Psychometric Evaluation of the Big Five Questionnaire for Children(BFQ-C): A Rasch Model Approach

Marcos Cupani; Valeria E. Morán; Fernanda B. Ghío; Ana E. Azpilicueta; Sebastián J. Garrido
marcoscup@gmail.com
(Instituto de Investigaciones Psicológicas Conicet-UNC; Facultad de Psicología, Universidad Nacional de Córdoba)

Cupani, M., Morán, V.E., Ghío, F.B. *et al.* Psychometric Evaluation of the Big Five Questionnaire for Children (BFQ-C): A Rasch Model Approach. *J Child Fam Stud* **29,** 2472–2486 (2020). https://doi.org/10.1007/s10826-020-01752-y

Published
24 June 2020

Issue Date
September 2020

Abstract

The Big Five Questionnaire for Children, (BFQ-C) is an instrument for personality assessment in children and adolescents widely used worldwide. The aim of this work was to study the psychometric properties of the instrument scores from the item response theory (IRT) perspective. We worked with a Partial Credit Rasch Model to analyze an Argentinean sample to validate the scale for its use in this population. We opted for an instrumental design, and for each factor we applied an item calibration plan consisting of different analysis: unidimensionality, classification of response categories, fit levels of items and persons, specific objectivity, and differential item functioning as regards sex. We worked with a sample of 1162 high school students aged 12–17 years. The five original subscales did not show satisfactory fit, so modifications were made to improve their properties. As a result, we could demonstrate that each subscale measures a single latent trait, meets the invariance assumption regarding the sample and the assumption of local independence, showing no sex differential item functioning (DIF). Finally, the ordinal scores were converted to an interval scale, which allows more accurate analysis and better confidence in outcomes. Our results showed that the five subscales corresponding to each factor were in line with the IRT key parameters, although we suggest further studies on both the test capacity to assess extreme scores and the relevanceof using a five-response category scoring.

Keywords: BFQ-C; Item response theory; Partial Credit Rasch model; Argentinean samples; High school

Highlights

- The items of BFQ-C have adequate psychometric properties based on Rasch model.
- The items do not show differential functioning in terms of sex.
- The BFQ-C allows to measure the Big Five personality factors in Argentinean population.

In the field of personality psychology, the Big Five factor model is well recognized by its explicative capacity of the individual differences as regards personality in all the development stages of human beings (John et al. 2008). Specifically, the importance of studying personality during childhood and adolescence lies in its relationship with the behavior that might impact in the adult age in areas such as the academic performance, substance use, interpersonal relationships, work achievements, and criminal behavior (Damian et al. 2015; Goretti et al. 2017; Mitsopoulou and Giovazolias 2015; Morizot 2015; Poropat 2014).

Currently, one of the most complete and widely used instruments to measure the personality dimensions for populations of children and adolescents is the Big Five Questionnaire for Children (BFQ-C) developed by Barbar- anelli et al. (2003). These authors studied the personality structure of 7- to 14-year old Italian children through dif- ferent informants (self-report, parents, and teachers); their results using the exploratory factor analysis revealed a five-factor structure in all of the cases. These factors were named by the authors as Energy/Extraversion, Agreeableness, Conscientiousness, Neuroticism/Emotional instability, and Openness/Intellect.

This instrument became so popular that it was adapted to different countries, such as The Netherlands (Muris et al. 2005), Germany (Essau et al. 2006), Spain (Holgado et al. 2009), and Argentina (Cupani and Ruarte 2008), among others (Markos and Kokkinos 2017; Olivier and Herve 2015), showing in all of these studies its adequate reliability and validity properties. In most of these works, the assessment of the BFQ-C internal structure was performed by exploratory and confirmatory factor analyses, convergent-discriminant validity, and criterion-related validity considering disadaptive behavioral syndromes and academic performance as the external criteria (Barbaranelli et al. 2003; Cupani and Pautassi 2013; Cupani and Ruarte 2008; Zuffian et al. 2013). Factor invariance studies considering the participants' sex and age (Del Barrio et al. 2006) were also performed.

Although the analyses of the BFQ-C psychometric properties were satisfactory from the point of view of the Classical Test Theory (CTT), it is worth mentioning that this psychometric model presents constrains and dis- advantages (Abedalaziz and Leng 2018; An and Yung 2014). In principle, it has demonstrated to be useful in the development and validation of instruments, although the statistical indices (e.g. difficulty and discrimination of items) depend on the characteristics of the sample used. Conversely, the models based on the Item Response Theory (IRT) allow overcoming some of the CTT constrains. The IRT-based models can get invariant measurements (Engelhard 2013), which means that the item parameters do not depend on the sample features and the person parameters do not depend on the items selected for their assessment. Additionally, the calibration procedure is also independent of the individuals to whom the test is administered, and the person measurements are independent of the applied test (Cupani and Cortez 2016). On the other hand, the mea- surement precision is estimated for each ability level in the variable, emphasizing the item analysis and a person's ability level by emitting a response to each of them; this is the reason why the measurement error is calculated for each item and for each person (Cupani and Cortez 2016).

Another advantage of the IRT is that it allows calculating the measurement error for each item and for each person; because of that, the measurement precision is estimated for each ability level in the variable. Currently, the IRT is becoming more widely used than the CTT as a model to create and validate instruments (Embretson and Reise 2000). Over the last twenty years, the Rasch mathematical model (Rasch 1960) has been widely chosen to assess the quality of health measurement instruments in IRT (Leung et al. 2014). This model is characterized by its simplicity. In fact, it considers only one parameter to specify the properties of each item, which grants getting the same item difficulty order for all the ability levels. Moreover, persons and items are placed in the same scale interval, allowing the direct comparison of person and item measurements. Likewise, the Rasch model is able to analyze polytomous data through the Rating Scale Model (RSM; Andrich 1978) or the Partial Credit Model (PCM; Masters 1982). To apply the RSM, the scales should have gradual response categories and the distance between thresholds should be equal across items (Bond and Fox 2015). On the other hand, for the PCM the distance between the response categories does not have to be constant across the items (Tennant and Conaghan 2007).

In the past few years, the psychometric properties of the personality questionnaires have started being assessed from the IRT perspective (e.g., Maples-Keller et al. 2017; Nieto et al. 2017). In fact, a study on the NEO-FFI scales from the IRT in a sample of adolescents have suggested that several items considered as personality measurement components cannot be used as reliable indicators, which hinders the internal validity of the instrument. Specifically, in the BFQ-C, Markos and

Kokkinos (2017) analyzed the psy- chometric properties in Greek samples using IRT to deter- mine which items provided more information about each one of the latent traits (factors). Based on these results, theydeveloped a short version of the instrument that was later analyzed under the precepts of the CTT. Similarly, Bore et al. (2018) analyzed a reduced version of the instrument in Australian samples. The results showed that most of the items allowed to differentiate lower levels of each factor except for Emotional Instability, in which uncertainties were observed in the parameters and in the information levels of the response categories. These uncertainties were also detected in some items of the remaining factors.

So far, little is known about the individual item char- acteristics of the long version of the BFQ-C, their difficulty levels, and whether they fit to a measurement model. Little attention has been given to the rating scales to measure personality traits and the differential item functioning. In addition, although the cited studies are valuable backgroundon the analysis of the BFQ-C from the IRT, they do not provide complete information according to the current trendwhen reporting the results of the Rasch model analysis (Leung et al. 2014). Because of this drawback, the aim of this work was to examine the BFQ-C psychometric prop- erties in a sample of Argentinean students using the Rasch model (Rasch 1960) to determine whether the precision of the personality measurement can be maximized to be used in clinical and behavioral studies in this age range.

Method

Participants

Argentinean adolescents (606 females, 556 males), age range 12-17 years (M_{age} = 13.9 years, *SD* = 0.90), enrolled in state (34%) and private (66%) educational institutions from the city of Córdoba, Argentina. Córdoba is the second largest city in the country in terms of population (1,329,694 inhabitants according to the Argentina Population and Housing Census, 2010). Participants represented two grade levels from the Argentinean high school system: eighth (29.9%) and ninth (70.1%) grades. The sample was repre- sentative of upper-middle and lower-middle socio-eco- nomical classes, considering the characteristics of the institutions participating in this study (the students attending there belonged to families of skilled workers, large- production farmers, professionals, and local traders) andthe classification given by the National Institute of Statisticsand Censuses (INDEC, for its acronym in Spanish), Argentina. The study was approved by the General Office of Secondary Education from the city of Córdoba. A note was sent to the students' parents explaining the aim of the research and requesting signature for informed consent.

Procedure

The data collection process took four years and yielded a database exhibiting a comparable number of girls and boys,who were students attending state and private institutions. Students participated during classroom time; tests wereadministered collectively during a regular school day.
Teachers stayed in the class to help monitor the students' behavior. The researcher provided detailed instructions about how to complete the survey, and students had the chance to ask questions. Participants' parents, after being informed about the aim of the study, signed a consent formthat stated the aim, the voluntary participation, and the confidential nature of the data.

Measures

Big Five Questionnaire for Children (BFQ-C). The BFQ-C(Barbaranelli et al. 2003) measures the five personalityfactors in 9- to 15-year-old children; it consists of 65 items,13 items for each factor. The factors are: Extraversion (e.g., "I easily make friends"), Agreeableness (e.g., "I trust in others"), Conscientiousness (e.g., "I like to keep all my school things in order"), Neuroticism/Emotional instability (e.g., "I easily get angry") and Openness/Intellect (e.g., "I easily learn what I study at school"). For each of the 65 items, participants rated on a 5-point scale the occurrence ofthe behavior reported in the item using a 5-point Likert scale ranging from 1 (=Almost never) to 5 (=Almost always). The original instrument has adequate reliability and validity (Barbaranelli et al. 2003). The Spanish adapted version (Cupani and Ruarte 2008) used in the present study has acceptable internal consistency (α = 0.70–0.78), sub- stantial temporal stability after 2 months (r = 0.71–0.84), and evidence of internal structure validity through exploratory and confirmatory factor analyses (GFI 0.91; CFI 0.90; RMSEA 0.06).

Data Analysis

The data analysis was performed with RUMM2030 software (Andrich et al. 2010), following the guidelines proposed by Leung et al. (2014). Because it is recommended that each subscale establishes a different latent trait (Barbaranelli et al. 2003), the results were presented for each factor of the BFQ-C. The calibration plan for the items involved the followinganalysis.

Rasch model for polytomous items

The use of the Partial Credit Model (PCM) was determined through the log-likelihood ratio test, because the result of this analysis was significant ($p < 0.001$) in all subscales.

Thresholds

The functioning and structure of the response categories were examined. The order of the thresholds was inspected. The term threshold refers to the limit between two response categories where any response is likely. The increasing order of response options should be able to demonstrate higher levels of the measured trait. However, a disorder occurred in the thresholds when the level of trait being measured results inconsistent with respondents' response choices (Pallant and Tennant 2007). This occurred when there were too many response options or they were confusing and misinterpreted by persons. In these cases, the RUMM2030 software allowed us to col- lapse response categories and then replicate the analyses to see if those modifications improved the fit to the model (Parkitny et al. 2012).

Unidimensionality

To evaluate the unidimensionality of each subscale, we used the method proposed by Smith (2002). Through the Prin- cipal Components Analysis (PCA) of the residuals, we established the relationship between the items and the first residual factor. From this procedure, two subsets of items were delimited: a group consisting of items with positive residual loadings and another group with negative residual loadings on the first principal component (±0.30). A paired *t*-test analysis of the subsets was performed to verify whe- ther the estimates of the persons in each subset differed significantly from each other. The proportion of people with differences in their measurements in the two subgroups was obtained. For the instrument to be one-dimensional, the percentage outside the range ±1.96 should not exceed 5% (Tennant and Pallant 2006).

Local independence

We evaluated local independence through the matrix of residual correlations. Values > 0.2 indicate the presence of local dependence (Andrich et al. 2010). In case of depen- dence between items, those with values higher than 0.2 were combined, creating "superitems" (Nilsson and Tennant 2011). After conducting this type of modifications, the analysis was replicated to observe changes in the correlation matrix.

Item and person residual fits

From the item-person interaction analysis, we obtained a general summary of the variations of items and persons in relation to the fit to the model. If data are consistent with what is expected by the model, the mean (M) must be close to 0 and the standard deviation (SD) close to 1. We used the standardized residual statistic to observe the behavior of items and persons; those values between ±2.5 indicated a good fit to the model. Residual values outside the range andextreme values should be identified, reported, and, if necessary, the misfitting items or persons should be elimi- nated, since a few cases with anomalous response patterns can affect the fit indices (Tennant and Conaghan 2007).

Statistic of item-trait interaction

We used the chi-square (X^2) as a test of fit between the data and the model. We obtained an overall test of statistic fit (total item chi-square) and a chi-square statistic for each item. A significant chi-square and a Bonferroni adjustment less than α 0.05, indicates that the relative item difficulty is not constant across the trait, compromising the requirement of invariance (Tennant and Conaghan 2007). When data fit the model, the item-trait interaction has a low X^2 value with $p > 0.05$ (Cavanagh and Waugh 2011). Note that this sta- tistic is sensitive when the estimated probabilities are close to 0 or 1, because it is usually modified according to the sample size and the number of class intervals. The X^2 assessment refers to a perfect fit to the model, so it is usually considered only as an alternative measure to observe the location of the items with extreme values. Thus, we calculated the root mean square error of approximation (RMSEA) to examine fit (Tennant and Pallant 2012). Values < 0.02 suggest the data fit the Rasch model. We also performed an Analysis of Variance (ANOVA) to determine differences in class intervals. This analysis is a more precise indicator of the fit of the items tothe model (Andrich et al. 2010).

Reliability index

The Person Separation Index (PSI) indicates the degree to which the instrument differentiates persons in the constructbeing measured. A value of 0.70 is optimal for using the instrument in groups, whereas a value of 0.85 is optimal forindividual use (Tennant and Conaghan 2007).

Differential item functioning (DIF)

We analyzed the DIF according to the participants' sex. DIF exists when the functioning of an item differs in different groups of people (e.g., women/men). That is, the item has a different meaning for each group when it should be inter- preted in the same way (Tennant and Conaghan 2007). The DIF was identified through the characteristic curve of the item and was statistically confirmed from the ANOVA (Bonferroni lower than α 0.05). Two types of biases can be identified: the uniform DIF (the group shows a systematic difference in the responses given to an item across mea- surement range of the attribute) and the nonuniform DIF (there is no uniformity in the differences between the groups). The first type of DIF is solved by the split item procedure between groups (female and male). The non-uniform DIF is difficult to solve; in general, the itemspresent in it are eliminated (Tennant and Pallant 2006).

Results

The Partial Credit Model (PCM) was chosen, as the log- likelihood test proved to be significant across all subscales.

Agreeableness

The items of this subscale presented M = 0.76 and SD =1.98, whereas the persons had M = −0.35 and SD = 1.51. A significant chi-square was obtained (X^2 = 118.18; df = 39; $p < 0.00$; RMSEA = 0.04), suggesting a poor fit of some ofthe items. In fact, item-level analyses showed residual valuesof ±2.5 in items 16 and 51, whereas items 27, 32, and 38 presented significant X^2 and F values (Bonferroni at α 0.05;Table 1). Disordered thresholds were observed in items 11, 38, and 64. The residual correlation matrix showed novalues > 0.20. A uniform DIF was identified in items 16, 47, and 60. On a person-level analysis, four participants (0.3%)had extreme values and 129 ones (11.1%) had a residual value ±2.5. The PSI was 0.81. The proportion of significant *t*-tests outside the range ±1.96 was 6.71% (Table 2), there- fore the unidimensionality assumption is not satisfied.

From these results, modifications were made to achieve abetter fit of the scale. We deleted items 16, 51, 13, and 27 as well as those with residual values ±2.5 and extreme cases. A disordered threshold was still observed in item 64 and a uniform DIF in item 45; however, neither the collapse of categories nor the division of the DIF item was considered because the results did not improve the fit to the model. Thesubsequent analysis yielded M = 0.4 and SD = 0.81 for items and M = −0.21 and SD = 1.02 for persons, with a sample distribution of −2.4 to +3.6 logits (Fig. 1). The chi-square statistic was significant, with values implying a betterfit to the model (X^2 = 43.30, df = 27, $p < 0.05$; RMSEA = 0.02). A PSI value of 0.74 indicated an acceptable reliability index. All items fitted to the model. The assumption of localindependence and one-dimensionality was fulfilled (Per C < 5% = 4.51%; Table 2).

Openness/Intellect

We obtained M = 1.28 and SD = 2.65 for items and M =−0.24 and SD = 1.37 for persons. The value of the chi-square interaction was significant (X^2 = 288.86, df = 39; p < 0.00; RMSEA = 0.07). A PSI of 0.77 was obtained, 7% (83) of the persons presented residual values ±2.5, and a single extreme case was observed. Five items presented residual values outside the acceptable range of residual fit. Eight items showed a significant chi-square (Bonferroni α < 0.05) (Table 3). Disorder was observed in the thresholds in items 24, 10, 33, 36, 46, and 59. In the residual correlation matrix, three pairs of items obtained values greater than 0.20 (30-12, 62-12, and 62-30). Uniform DIF was observed in items 5, 10, 24, 33, 36, and 59. The proportion of significant *t*-tests was 16.01%.

We eliminated cases with residual values ±2.5 and extreme cases. The response categories of the items with disorder in the thresholds were recoded, reducing items 3, 4, 10, 33, 36, 46, and 59 to three response categories. We created a superitem between items 30 and 12 and eliminated items 36, 43, and 62. These modifications showed that M (items = 0.34, person = −0.20) and SD (items = 2.27, person = 0.97) approached acceptable values. In addition, it was graphically observed that the sample and the items presented an adequate distribution, with the exception of some cases (Fig. 2). The chi-squared value improved with respect to the initial analysis (X^2 = 49.48, df = 27, p < 0.01; RMSEA = 0.03) and the PSI was 0.71. No dependence was observed among the items, although there was evidence of disorder in the thresholds of the created superitem. Item 10 and the superitem (BFQ-C30-BFQ-C12) showed residual values outside the acceptable range, as well as item 18 that presented a significant *F* value. Items 5, 24, and 33 showed DIF-uniform. The assumption of one-dimensionality was checked (Per C < 5% = 4.06%; Table 2).

Energy/Extraversion

In the initial analysis, we obtained M = 0.80 and SD = 2.37 for items, and M = −0.26 and SD = 1.30 for persons. The item-trait interaction resulted in a significant chi-square value, which indicates that there are certain imbalances of the data with respect to the model (X^2 = 270.13, df = 39, p < 0.00; RMSEA = 0.07). The PSI index was 0.72. Three items showed Bonferroni test at α 0.05 (Table 4). Most of the items presented disorder in the thresholds. We observed that 6% (70) of the persons obtained a residual value ±2.5; a value greater than 0.2 was observed in the matrix of corre- lations of the residuals between items 23 and 26. Items 1, 9, 19, 40, and 55 showed uniform DIF. The one-dimensional assumption was fulfilled (Per C < 5% = 4.99%).

A new analysis was performed to improve the fit to the model of the Extraversion subscale. We eliminated items 9 and 40, persons with a residual fit outside the acceptable range, and extreme cases, and created a superitem with items 23 and 26. The results showed M = 0.64 and SD = 0.84 for the items with acceptable values as well as for persons (M = −0.21; SD = 1.03), with an approximate sample distribution of −1.6 to +3.5 logits (Fig. 3). The item-trait interaction fit showed significant chi-square values; however, these values were lower than in the initial analysis (X^2 = 72.61, df = 30, p < 0.01; RMSEA = 0.03). The PSI presented a value of .68 and did not meet the criteria established for X^2 and *F*. The assumption of local independence was fulfilled. Uniform DIF was observed in items 19 and 55; nevertheless, no improvements were obtained in the model when replicating the analysis separ- ating the sample into females and males. Regarding the disorder in the thresholds, the original response categories were maintained because the collapse of categories did not improve the fit to the model. The one-dimensional assumption was fulfilled (Per C < 5% = 1.38%; Table 2).

Emotional Instability

We obtained M = 0.88 and SD = 2.55 for items, and M = −0.29 and SD = 1.49 for persons. A significant chi-square was observed (X^2 = 202.45; df = 39; p < 0.00; RMSEA = 0.06), and the PSI was 0.79. In this first analysis, five items had residual values outside an acceptable range. Table 5 shows both significant X^2 and significant *F*. As regard participants, 8.8% (102) obtained residual values ±2.5. We did not observe order in the thresholds of any of the items. The assumption of local independence was fulfilled. Items 8, 17, 54, and 58 presented uniform DIF. Unacceptable values were observed in relation to the one-dimensional assumption (Per C < 5% = 7.23%).

We eliminated outlier cases and those with residual values outside the acceptable limits. Items 15 and 58 were also eliminated because they showed non-uniform DIF. We also removed item 61 because it presented residual pro- blems and significant chi-square. In addition, the response categories of items 4 and 31 were reduced to three points. The other items had a scale of four-point response cate- gories. We performed a new analysis in which we obtained M = 0.50 and SD = 1.87 for items, and M = −18 and SD = 1.01 for persons, with a sample distribution of −3.2 to +3.3 logits (Fig. 4). The model did not fit (X^2 = 89.67; df = 30; p < 0.00; RMSEA = 0.04), so it was necessary to perform a deeper analysis. The results of this procedure showed that item 49 had a residual value outside the acceptable range, and that X^2 and F were significant. In addition, item 54 presented a residual fit lower than 2.5, whereas item 29 presented a significant F value. A relia- bility index of 0.74 was obtained. Local dependence between the items could not be observed. Items 4, 8, 31, and 54 showed uniform DIF. The assumption of one- dimensionality was met (Per C < 5% = 2.46%; Table 2).

Conscientiousness

The residual fit yielded M = 1.21 and SD = 4.46 for items, and M = −0.27 and SD = 1.48 for persons. The item-trait interaction resulted in a significant chi-square (X^2 = 530.72; df = 91; p < 0.00; RMSEA = 0.06). Five items presented residual values ±2.5, and six items had a significant value of X^2 and F (Table 6). A good reliability index was obtained (PSI = 0.83). Out of the total sample, 9.5% (110) presented residual values higher or lower than 2.5. Disorder was observed in the thresholds in seven items and in local dependence between items 20 and 34. Items 20 and 28 presented uniform DIF. The proportion of significant *t*-tests with values outside the range ±1.96 was 8.95% (Table 2). We eliminated extreme cases and those with a residual fit outside the acceptable range. Items 22, 44, 56, and 65 were discarded; items 25 and 37 were collapsed into three cate- gories, and items 3 and 53 were collapsed into four ones. Hence, the new analysis showed an improvement in the values of items (M = 0.42; SD = 1.25) and persons (M = −0.22; SD = 1.07), with an approximate sample distribution of −3.8 to 3.6 logits (Fig. 5). The item-trait interaction showed a significant chi square (X^2 = 57.51, df = 27, p < 0.01; RMSEA = 0.03). Item 37 did not meet the criteria established with respect to the value of the residuals. On the other hand, item 3 showed significant X^2 and F values. A PSI of 0.78 was obtained. There was no evidence of dis- order in the thresholds. The assumption of local independence and unidimensionality was fulfilled (Per C < 5% = 2.76%; Table 2). Although item 28 presented a uniform DIF, it was not partitioned by sex; this modification did not improve the fit to the model. We decided to keep items 3 and 37 because the various tests did not improve the fit to the model.

Finally, ordinal-to-interval rescoring algorithms were generated for the five original subscales. Using the con- version table (Table 7), the ordinal scores resulting from the addition of item scores to domain scores could be converted to interval-level scores, making the domain scores suitable for parametric statistics.

Discussion

The validation of tests in different cultures is a common pro-cess in the field of psychometrics, not only to obtain more empirical evidence of the universality of the theories, but mainly to reduce measurement biases that may occur because of an incorrect use of the tests. The validity of intercultural studies can be threatened by methodological difficulties and semantic inconsistencies derived from the translation processes of the components of the measurement instruments (Sperber et al. 1994). In fact, the equivalence of the psychometric properties among the different languages of a test cannot be simply assumed by translating its components, since the meaning and interpretation of the items in different cultures can vary widely, even if they are written in the same language (Chahín-Pinzón et al. 2012; Geisinger 1994). When adapting an item, one must guarantee that the language is adequate in its linguistic and cultural aspects; the used vocabulary and writing style must be comparable with the original one as regard the level of difficulty, complexity, style, length, etc. (Chahín- Pinzón 2014; Ramada-Rodilla et al. 2013).

The BFQ-C is the first instrument specifically created to assess the personality styles in children and adolescents through a self-report measure according to the Big Five factor theory. Although it is a widely used questionnaire, psychometric studies of the long version of this scale from the IRT have not been performed yet, only studies in reduced versions have been reported (Bore et al. 2018; Markos and Kokkinos 2017). Because of this drawback, this work aimed at analyzing the individual functioning of each item, the fit to the model, and the used classification structure functioning through the Rasch model. Based on our results, we made proposals for the scale improvement and expanded its application at a transcultural level.

Initially, none of the five scales totally satisfied the fit criteria for the Rasch model. To achieve reasonably good fit, it was necessary to make some modifications, such as removing items or creating superitems, and resolve local dependency, which provides strong evidence of uni- dimensionality. Although removing items from existingscales is a controversial practice, it must be considered thatthe psychometric properties resulting from the analysis under the CTT assumptions are not directly replicable fromthe Rasch model, which is based on more stringent assumptions and criteria. In this sense, because the BFQ-C is an instrument whose psychometric properties are still under study and, therefore, they are not fully defined, to make substantial modifications to it is not only convenient but also not problematic, as it could be when considering anestablished questionnaire (Shea et al. 2009).

Unidimensionality was corroborated separately for each of the five personality dimensions, although previous stu- dies supported that BFQ-C fits to the five-factor model (Barbaranelli et al. 2003; Cupani and Ruarte 2008; Essauet al. 2006; Muris et al. 2005). It could be observed that in four out of the five factors, the assumptions were not met inthe original items, but were reached after the modifications, as mentioned above. These results support the compliance with the assumption of unidimensionality, although it is known that perfect unidimensionality is difficult to obtain in practice (Zickar and Broadfoot 2009).

Individual separation indices were adequate before and after making changes to the instrument (except for the Energy/Extraversion trait). However, the values suggest that the scale should be used in groups and would not be ade- quate enough to make decisions at individual level. The item difficulty parameters considering the five factors are located in the central zone of the continuum and do not cover the latent trait completely. Thus, the items might not allow measuring low and high levels of the personality traits in adolescents (Wright and Stone 2004). Therefore, we can assume that in a population of adolescents where the per- sonality traits are distributed in a normal way (Z~N [0,1]), the personality questionnaire might measure adequately 65% of the population (between ±1 score z). This indicates that for adolescents who have extreme levels in the latent trait that the instrument cannot assess, some items might beadded to assess extreme levels related to personality.

Concerning the original structure of assessment of the five-category responses, it could be seen that it was not adequate for three out of the five scales that allow assessing the questionnaire. Specifically, we observed that the distances among the consecutive thresholds were not wide enough todescribe different ranges in the measured variable. Reducingthe categories to a three-point scale is more effective (except for the Agreeableness trait) in providing more precise infor- mation about the latent trait. Twiss et al. (2016) indicated that it is usual to find problems with the functioning of the response categories when applying the Rasch model analysisto existing scales. About this drawback, Bond and Fox (2015)stated that a simple solution is the rescoring that reduces theresponse categories without modifying the original appear- ance of the test and respecting its construction process.

The DIF was, in general, uniform; it could be assumed that the scales work in a similar way for men and women, which agrees with other studies that found that the items ofthe scale are invariant between men and women (Del Barrio et al. 2006). Nevertheless, two items showed no uniform DIF, representing 3% of the entire instrument.

When using probabilistic models, Rasch's analysis makes possible to determine the degree to which the items of a scale function as interval measurements of a latent trait (Tennant et al. 2004). When data fit this model, it is possible to assume that the test scores meet basic assumptions of psychological measurement such as unidimensionality and invariance. When assessed using ordinal scales, the measures obtained not only do not meet these criteria but are not enough precise and suitable for the calculation of parametric statistics (Medvedev et al. 2018). The ordinal to interval conversion provided for each subscale in this study allowed us to obtain continuous scores, which are feasible to be used for complex arithmetic operations. In addition, continuous scales guarantee that the distances between the values are equal and constant, making the measure more reliable and less ambiguous (Granberg-Rademacker 2010; Harwell and Gatti 2001).

Practical Implications

This research is of great value since it contributes to clinical psychology by presenting a useful instrument for this field. Also, based on the fact that it is a self-administered test thatcan be applied in groups, it is a valuable resource to be considered to promote local research on the personality field. On the other hand, the performed Rasch model ana- lyses indicate the items can be included in specialized programs to develop computerized versions of the tests, facilitating shorter measurements without losing precision and reliability (Abad et al. 2010).

Limitations and Future Research Directions

It is worth mentioning that the results obtained in the pre- sent study should be analyzed considering a series of con- straints. Firstly, the data analyses were performed on the basis of a sample of 12- to 17-year-old students, and no children were considered in this study. In fact, this ques- tionnaire can also be used for 9- to 12-year-old children; because of this, the analyses should be replicated in this population to establish whether the items behave similarly. It should also be mentioned that complementary studies such as focus groups were not performed, to establish an explanation of the reasons why some items presented DIF. It should be analyzed whether the responses informed by participants were related to their attribute level or responded to what is socially accepted (for example, in the item “I weep”). Finally, considering that the internal validity study was based on the fit model of each dimension separately, future research studies should complement these analyses applying a one-parameter multidimensional model (Briggs and Wilson 2003) as well as other IRT-based models, such as the Samejima two-parameter model (Samejima 1997).

In summary, through the Rasch model we obtained relevant results about the BFQ-C psychometric properties; this ques- tionnaire is one of the most widely used psychological assessment tools to assess personality traits in children and adolescents. However, it should be mentioned that the mod- ifications suggested in this study only constitute an improve- ment proposal to optimize its functioning, and that further study is needed. One way to expand these results, according to what the authors of the test propose, is to compare the results obtained in the sample of Argentinean adolescents with sam- ples from other countries, such as Spain or Italy.

Acknowledgement. This work was supported by the National Scien- tific and Technical Research Council—Argentina (CONICET for its acronym in Spanish).

Author Contributions. M.C. designed and executed the study, and con- tributed in writing all parts of the manuscript. V.E.M. analyzed the data and wrote and edited all parts of the manuscript. F.B.G. collaborated in the data collection, assisted with the data analyses and the writing and editing of the final manuscript. A.E.A. and S.J.G. collaborated with data collection and writing the literature review and discussion.

Compliance with Ethical Standards

Conflict of Interest. The authors declare that they have no conflict of interest.

Ethical Approval. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the Declaration of Helsinki (1964) and its later amendments or comparable ethical standards. The National University of Cordoba provided IRB approval for this study.

Informed Consent. It was obtained from all participants in the study and their parents.

References

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT (Deterioration of the para- meters of the items in computerized adaptive tests: study with eCAT). *Psicothema*, *22*(2), 340–347. http://goo.gl/fAjIj6.

Abedalaziz, N., & Leng, C. H. (2018). The relationship between CTT and IRT approaches in analyzing item characteristics. *Malaysian Online Journal of Educational Sciences, 1*(1), 64–70. https://bit. ly/2Z2IgKC.

An, X., & Yung, Y.F. (2014). *Item response theory: what it is and how you can use the IRT procedure to apply it*. Cary: SAS Institute. SAS364-2014. https://bit.ly/2WzVfBJ.

Andrich, D. (1978). Rating formulation for ordered response cate- gories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/ bf02293814.

Andrich, D., Sheridan, B. & Luo, G. (2010). *Rasch models for mea- surement: RUMM2030*. Perth: RUMM Laboratory Pty Ltd.

Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences*, *34*(4), 645–664. https:// doi.org/10.1016/S0191-8869(02)00051-X.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: fun- damental measurement in the human sciences* (3rd ed.). New York: Routledge/Taylor & Francis Group.

Bore, M., Laurens, K. R., Hobbs, M. J., Green, M. J., Tzoumakis, S., Harris, F., & Carr, V. J. (2018). Item response theory analysis of the Big Five Questionnaire for Children–Short Form (BFC-SF): a self-report measure of personality in children aged 11–12 years. *Journal of Personality Disorders*, 1–24. https://doi.org/10.1521/pedi_2018_32_380.

Briggs, D. C., & Wilson, M. (2003). An introduction to multi- dimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100. https://bit.ly/2LCIoZ7.

Cavanagh, R. F., & Waugh, R. F. (Eds.). (2011). *Applications of rasch measurement in learning environments research* (Vol. 2). Rot- terdam: Springer Science & Business Media. https://doi.org/10. 1007/978-94-6091-493-5.

Chahín-Pinzón, N. (2014). Aspectos a tener en cuenta cuando se realiza una adaptación de test entre diferentes culturas (Aspects to take into account when performing a test adaptation between different cultures). *Psychología: Avances de la Disciplina*, *8*(2), 109–112. https://doi.org/10.21500/19002386.1225.

Chahín-Pinzón, N., Lorenzo-Seva, U., & Vigil-Colet, A. (2012). Características psicométricas de la adaptación colombiana del Cuestionario de Agresividad de Buss y Perry en una muestra de preadolescentes y adolescentes de Bucaramanga (Psychometric characteristics of the Colombian adaptation of the Buss and Perry Aggression Questionnaire in a sample of preadolescents and adolescents from Bucaramanga). *Universitas Psychologica*, *11*(3), 979–988. https://bit.ly/3cEsysU https://bit.ly/3cEsysU.

Cupani, M., & Cortez, F. D. (2016). Análisis psicométricos del Subtest de Razonamiento Numérico utilizando el Modelo de Rasch (Psychometric analysis of the subtest of numerical reasoning using the Rasch model). *Revista de psicología (Santiago)*, *25*(2), 1–16. https://doi.org/10.5354/0719-0581.2016.44558.

Cupani, M., & Pautassi, R. M. (2013). Predictive contribution of personality traits in a sociocognitive model of academic perfor- mance in mathematics. *Journal of Career Assessment*, *21*(3), 395–413. https://doi.org/10.1177/1069072712475177.

Cupani, M., & Ruarte, M. (2008). Propiedades psicométricas del Cuestionario de los Cinco Factores para Niños (BFQ-C) en una muestra de adolescentes argentinos (Psychometric properties of the Five-Factor Questionnaire for Children (BFQ-C) in a sample of Argentine adolescents). *Estudios de Psicología*, *29*(3), 351–364. https://doi.org/10.1174/021093908786145421.

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., & Roberts, B. W. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adult- hood. *Journal of Personality and Social Psychology*, *109*(3), 473 https://doi.org/10.1037/pspp0000024.

Del Barrio, M. V., Carrasco, M. A., & Holgado, P. (2006). *BFQ-NA cuestionario de los Cinco Grandes para niños y adolescentes (adaptación a la población española) [BFQ-NA questionnaire of the Big Five for children and adolescents (adaptation to the Spanish population)]*. Madrid: TEA.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: LEA.

Engelhard Jr., G. (2013). *Invariant measurement: using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*(4), 454–469. https://doi.org/10.1177/1073191106287354.

Geisinger, K. F. (1994). Cross-cultural normative assessment: trans- lation and adaptation issues influencing the normative inter- pretation of assessment instruments. *Psychological Assessment*, *6*(4), 304–312. https://doi.org/10.1037//1040-3590.6.4.304.

Goretti, S., Sanchéz, M. S., Borja, P. L., Rivera, G. B., & Lara, M. R. (2017). The relationship between personality disorders and sub- stance abuse disorders. *European Psychiatry*, *41*, S473–S474. https://doi.org/10.1016/j.eurpsy.2017.01.547.

Granberg-Rademacker, J. S. (2010). An algorithm for converting ordinal scale measurement data to interval/ratio scale. *Educa- tional and Psychological Measurement*, *70*(1), 74–90. https://doi.org/10.1177/0013164409344532.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131. https://doi.org/10.3102/00346543071001105.

Holgado, F. P., Carrasco, M. Á., del Barrio, M. V., & Chacón, S. (2009). Factor analysis of the Big Five Questionnaire using polychoric correlations in children. *Quality & Quantity*, *43*(1), 75–85. https://doi.org/10.1007/s11135-007-9085-3.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: history, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: theory and research* (pp.114–158). New York: Guilford Press.

Leung, Y.-Y., Png, M.-E., Conaghan, P., & Tennant, A. (2014). A systematic literature review on the application of Rasch analysis in musculoskeletal disease—a special interest group report of OMERACT 11. *The Journal of Rheumatology*, *41*(1), 159–164. https://doi.org/10.3899/jrheum.130814.

Maples-Keller, J. L., Williamson, R. L., Sleep, C. E., Carter, N. T., Campbell, W. K., & Miller, J. D. (2017). Using item response theory to develop a 60-item representation of the NEO PI–R using the international personality item pool: development of the IPIP–NEO–60. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2017.1381968.

Markos, A., & Kokkinos, C. M. (2017). Development of a short form of the Greek Big Five Questionnaire for Children (GBFQ-C-SF): validation among preadolescents. *Personality and Individual Differences*, *112*, 12–17. https://doi.org/10.1016/j.paid.2017.02.045.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/bf02296272.

Medvedev, O. N., Krägeloh, C. U., Titkova, E. A., & Siegert, R. J. (2018). Rasch analysis and ordinal-to-interval conversion tables for the depression, anxiety and stress scale. *Journal of Health Psychology*. https://doi.org/10.1177/1359105318755261.

Mitsopoulou, E., & Giovazolias, T. (2015). Personality traits, empathy and bullying behavior: a meta-analytic approach. *Aggression and Violent Behavior*, *21*, 61–72. https://doi.org/10.1016/j.avb.2015.01.007.

Morizot, J. (2015). 10 The Contribution of temperament and personality traits to criminal and antisocial behavior development and desis- tance. In J. Morizot & L. Kazemian (eds), *The development of criminal and antisocial behavior* (pp. 137–165). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-08720-7. Muris, P., Meesters, C., & Diederen, R. (2005). Psychometric properties of the Big Five Questionnaire for Children (BFQ-C) in a Dutch sample of young adolescents. *Personality and Individual Differences*, *38*(8), 1757–1769. https://doi.org/10.1016/j.paid.2004.11.018.

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema*, *29*(3), 390–395. https://bit.ly/2T7bJiw.

Nilsson, L. A., & Tennant, A. (2011). Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. *Journal of Rehabilitation Medicine*, *43*(10), 884–891. https://doi.org/10.2340/16501977-0871.

Olivier, M., & Herve, M. (2015). The Big Five Questionnaire for Children (BFQ-C): a French validation on 8 to 14-year-old children. *Personality and Individual Differences*, *87*, 55–58. https://doi.org/10.1016/j.paid.2015.07.030.

Pallant, J., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychol- ogy*, *46*, 1–18. https://doi.org/10.1348/014466506X96931.

Parkitny, L., McAuley, J. H., Walton, D., Costa, L. O. P., Refshauge, K. M., Wand, B. M., Di Pietro, F., & Moseley, G. L. (2012). Rasch analysis supports the use of the depression, anxiety, and stress scales to measure mood in groups but not in individuals with chronic low back pain. *Journal of Clinical Epidemiology*, *65*(2), 189–198. https://doi.org/10.1016/j. jclinepi.2011.05.010.

Poropat, A. E. (2014). Other-rated personality and academic perfor- mance: evidence and implications. *Learning and Individual Dif- ferences*, *34*, 24–32. https://doi.org/10.1016/j.lindif.2014.05.013. Ramada-Rodilla, J. M., Serra-Pujadas, C., & Delclós-Clanchet, G. L. (2013). Adaptación cultural y validación de cuestionarios de salud: revisión y recomendaciones metodológicas (Cultural adaptation and validation of health questionnaires: review and methodological recommendations). *Salud pública de México*, *55*, 57–66. https://doi.org/10.1590/s0036-36342013000100009.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Danish Institute for Educational Research.

Samejima, F. (1997). Graded response model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Shea, T. L., Tennant, A., & Pallant, J. F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry*, *9*(1). https://doi.org/10.1186/1471-244x-9-21.

Smith, E. V. (2002). Detecting and evaluating the impact of multi- dimensionality using item fit statistics and principal component ana-lysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation. *Journal of Cross-Cultural Psychology*, *25*(4), 501–524. https://doi.org/10.1177/0022022194254006.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied and what should one look for in a Rasch paper? *Arthritis Rheum*, *57*(8), 1358–1362. https://doi.org/10.1002/art.23108.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, *7*, S22–S26. https://doi.org/10.1111/j.1524-4733.2004.7s106.x.

Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions*, *20*, 1048–1051.

Tennant, A., & Pallant, J. F. (2012). The root mean square error of

approximation (RMSEA) as a supplementary statistic to deter- mine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions*, *25*, 1348–1349.

Twiss, J., McKenna, S. P., Graham, J., Swetz, K., Sloan, J., & Gomberg-Maitland, M. (2016). Applying Rasch analysis to evaluate measurement equivalence of different administration formats of the Activity Limitation scale of the Cambridge Pul- monary Hypertension Outcome Review (CAMPHOR). *Health and Quality of Life Outcomes*, *14*(1). https://doi.org/10.1186/ s12955-016-0462-2.

Wright, B. D., & Stone, M. H. (2004). *Making measures*. Chicago: Phaneron Press.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C.

E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: doctrine, verity, and fable in the organi-zational and social sciences* (pp. 37–59). New York: Routledge.

Zuffian, A., Alessandri, G., Gerbino, M., Luengo Kanacri, B. P., Di

Giunta, L., Milioni, M., & Caprara, G. V. (2013). Academic achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learning and Individual Differences*, *23*(1), 158–162. https://doi.org/10.1016/j.lindif.2012.07.010.

Table 1 Fit index of the items of the Agreeableness subscale (initial analysis)

Item statistics	Location (logits)	*SE*	Fit residual	X^2	X^2 (*p*)	*F*	*F* (*p*)
(2) I share my things with other people	−0.21	0.03	0.90	4.56	0.207	1.64	0.179
(11) I behave correctly and honestly with others	−0.24	0.03	0.04	4.46	0.216	1.35	0.256
(13) I understand when others need my help	−0.24	0.03	1.78	0.84	0.841	0.24	0.869
(16) I like to give gifts	0.35	0.03	4.51	14.82	0.002	5.41	0.001
(21) If someone commits an injustice to me, I forgive her/him	0.12	0.03	1.18	3.23	0.358	1.21	0.305
(27) I treat my peers with affection	−0.14	0.03	−1.96	25.48	0.000	10.15	0.000
(32) I behave with others with great kindness	−0.004	0.03	−1.26	17.28	0.001	6.75	0.000
(38) I am polite when I talk with others	−0.29	0.03	−1.80	19.01	0.000	7.63	0.000
(45) If a classmate has some difficulty I help her/him	−0.15	0.03	−1.09	5.80	0.122	2.34	0.072
(47) I trust in others	0.11	0.03	1.98	2.91	0.406	0.79	0.500
(51) I treat kindly also persons who I dislike	0.33	0.03	3.33	12.16	0.007	3.77	0.010
(60) I think other people are good and honest	0.26	0.03	0.36	1.84	0.607	0.65	0.580
(64) I let other people use my things	0.10	0.03	1.93	5.80	0.122	2.3	0.076

In bold type, the items that did not meet the criteria established in some indices
SE standard error, X^2 chi-square

Table 2 Summary of the fit of the subscales of the BFQ-C initial and final versionsResidual fit ofitems
Residual fit ofpersons

Item–Trait Interaction				PSI	Unidimensionality			
	M	*SD*	*M*	*SD*	X^2 (*df*)	*p*	*RMSEA*	Per C < 5%
greeableness								
Initial	0.76	1.98	−0.35	1.51	118.18 (39)	0.000	0.04	0.816.71%
Final	0.40	0.81	−0.21	1.02	43.30 (27)	0.024	0.02	0.744.51%
Openness/Intellect								
Initial	1.28	2.65	−0.24	1.37	288.86 (39)	0.000	0.07	0.7716.01%
Final	0.34	2.27	−0.20	0.97	49.48 (27)	0.005	0.03	0.714.06%
Energy/Extraversion								
Initial	0.80	2.37	−0.26	1.30	270.13 (39)	0.000	0.07	0.724.99%
Final	0.64	0.84	−0.21	1.03	72.61 (30)	0.000	0.03	0.661.38%
Emotional Instability								

Initial	0.88	2.55	−0.29	1.49	202.45 (39)	0.000	0.06	0.797.23%
Final	0.50	1.87	−0.18	1.01	89.67 (30)	0.000	0.04	0.742.46%
Conscientiousness								
Initial	1.21	4.46	−0.27	1.48	530.72 (91)	0.000	0.06	0.838.95%
Final	0.42	1.25	−0.22	1.07	57.51 (27)	0.001	0.03	0.782.76%

M mean, *SD* standard deviation, *df* degrees of freedom, *PSI* person separation index, *per C < 5%* proportion of *t*-tests that were significant at thelevel of significance of 0.05

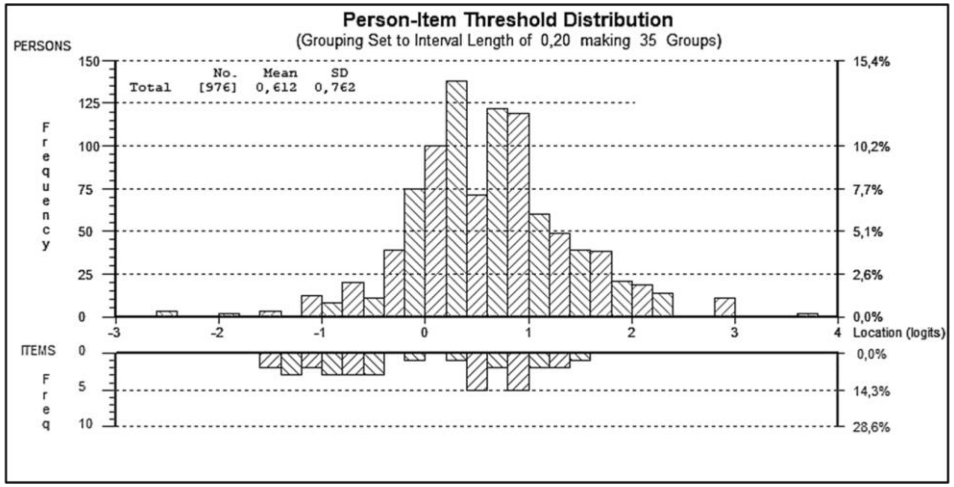

Fig. 1 Person-item threshold map. Final version (agreeableness)

Table 3 Fit index of the items of the Openness/Intellect subscale (initial analysis)

Item statistics	Location (logits)	*SE*	Fit residual	X^2	X^2 (*p*)	*F*	*F* (*p*)
(5) I know many things	−0.14	0.03	0.03	12.84	0.005	4.72	0.003
(10) I have a great deal of fantasy	−0.03	0.03	4.11	17.50	0.001	5.38	0.001
(12) I easily learn what I study at school	−0.19	0.03	−1.38	39.10	0.000	15.96	0.000
(18) When the teacher asks questions I am able to answer correctly	0.12	0.03	−0.41	25.47	0.000	9.55	0.000
(24) I like to read books	0.54	0.03	2.26	16.35	0.001	5.33	0.001
(30) When the teacher explains something I understand it immediately	0.04	0.03	−2.21	43.46	0.000	18.75	0.000
(33) I like scientific TV shows	0.44	0.03	3.38	4.57	0.206	1.39	0.244

(36) I like to watch TV news and to know what happens in the world	0.12	0.03	5.08	19.91	0.000	6.66	0.000
(43) I am able to create new games and entertainments	0.14	0.03	4.24	10.17	0.017	3.26	0.021
(46) I am able to solve mathematics problems	0.29	0.03	3.13	2.87	0.411	0.92	0.433
(52) I like to know and to learn new things	−0.49	0.03	−1.75	26.04	0.000	10.73	0.000
I would like very much to travel and to know the habits of people from other countries	−0.66	0.03	2.25	25.58	0.000	7.30	0.000
(62) I understand immediately	−0.17	0.03	−1.98	44.99	0.000	19.30	0.000

In bold type, the items that did not meet the criteria established in some indices
SE standard error, X^2 chi-square

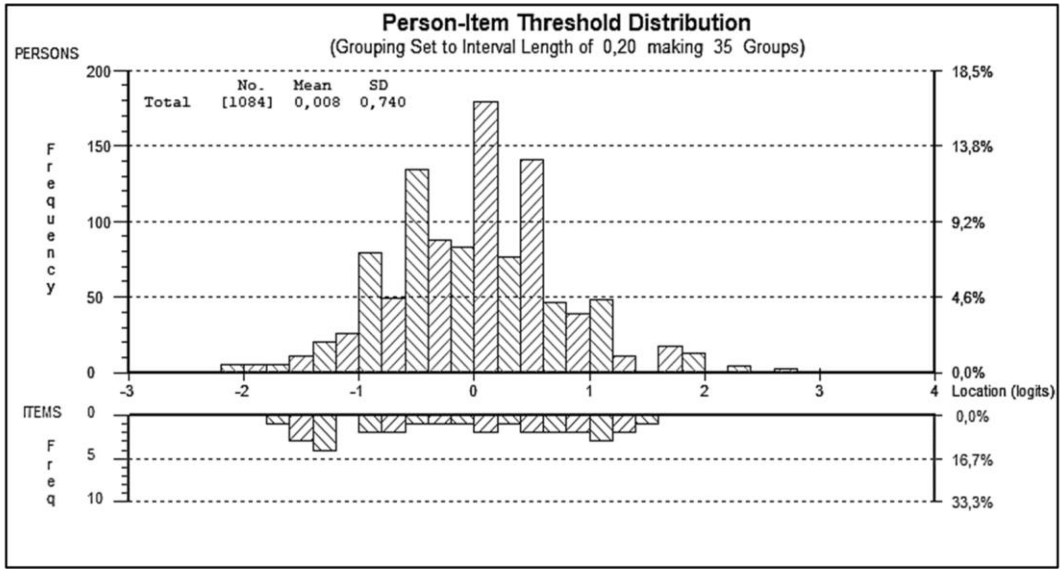

Fig. 2 Person-item threshold map. Final version (openness/intellect)

Table 4 Fit index of the items of the Energy/Extraversion subscale (initial analysis)

Item statistics	Location (logits)	*SE*	Fit residual	X^2	X^2 (*p*)	*F*	*F* (*p*)
(1) I like to meet with other people	−0.52	0.03	−0.35	10.16	0.017	2.49	0.059
(9) I like to compete with others	0.84	0.02	7.25	122.21	0.000	33.85	0.000

I like to move and to do a great deal of activity	−0.10	0.03	0.53	4.53	0.209	1.07	0.359
(19) I like to be with others	−0.26	0.03	−0.28	11.16	0.011	3.01	0.029
(23) I can easily tell others what I think	0.30	0.03	2.55	1.10	0.778	0.24	0.869
(26) I say what I think	0.07	0.03	0.59	9.56	0.023	2.87	0.035
(35) I do something not to get bored	−0.29	0.03	0.49	4.19	0.242	0.53	0.664
(40) I like to talk with others	−0.49	0.03	−2.74	48.46	0.000	18.08	0.000
I am able to convince someone of whatI think	0.03	0.03	−0.63	13.19	0.004	4.99	0.002
(50) When I speak, the others listen to me and do what I say	0.75	0.03	2.03	7.64	0.054	2.47	0.061
(55) I like to joke	0.09	0.03	1.71	1.62	0.654	0.25	0.863
(57) I easily make friends	−0.02	0.03	−0.84	24.91	0.000	8.55	0.000
(63) I am happy and lively	−0.40	0.03	0.17	11.38	0.010	3.01	0.029

In bold type, the items that did not meet the criteria established in some indices

SE standard error, X^2 chi-square

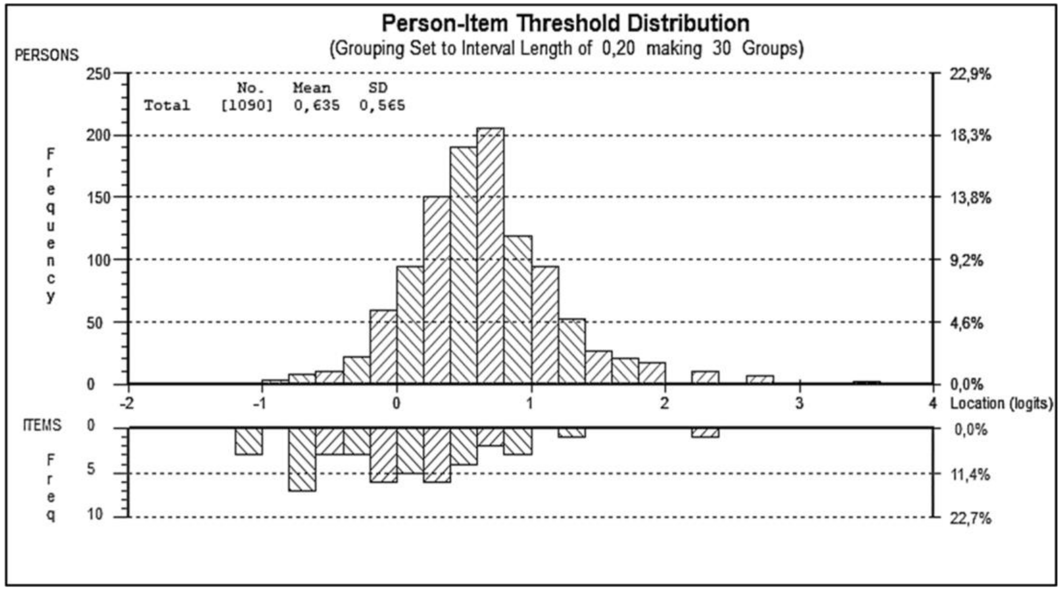

Fig. 3 Person-item threshold map. Final version (extraversion)

Table 5 Fit index of the items of the Neuroticism/Emotional instability subscale (initial analysis)

Item statistics	Location (logits)	*SE*	Fit residual	X^2	X^2 (*p*)	*F*	*F* (*p*)
(4) I get nervous for silly things	0.05	0.02	1.27	4.09	0.252	1.24	0.294
(6) I am in a bad mood	0.48	0.03	−0.78	13.75	0.003	4.68	0.003
(8) I argue with others with excitement	0.18	0.03	1.12	1.34	0.719	0.35	0.786
(15) I easily get angry	−0.07	0.03	−3.73	51.29	0.000	24.07	0.000
(17) I quarrel with others	0.54	0.03	0.51	6.08	0.108	1.89	0.129
(29) I easily get offended	0.10	0.03	−1.80	22.68	0.000	9.60	0.000
(31) I am sad	0.48	0.03	1.45	12.39	0.006	3.90	0.009
If I want to do something, I am not capable of waiting and I have to do it immediately	−0.75	0.03	4.28	20.89	0.000	6.67	0.000
(41) I am not patient	−0.59	0.03	0.12	3.30	0.347	1.15	0.326
(49) I easily lose my temper	0.03	0.03	−2.20	27.01	0.000	12.44	0.000
(54) I do things with agitation	−0.15	0.03	3.89	12.20	0.007	4.03	0.007
(58) I weep	0.05	0.02	3.13	17.46	0.001	5.64	0.001
(61) I worry about silly things	−0.34	0.03	4.15	9.97	0.019	3.35	0.018

In bold type, the items that did not meet the criteria established in some indices
SE standard error, X^2 chi-square

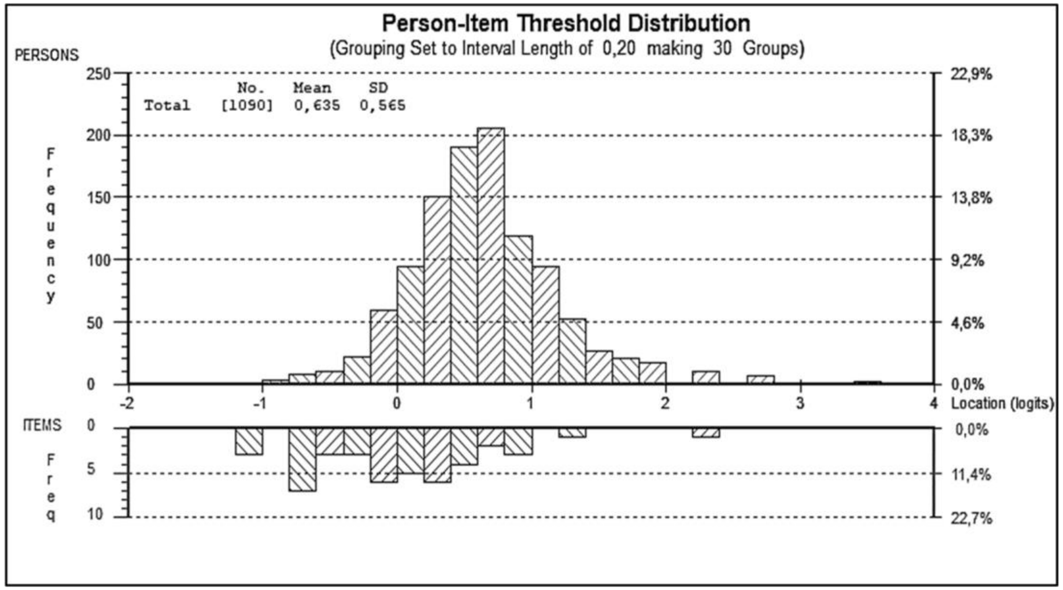

Fig. 4 Person-item threshold map. Final version (emotional instability)

Table 6 Fit index of the items of the Conscientiousness subscale (initial analysis)

Item statistics	Location (logits)	*SE*	Fit residual	X^2	X^2 (*p*)	*F*	*F* (*p*)
(3) I do my job without carelessness and inattention	0.27	0.03	−2.59	51.13	0.000	9.25	0.000
(7) I work hard and with pleasure	0.10	0.03	−0.38	15.37	0.031	2.45	0.017
(20) I engage myself in the things I do	−0.78	0.03	−0.36	15.54	0.030	2.50	0.015
(22) During class-time I am concentrated on the things I do	−0.05	0.03	−2.49	66.99	0.000	12.59	0.000
When I finish my homework, I check it many times to see if I did it correctly	0.95	0.03	−0.41	8.68	0.276	1.30	0.249
(28) I respect rules and order	−0.36	0.03	0.73	8.29	0.307	1.224	0.286
(34) If I engage in something I commit myself to it	−0.87	0.03	1.16	9.48	0.220	1.37	0.215
(37) My room is in order	0.08	0.02	5.24	30	0.000	3.46	0.001
(44) When I start to do something I have to finish it at all costs	−0.15	0.03	8.73	85.89	0.000	10.79	0.000
(48) I like to keep all my school things in order	0.10	0.03	−0.64	13.15	0.068	2.19	0.033
(53) I play only when I finish my homework	0.45	0.03	−0.12	7.29	0.399	1.13	0.339
(56) It is unlikely that I divert my attention	0.32	0.05	11.03	157.66	0.000	16.14	0.000
(65) I do my own duty	−0.07	0.03	−4.12	61.29	0.000	12.01	0.000

In bold type, the items that did not meet the criteria established in some indices
SE standard error, X^2 chi-square

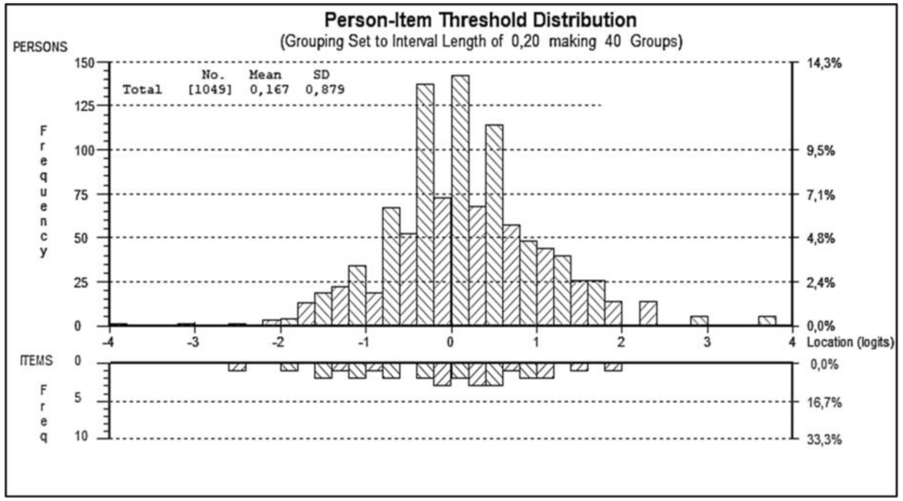

Fig. 5 Person-item threshold map (conscientiousness)

Table 7 Ordinal-interval conversion

Raw score	Agreeableness		Openness/Intellect		Energy/Extraversion		Emotional Instability		Conscientiousness	
Ordinal	Logit	Interval	Logit	Interval	Logit	Interval	Logit	Interval	Logit	Interval
13	−4.79	13.00	−4.46	13.00	−4.38	34.27	−4.41	13.00	−4.68	13.00
14	−3.62	19.19	−3.32	19.56	−3.24	35.41	−3.24	20.04	−3.51	19.39
15	−2.96	22.67	−2.69	23.19	−2.61	36.04	−2.59	23.98	−2.86	22.97
16	−2.58	24.69	−2.33	25.30	−2.25	36.40	−2.22	26.25	−2.49	25.05
17	−2.31	26.12	−2.07	26.80	−2.00	36.65	−1.95	27.86	−2.22	26.53
18	−2.10	27.24	−1.87	27.98	−1.80	36.85	−1.75	29.11	−2.00	27.69
19	−1.92	28.16	−1.70	28.96	−1.64	37.01	−1.58	30.13	−1.83	28.66
20	−1.77	28.97	−1.55	29.80	−1.50	37.15	−1.43	31.01	−1.68	29.49
21	−1.64	29.68	−1.42	30.55	−1.38	37.28	−1.30	31.78	−1.54	30.23
22	−1.52	30.32	−1.31	31.22	−1.27	37.39	−1.19	32.47	−1.42	30.90
23	−1.40	30.91	−1.20	31.84	−1.17	37.49	−1.09	33.10	−1.31	31.52
24	−1.30	31.47	−1.10	32.41	−1.07	37.58	−0.99	33.68	−1.20	32.09
25	−1.20	31.99	−1.01	32.95	−0.99	37.67	−0.90	34.22	−1.11	32.63
26	−1.11	32.48	−0.92	33.45	−0.90	37.75	−0.82	34.73	−1.01	33.14
27	−1.02	32.96	−0.84	33.93	−0.82	37.83	−0.74	35.21	−0.92	33.63
28	−0.93	33.42	−0.76	34.39	−0.75	37.90	−0.66	35.66	−0.84	34.10
29	−0.85	33.87	−0.68	34.83	−0.68	37.98	−0.59	36.10	−0.76	34.55
30	−0.76	34.30	−0.61	35.25	−0.61	38.05	−0.52	36.52	−0.68	34.99
31	−0.68	34.73	−0.54	35.67	−0.54	38.12	−0.45	36.93	−0.60	35.41
32	−0.60	35.15	−0.47	36.07	−0.47	38.18	−0.39	37.32	−0.52	35.83
33	−0.52	35.57	−0.40	36.47	−0.41	38.25	−0.33	37.70	−0.45	36.24
34	−0.45	35.99	−0.33	36.86	−0.34	38.31	−0.26	38.08	−0.38	36.64
35	−0.37	36.40	−0.27	37.24	−0.28	38.38	−0.20	38.45	−0.30	37.04
36	−0.29	36.81	−0.20	37.62	−0.21	38.44	−0.14	38.81	−0.23	37.44
37	−0.21	37.22	−0.14	38.00	−0.15	38.50	−0.08	39.17	−0.16	37.83
38	−0.13	37.64	−0.07	38.37	−0.09	38.57	−0.02	39.53	−0.09	38.22
39	−0.05	38.05	−0.01	38.75	−0.03	38.63	0.03	39.88	−0.02	38.61
40	0.02	38.47	0.06	39.12	0.04	38.69	0.09	40.23	0.05	39.01
41	0.10	38.90	0.12	39.50	0.10	38.75	0.15	40.58	0.12	39.40
42	0.18	39.32	0.19	39.87	0.16	38.82	0.21	40.93	0.20	39.80
43	0.27	39.75	0.25	40.25	0.23	38.88	0.27	41.29	0.27	40.20
44	0.35	40.19	0.32	40.64	0.29	38.95	0.33	41.65	0.34	40.60
45	0.43	40.64	0.39	41.03	0.36	39.02	0.39	42.01	0.42	41.01
46	0.52	41.09	0.46	41.43	0.43	39.08	0.45	42.37	0.49	41.43
47	0.61	41.55	0.53	41.84	0.50	39.15	0.51	42.75	0.57	41.86
48	0.69	42.02	0.60	42.26	0.57	39.23	0.57	43.13	0.65	42.30

49	0.7942.50	0.6842.69	0.6539.30	0.6443.53	0.7342.75
50	0.8843.00	0.7543.14	0.7239.38	0.7043.93	0.8243.22
51	0.9843.52	0.8343.60	0.8039.46	0.7744.35	0.9143.70
52	1.0844.05	0.9244.08	0.8939.54	0.8544.80	1.0044.21
53	1.1844.61	1.0044.59	0.9839.63	0.9245.26	1.1044.74
54	1.3045.20	1.1045.13	1.0739.72	1.0045.75	1.2045.30
55	1.4145.83	1.2045.71	1.1739.82	1.0946.28	1.3145.90
56	1.5446.50	1.3046.33	1.2839.93	1.1946.85	1.4246.55
57	1.6847.22	1.4247.01	1.4040.05	1.2947.47	1.5547.25
58	1.8348.02	1.5547.77	1.5340.18	1.4048.17	1.6948.03
59	2.0048.92	1.7048.62	1.6840.33	1.5348.96	1.8548.91
60	2.1949.96	1.8749.62	1.8540.51	1.6949.88	2.0449.92
61	2.4351.20	2.0850.82	2.0640.72	1.8751.01	2.2651.14
62	2.7352.77	2.3552.36	2.3340.98	2.1152.46	2.5452.69
63	3.1454.95	2.7254.53	2.7041.36	2.4654.54	2.9354.84
64	3.8358.63	3.3758.28	3.3542.00	3.0758.22	3.6058.52
65	5.0465.00	4.5365.00	4.5043.16	4.1965.00	4.7865.00