by the program *Winsteps* .

Results and Discussion

In spite of classical tests theory, which uses total scoring by summing the items in a scale, the item response theory (IRT) allows different interpretation of test scores according to personal abilities assessed in each response. A comparison of the three IRT models made by Wright (1992) pointed out some advantages of the Rasch model over the other two models. The Rasch model was derived for defined measures specifying the measure requirements. Its mathematical formulation is solid and provides statistics that enable reaching linear and objective measurement. Also, in the Rasch model guessing is not accepted for being considered as unreliable. Not everyone makes use of guessing, let alone for the same items. Moreover, variations on item discrimination are rejected by the Rasch model for being considered a symptom of item bias. The items discrimination variance is affected by items bias or by extra dimensions.

In terms of Socialization, the difficulty parameter reveals that the more difficult an item is, the less agreement is displayed on it, and the less incidence of that trait is presented. That is, a subject presenting high levels of agreement on items that extreme, some trait will probably present a higher incidence of this sort of characteristics, also enhancing the probability of presenting a high level of agreement on items that measure lower levels on that personality trait (easier items).

The accuracy assessed by this model provided an index of 0.97 for items and 0.70 for persons, which favors the interpretation that subjects provided more data about the items than they did about their behaviors. Despite this, it is precise both for items and for people. The average measurement error was 0.15 for items and 0.06 for persons. The subject's precision indicates the possibility of finding similar results in the same sample whether submitted to another group of items with same characteristics and that assess the same underlying construct. This fact is more likely to happen in cases with less significant measurement errors, as was the case in the present study. The items precision informs the replicability of the items if the same indicators were employed in another sample with similar levels of ability in the latent trait (Bond & Fox,

2001). This, as well as additional information, is featured in Table 1.

TABLE 1

Summary statistics of Rasch model for the Socialization items with 5-point scale .

Summary Statistics of Rasch Model for the Socialization Items with 5 Point Scale

	Raw score	Count	Measure	Model error	Infit mnsq	Infit zstd	Outfit mnsq	Outfit zstd
People								
Mean	71.3	17.9	0.96	0.29	1.13	0.2	1.07	0.1
S.D.	7.6	0.4	0.63	0.06	0.54	1.3	0.71	1.4
Max.	87	18	3.02	0.61	2.97	3.7	6.48	7.8
Min	46	16	-0.52	0.21	0.23	-3	0.22	-2.9
Reliab.	0.77							
Items								
Mean	519.2	130.2	0	0.11	1	0	1.07	0.0
S.D.	74.3	1.5	0.62	0.02	0.43	2.8	0.65	3.1
Max.	600	131	1.81	0.15	2.45	8.8	3.51	9.9
Min	279	125	-0.96	0.08	0.56	-3.1	0.55	-3.2
Reliab.	0.97							

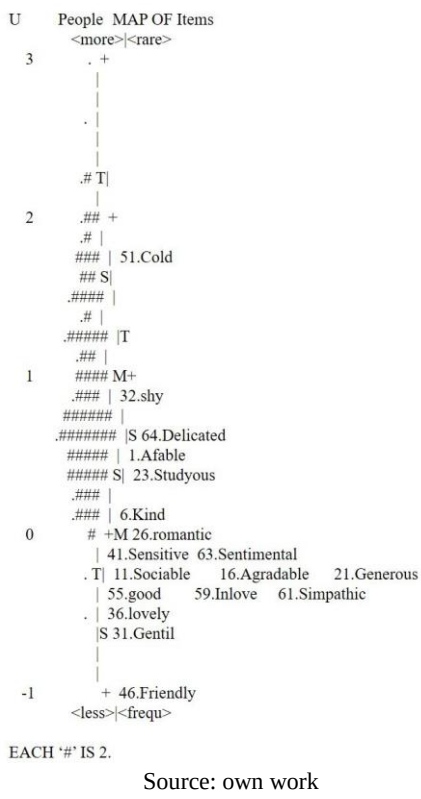
Source: own work

Concerning the fit to Rasch model, two procedures must be observed: (a) the infit, which informs the discrepancy between the observed and expected data in the area of the item characteristic curve (ICC), in which the probability of a high level of agreement with the item is close to 50% and (b) to outfit, which corresponds to the unexpected pattern of response at the extreme regions of the same curve, with high and low probabilities of agreement with the item. This adjustment is performed to the items as to the persons, and Bond and Fox (2001) considered that 1.00 would be an expected value for each of these measures. However, our study followed the Linacre (2002) suggestion that the 1.50 level would be considered the maximum limit of acceptance for an item. Only two items presented *infit* and *outfit* problems, above 1.5 (items Shy and Cold), suggesting that higher agreement rates on these are unexpected considering the latent trait Socialization. Indeed, this amount of fit problems may be considered small. Regarding persons fit to Rasch model (in terms of infit and outfit), it has generally been above the expected values. About 10% of people presented infit and outfit indicators above 1.50 that can be considered a very small amount.

A detailed analysis of people's Socialization and the items difficulty (level of agreement on the trait assessed by the indicators) suggested that all items have measured mean levels of Socialization (Figure 1). All items ranged within –1 and 1, which assess persons with more accuracy (Bond & Fox, 2001). The mean scores of item measures is higher than the persons mean measures scores which suggests that items assess higher levels of Socialization than the athletes sample actually has. The items with the higher agreement were Friendly and Gentle adjectives and the item with lower agreement was the adjective Cold.

Figure 1

Map of Item and Person.



Source: own work

We examined the differential item functioning by gneder. In this analysis the intensity of the characteristics in persons are estimated and subtracted by group. Then, it is supposed that the intensity of each item must be statistically equivalent in the studied groups (gender, in this case). This procedure estimates the quantity of DIF measure added to item positively or negatively and calculates the probability of this difference to be aleatory and furnishes a Student t measure. Draba (1977) considers that 2 is a good point to determine statistical significance in analyzing less than 20 items (as in this case). Table 2 presents these results.

TABLE 2

DIF analysis criteria to the Big Five Socialization scale items .

DIF Analysis Criteria to the Big Five Socialization Scale Items

People	Dif measure	Dif S.E.	People	Dif measure	Dif S.E.	Dif contrast	Joint S.E.	t	d.f.	Prob.	M.H. Prob.	M.H. Size	Item Number
1	0.51	0.11	2	0.40	0.15	0.11	0.19	0.60	123	0.546	0.532	0.27	1
2	0.40	0.15	1	0.51	0.11	-0.11	0.19	-0.60	123	0.546	0.532	-0.27	1
1	0.16	0.12	2	-0.01	0.17	0.17	0.21	0.81	129	0.420	0.436	-0.06	6
2	-0.01	0.17	1	0.16	0.12	-0.17	0.21	-0.81	129	0.420	0.436	0.06	6
1	-0.28	0.14	2	-0.23	0.19	-0.05	0.23	-0.23	129	0.819	0.762	-0.23	11
2	-0.23	0.19	1	-0.28	0.14	0.05	0.23	0.23	129	0.819	0.762	0.23	11
1	-0.37	0.15	2	-0.10	0.18	-0.27	0.23	-10.17	127	0.244	0.330	-0.32	16
2	-0.10	0.18	1	-0.37	0.15	0.27	0.23	10.17	127	0.244	0.330	0.32	16
1	-0.30	0.14	2	-0.16	0.18	-0.14	0.23	-0.61	129	0.542	0.771	10.18	21
2	-0.16	0.18	1	-0.30	0.14	0.14	0.23	0.61	129	0.542	0.771	-10.18	21
1	0.42	0.11	2	0.39	0.15	0.03	0.18	0.15	129	0.878	0.929	-0.48	23
2	0.39	0.15	1	0.42	0.11	-0.03	0.18	-0.15	129	0.878	0.929	0.48	23
1	0.09	0.12	2	-0.30	0.19	0.39	0.23	10.72	129	0.088	0.351	0.35	26
2	-0.30	0.19	1	0.09	0.12	-0.39	0.23	-10.72	129	0.088	0.351	-0.35	26
1	-0.64	0.16	2	-0.46	0.20	-0.18	0.26	-0.70	129	0.484	0.311	-0.45	31
2	-0.46	0.20	1	-0.64	0.16	0.18	0.26	0.70	129	0.484	0.311	0.45	31
1	0.85	0.10	2	0.90	0.13	-0.04	0.17	-0.25	128	0.799	0.999	0.18	32
2	0.90	0.13	1	0.85	0.10	0.04	0.17	0.25	128	0.799	0.999	-0.18	32
1	-0.56	0.16	2	-0.26	0.19	-0.30	0.24	-10.22	129	0.223	0.128	-0.05	36
2	-0.26	0.19	1	-0.56	0.16	0.30	0.24	10.22	129	0.223	0.128	0.05	36
1	-0.30	0.14	2	0.02	0.17	-0.32	0.22	-10.47	129	0.142	0.095	-0.77	41
2	0.02	0.17	1	-0.30	0.14	0.32	0.22	10.47	129	0.142	0.095	0.77	41
1	-0.89	0.18	2	-10.09	0.26	0.20	0.32	0.62	129	0.535	0.875	-0.54	46
2	-10.09	0.26	1	-0.89	0.18	-0.20	0.32	-0.62	129	0.535	0.875	0.54	46
1	10.50	0.10	2	20.39	0.17	-0.89	0.20	-40.56	129	0.000	0.000	-0.30	51
2	20.39	0.17	1	10.50	0.10	0.89	0.20	40.56	129	0.000	0.000	0.30	51
1	-0.52	0.15	2	-0.18	0.18	-0.34	0.24	-10.40	128	0.163	0.097	-10.10	55
2	-0.18	0.18	1	-0.52	0.15	0.34	0.24	10.40	128	0.163	0.097	10.10	55
1	-0.24	0.14	2	-0.46	0.20	0.21	0.25	0.87	129	0.388	0.488	+0.20	59
2	-0.46	0.20	1	-0.24	0.14	-0.21	0.25	-0.87	129	0.388	0.488	-0.10	59
1	-0.40	0.15	2	-0.40	0.20	-0.01	0.25	-0.03	126	0.972	0.877	-0.20	61
2	-0.40	0.20	1	-0.40	0.15	0.01	0.25	0.03	126	0.972	0.877	0.20	61
1	0.03	0.13	2	-0.54	0.21	0.57	0.24	20.34	128	0.029	0.046	0.93	63
2	-0.54	0.21	1	0.03	0.13	-0.57	0.24	-20.34	128	0.020	0.046	-0.93	63
1	0.85	0.10	2	0.05	0.17	0.80	0.20	40.11	129	0.001	0.000	0.65	64
2	0.05	0.17	1	0.85	0.10	-0.80	0.20	-40.11	129	0.000	0.000	-0.65	64

Size of Mantel-Haenszel slice = 0.100 logits

Size of Mantel-Haenszel slice = 0.100 logits

Source: own work

Three items from 16 (18%) yielded t scores above the suggested point, revealing that some items favored one of the groups and are biased by characteristics of one of the genders that are independent of Socialization level. The items Sentimental and Delicate favored females and are easier to agree in this group than for males. The item Cold favored male athletes.

We investigated whether the number of item categories is representative of socialization (Table 3, Figure 2). The progression of Rasch measure indicated an increased ability of people in each response categories. The Outfit level, suggested good fits in all categories, except for category 1 (totally disagree). Finally, by analyzing the thresholds graph, there was a discontinuity in categories 2 and 3 (neither agree nor disagree and partially agree, respectively). In this context, one solution would be to combine the responses of categories 2 and 3.

TABLE 3

Tables structure of categories of the five-point structure .

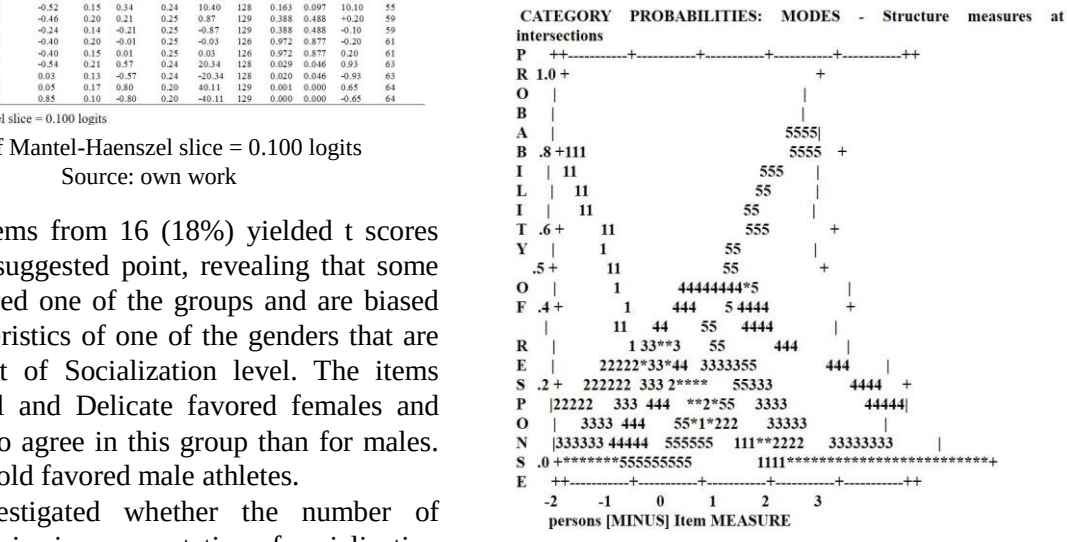
Tables Structure of Categories of the Five Point Structure

Category score	Observerd Count	%	Observed average	Sample expect	Infit mnsq	Outfit mnsq	Structure calibrat	Categoy measure	Label
1	131	6	-0.30	-0.54	1.28	1.37	None	-1.93	1 disagreestrongly
2	124	5	-0.01	0.02	0.96	0.85	-0.21	-0.85	2 Disagreesomewhat
3	321	14	0.47	0.52	0.97	0.93	-0.68	-0.14	3 Neitheragreeordisagree
4	837	35	0.87	0.97	1.12	0.99	-0.22	0.73	4 partiallyagree
5	931	39	1.52	1.45	1.02	1.13	1.10	2.37	5 stronglyagree

Source: own work

Figure 2

Probability chart categories for the socialization scale in the 5-point model



Source: own work

These results revealed more appropriate data with better outfit and continuity of the proposed categories as suggested in the graph. However, it has not been satisfactory since categories 2 and 4 did not discriminate well the participant's answers yet. To ensure that this number of categories are good to represent data assessed with this test, validity evidences with this test format are necessary (Bond & Fox, 2001; Wright & Masters, 1982). So we performed Rasch analysis again and checked the adjustment of items with this new format in the Rasch scale to provide further validity evidences of the internal structure of the items (Table 4, Figure 3).

TABLE 4

Categories structure for the socialization scale with four categories

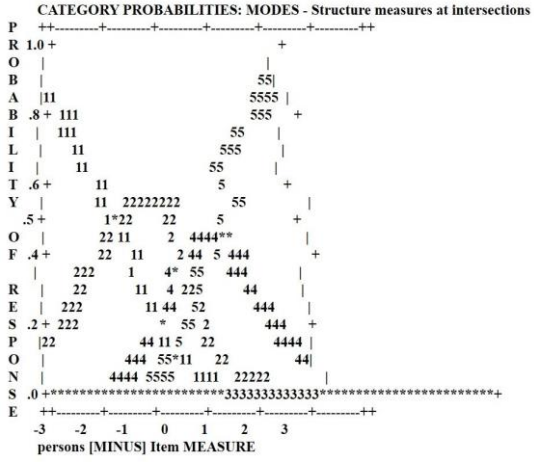
Categories Structure for the Socialization Scale with Four Categories

Category score	Observerd Count	%	Observed avrge	Sample expect	Infit mnsq	Outfit mnsq	Structure calibrat	Categoy measure	Label
1	131	6	-0.30	-0.54	1.28	1.37	NONE	-1.93	1 disagreestrongly
2	445	19	0.22	0.15	1.07	1.08	-1.29	-0.63	2 Disagreesomewhat
3	0	0			0.00	0.00	NULL	0.08	
4	837	35	0.72	0.84	1.13	1.07	0.38	0.72	4 partiallyagree
5	931	39	1.27	1.20	.96	1.14	0.91	2.09	5 stronglyagree

Source: own work

Figure 3

Probability chart categories for the socialization scale with four categories



Source: own work

The validity evidences for these items with four category format were slightly better than with the five category format. Indeed, the four category format provided almost the same reliability results as well as the same amount of items with infit and outfit problems and similar standard errors. The same is applicable for people. As no significant differences were found in difficulty level and persons' abilities between the two items format, no important differences could be found in the item-person map. Table 5 presents the summary of the Rasch model of the items and persons with this format.

TABLE 5

Summary statistics of Rasch model for the Socialization items with 4-point scale

Summary Statistics of Rasch Model for the Socialization Items with 4 Point Scale

	Raw score	Count	Measure	Model error	Infit mnsq	Infit zstd	Outfit mnsq	Outfit zstd
People								
MEAN	68.9	170.9	0.79	0.25	10.09	0.1	10.09	0.1
S.D.	9.1	0.4	0.56	0.07	0.46	10.2	0.65	10.4
MAX.	87.0	180.0	20.71	0.60	30.09	30.2	40.72	60.2
MIN.	44.0	160.0	-0.36	0.19	0.35	-30.0	0.28	-20.5
Reliab.	0.78							
Items								
MEAN	501.3	1300.2	0.00	0.09	10.02	0.0	10.08	0.2
S.D.	78.6	10.5	0.53	0.01	0.40	20.4	0.61	20.6
MAX.	596.0	1310.0	10.56	0.13	20.45	70.2	30.37	80.5
MIN.	265.0	1250.0	-0.83	0.07	0.58	-20.8	0.57	-20.5
Reliab.	0.97							

Source: own work

Also, the same DIF items with other format presented DIF again. Hence, the only advantage with this format is better information of each category. It does not make sense to include a category in the items that does not yield any better information on the latent trait (Bond & Fox, 2001; Smith, Wakely, de Kruif, & Swartz, 2003; Wright & Masters, 1982). Since the results with four categories were slightly better than the five-point format, we constraint another category to ascertain if better validity results are obtained. Then we combined categories two and four and the results are presented in Table 7. This item scale showed the best discrimination between the categories with better outfit and continuity of the proposed categories, as indicated in the graph. Nevertheless, validity evidences showed worst results with lower levels of reliability and more items and persons with infit and outfit problems (Table 6, Figure 4). Despite only two items (loving, cold) have presented such aspect, infit and outfit indices were higher than with other scales.

TABLE 6

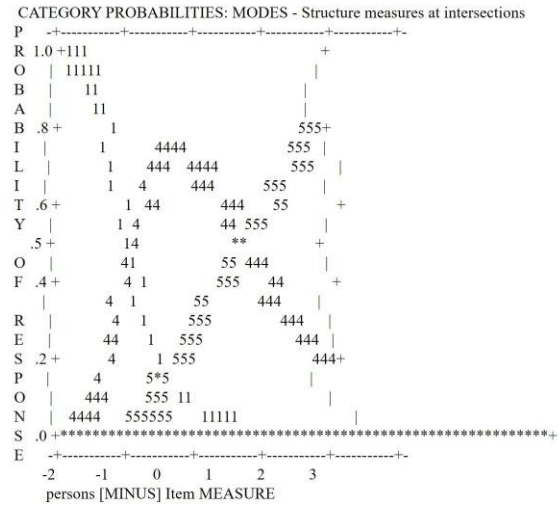
Structure of categories and probability chart categories for the socialization scale with tree categories

Structure of Categories and Probability Chart Categories for the Socialization Scale with Three Categories

Category score	Observerd Count	%	Observed avrge	Sample expect	Infit mnsq	Outfit mnsq	Structure calibrat	Categoy measure	Label
1	131	6	-0.14	-0.42	10.37	10.72	NONE	-10.34	1 disagreestrongly
2	0	0			0.00	0.00	NULL	-0.80	
3	0	0			0.00	0.00	NULL	-0.37	
4	1282	55	0.83	0.94	10.26	0.89	-10.56	0.27	4 partiallyagree
5	931	40	10.69	10.56	0.96	10.07	10.56	20.66	5 stronglyagree

Source: own work

Figure 4
Probability chart categories for the socialization scale with three categories



Source: Own work.

Within this item scale, 33% of people presented infit and outfit higher than 1.5, suggesting a larger amount of misfits than with 4 or 5-point scale. Also, the same items that presented DIF with other scales did so, but with one more item (Romantic, t=2.88, favoring women). The other tendencies were the same but with higher t values, suggesting greater misfits (Table 7). It is important to note that some adjectives in the socialization scale were misfits, presented DIF problems, or both such as Cold, Shy, and Delicate. These items showed unexpected agreement with higher categories when lower athletes' socialization was evidenced. Hence, they should be removed from this scale to assess such traits in athletes. Despite the three-category structure in the items was the one with better thresholds, outfit level, progression of socialization mean and most discriminative curves, its validity evidences were worse than the other two structures (with four and five levels) regarding DIF, model fit, and reliability making it unfeasible (Bond & Fox, 2001; Wright & Masters, 1982).

TABLE 7
Summary statistics of Rasch model for the Socialization items with 3-point scale

Summary Statistics of Rasch Model for the Socialization Items with 3 Point Scale

	Raw score	Count	Measure	Model error	Infit mnsq	Infit zstd	Outfit mnsq	Outfit zstd
People								
MEAN	750.7	170.9	10.11	0.36	10.32	0.5	10.20	0.3
S.D.	50.6	0.4	0.71	0.09	0.88	10.4	0.93	10.3
MAX.	870.0	180.0	30.27	0.65	40.25	40.8	60.95	50.0
MIN.	560.0	160.0	-0.28	0.19	0.18	-20.3	0.19	-10.6
Reliab.	0.73							
Items								
MEAN	5500.8	1300.2	0.00	0.14	10.08	0.5	10.19	0.5
S.D.	570.8	10.5	0.57	0.03	0.47	20.5	0.86	20.7
MAX.	6060.0	1310.0	10.56	0.19	20.45	90.2	40.52	90.9
MIN.	3430.0	1250.0	-10.04	0.07	0.54	-10.9	0.60	-20.0
Reliab.	0.94							

Source: own work

Indeed, the four-point scale was the best regarding validity evidences, despite similar amount of misfits, the coefficients were better than with the five-point scale as well as the category information with more adequate thresholds, outfit levels, progression of socialization mean, and good discriminative curves. Hence, this structure could be adopted in this scale aiming towards better socialization assessment in athletes. Category 4, partially agree, is the one that yielded worst thresholds and discrimination level but the level of socialization of the sample was low and this fact can be somehow expected. New studies could focus on higher athlete samples with higher levels of socialization to assure that this category can be adjusted aiming test standardization.

It is important to note that, although this analysis provides some guidelines regarding which categories have potential problems, it should be emphasized that the final decision to merge or delete a given category should be made not only based on statistical criteria but on assumptions provided in the variable under investigation. Furthermore, the optimization depends on the scales and sample to be tested again with a fresh sample of the same population (Smith et al., 2003). These preliminary evidences must be set into the Brazilian context, where no adequate instruments with good psychometric properties to assess athletes' personality are found (Bartholomeu et al., 2013; Brandão, 2007; Moraes, 2007). Further studies with this test (not only with the socialization scale, but all other five factors) can make it proper for use in the Brazilian context. Also, this test was first developed to assess personality traits in

college students, and comparisons on the Likert scale structure between these two samples can be valorous, since personality expression varies between contexts (Sinn & Moltschaniswskyj, 2005).

One possible explanation to these data can be that the use of adjectives to describe personality, despite being easier and quicker to use, they can be less informative than the use of phrases with context information of behaviors. Hence, new studies can investigate differences on category quantity in personality assessment by adjectives and phrases.

Final Considerations

The objectives of this study were to optimize scales of items in a scale of five great factors, socialization, assessed by means of adjectives, and administer them to a sample of athletes. The results showed that, although the structure of the items of three levels of response have better statistical properties shown with a progression of averages and thresholds, top anxiety, as well as outfit values and probability plot curves present the most discriminative for each of the categories. Wright and Masters (1982) point out that when comparing various forms of assessment items you must demonstrate an improvement in Indices of reliability and validity by means of adjustment measures, InFit, outfit, and differential item functioning. In other words, the new structure levels of the items should show a better functioning, which show as a reduction of errors and biases in the measurement and improvement in validity evidence. In this sense, it may point to the need to characterize and differentiate groups of responses in this type of instrument for this population and, thus, differentiate the different levels of performance that can bring a difference for this type of analysis, especially regarding personality characteristics.

References

Bartholomeu, D., Montiel, J. M., & Machado, A. A. (2013). Avaliação da Escala Likert dos itens do CSAI-2 em atletas. *Interação em Psicologia, 17* (1), 79-89.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rash Model: Fundamental measurement in the human sciences* . London: Erlbaum.

Brandão, M. R. F. (2007). A psicologia do exercício e do esporte e seus desafios para o milênio. In M. R. F. Brandão & A. A. Machado. *Coleção psicologia do esporte e do exercício: teoria e aplicação* (Vol. 1, Cap. 7, pp. 12-24). São Paulo: Atheneu.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Minnesota Multiphasic Personality Inventory – A (MMPI-A): Manual for administration, scoring, and interpretation* . Minneapolis: University of Minnesota Press.

Cattell, R. B., & Cattell, H. E. P. (1995). Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement, 55* , 926-937.

Cratty, B. J. (1984). *Psychology in contemporary sport* . Englewood Cliffs: Prentice-Hall.

Draba, R. E. (1977). *The identification and interpretation of item bias. Rasch Measurement Transactions* , MESA Memorandum, 25. Retrieved from http://www.rasch.org/rmt/rmt122m.htm

Goldberg, L. R. (1982). From ace to zombie: Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203-234). Hillsdale, NJ: Erlbaum.

Hutz, C. S., Nunes, C. H. S. S., Silveira, A. D., Serra, J., Anton, M., & Wieczorek, L. S. (1998). O desenvolvimento de marcadores para a avaliação da personalidade nomodelos dos Cinco Grandes Fatores. *Psicologia: Reflexão e Crítica , 11* , 395-409.

Linacre, J. M. (1997). *Guidelines for rating scales* . Retrieved from http://mesa.spc.uchicago.edu/rn2.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement* , *2* , 103-122.

Linacre, J. M. (2002). What do infit and outfit, mean-squared and standardized mean? *Rasch Measurement Transactions, 16* (2), 878. Retrieved from http://209.238.26.90/rmt/rmt82a.htm

McCrae, R. R., Costa, P. T., & Piedmont, R. L. (1993). Folk concepts, natural language, and psychological constructs: The California Psychological Inventory and the Five-Factor Model. *Journal of Personality, 61* , 1-26.

Moraes, L. C. C. A. (2007). Emoções no Esporte e na Atividade Física. In M. R. F. Brandão & A. A. Machado. *Coleção Psicologia do esporte e do exercício: Teoria e aplicação* (Vol. 1, Cap. 4, pp. 57-65). São Paulo: Atheneu.

Muñiz, J. (1990). *Teoría de respuesta a los ítens: Un nuevo enfoque en la evolución psicológica y educativa* . Madrid: Ediciones Pirámide, S. A.

Sinn, D. L., & Moltschaniwskyj, N. A. (2005). Personality traits in dumpling squid (Euprymna tasmanica): Context-specific traits and their correlation with biological characteristics. *Journal of Comparative Psychology, 119* (1), 99-110.

Smith, E. V., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63* (3), 369-391.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement* . Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1988). *Validity in Rasch measurement: Research memorandum 54* . Chicago: MESA Press.

Wright, B. D. (1992). *IRT in the 1990s: Which models work best? Rasch Measurement Transactions, 6,*196-200.

Notes

***** Research article