Psicothema

Psicothema 2017, Vol. 29, No. 3, 390-395
doi: 10.7334/psicothema2016.391

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Calibrating a new item pool to adaptively assess the Big Five

María Dolores Nieto[1], Francisco J. Abad[1], Alejandro Hernández-Camacho[1], Luis Eduardo Garrido[2], Juan Ramón Barrada[3], David Aguado[1] and Julio Olea[1]

[1] Universidad Autónoma de Madrid, [2] Universidad Iberoamericana en República Dominicana and [3] Universidad de Zaragoza

Abstract

Background: Even though the Five Factor Model (FFM) has been the dominant paradigm in personality research for the past two decades, very few studies have measured the FFM adaptively. Thus, the purpose of this research was the building of a new item pool to develop a computerized adaptive test (CAT) for personality assessment. **Method:** A pool of 480 items that measured the FFM facets was developed and applied to 826 participants. Facets were calibrated separately and item selection was performed being mindful of the preservation of unidimensionality of each facet. Then, a post-hoc simulation study was carried out to test the performance of separate CATs to measure the facets. **Results:** The final item pool was composed of 360 items with good psychometric properties. Findings reveal that a CAT administration of four items per facet (total length of 120 items) provides accurate facets scores, while maintaining the factor structure of the FFM. **Conclusions:** An item pool with good psychometric properties was obtained and a CAT simulation study demonstrated that the FFM facets could be measured with precision using a third of the items in the pool.

Keywords: Item pool, computerized adaptive testing, personality assessment, Five Factor Model, graded response model.

Resumen

Nuevo banco de ítems para evaluar adaptativamente los Cinco Grandes. **Antecedentes:** a pesar de que el Modelo de los Cinco Factores (MCF) ha sido el paradigma predominante durante las últimas dos décadas, muy pocos estudios han medido el MCF de forma adaptativa. El objetivo de esta investigación fue construir un nuevo banco de ítems para desarrollar un test adaptativo informatizado (TAI) para evaluar la personalidad. **Método:** se desarrolló un banco de 480 ítems para evaluar las facetas del MCF y se aplicó a 826 participantes. Cada faceta se calibró por separado y la selección de ítems se realizó atendiendo a que cada faceta fuese unidimensional. Después se realizó un estudio de simulación post-hoc para evaluar la eficiencia de TAIs a nivel de facetas. **Resultados:** el banco final estaba formado por 360 ítems con buenas propiedades psicométricas. Los resultados demostraron que la aplicación adaptativa de cuatro ítems por faceta proporciona puntuaciones precisas en las mismas, al mismo tiempo que se mantiene la estructura factorial del MCF. **Conclusiones:** el banco final está formado por ítems con buenas propiedades psicométricas. La aplicación adaptativa del banco permite medir la personalidad de forma eficiente a nivel de facetas utilizando una tercera parte de los ítems.

Palabras clave: banco de ítems, test adaptativo informatizado, evaluación de la personalidad, Modelo de los Cinco Factores, modelo de respuesta graduada.

Over the past 25 years the Five Factor Model (FFM) of personality traits (also called 'Big Five') has been established as the dominant paradigm in personality research, exceeding 300 publications per year (John, Naumann, & Soto, 2008). The FFM assumes a multifaceted structure with five broad personality traits (i.e., domains) each one containing several narrower traits (i.e., facets).

Although in personality research there is a debate about the measurement of facets versus domains, many studies have shown that narrow measures contribute to the prediction of several outcomes in various contexts (e.g., Ashton, Paunonen, & Lee, 2014). Thus, most personality tests developed to measure the FFM are based on facets. This is the case for the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) and the International Personality Item Pool Representation of the NEO PI-R (IPIP-NEO; Goldberg, 1999).

Because the FFM contain many facets, these questionnaires are usually very long (e.g., 240 items for the NEO PI-R), resulting in individual assessments that are oftentimes time consuming and inefficient. As a counter measure, short versions of such scales have been proposed but these have been designed to assess the broad domains, thereby ignoring the individual facet scores and even excluding facets. For example, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa Jr., 2007) is a version of the NEO PI-R with 60 items taken from 28 of the 30 facet scales. Another characteristic of some personality tests like the IPIP-NEO is that the items are placed in the public domain. Although this has given rise to great advances in personality research, its use could not be recommended in evaluation contexts where examinees must not know the item content prior to the administration.

Advances in measurement with item response theory (IRT) have allowed the application of computerized adaptive testing (CAT) as an alternative to traditional tests in a variety of contexts,

Received: December 22, 2016 • Accepted: April 6, 2017
Corresponding author: Francisco José Abad
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: fjose.abad@uam.es

including the study of personality. Pioneer attempts have been carried out recently to measure the Big Five adaptively. Two studies have performed real-data simulations using responses to the NEO-PI-R items. First, Reise & Henson (2000) found that administrating separate CATs for evaluating the FFM facets provided accurate estimates with half of the NEO PI-R items. More recently, Makransky, Mortensen, and Glas (2012) applied separate multidimensional CATs in order to measure the facets on each domain and obtained increases in the reliability of the facet scores. Also, the Tailored Adaptive Personality Assessment System (TAPAS) is a CAT used to measure the FFM in military settings in the United States (e.g., Stark, Chernyshenko, Drasgow, & White, 2012). Recently in Spain, Pedrosa, Suárez-Álvarez, García-Cueto, and Muñiz (2016) developed a CAT to assess specific personality traits of enterprising personality in young people.

The main core of a CAT is the wide pool of items that is calibrated with an IRT model (i.e., the person and item parameters are known). In the Reise and Henson (2000) and Makransky et al. (2012) studies the items of the NEO-PI-R were calibrated, thereby creating an item pool. However, because a number of phases are involved in an item pool construction, the current psychometric literature recommends other rigorous analyses that should be performed before starting the calibration such as testing the unidimensionality of the constructs and the fit at the item level (e.g., Revicki, Chen, & Tucker, 2015).

In view of all the above, we present in this study the development of an item pool to constitute the basis for the first Spanish CAT to measure the FFM facets efficiently. To do so, we identify four major steps: (a) develop items of each facet and obtain evidence for content validity, (b) calibrate each facet separately, checking the unidimensionality assumption and IRT fit, (c) test the performance of separate facet CATs, and (d) obtain evidences for internal structure and convergent validity. Thus, the specific purposes of this study were (a) to design, calibrate, and validate a new item pool based on the FFM and (b) to study the performance of CATs to measure the FFM facets more efficiently.

Method

Participants

A sample of 871 psychology undergraduate students participated voluntarily in the study. The sampling was intentional. Preliminary analyses revealed that a low percentage of the participants (45 respondents, 5.16% of the initial sample) presented careless, invalid or atypical responses according to multiple criteria described in the data analysis section and were consequently excluded. The final sample was composed of 826 individuals aged 17 to 50 years (M = 20.06, SD = 3.73), of which 696 were female (70.91%). For some analyses, the whole sample was randomly divided into two datasets with equal size (n = 413), one for applying exploratory statistical analysis (model-derivation sample) and the other one for validating statistical results (validation sample). The University Research Ethics Committee granted approval for the present study. The full anonymized data set is available from the authors upon request.

Instruments

Personality item pool. According to the traditional descriptions of the FFM facets, four independent experts in personality assessment and psychometrics developed an initial pool of 480 items (16 per facet) in Spanish language. The recommendations for item pool building were followed (e.g., Revicki et al., 2015). Then, each expert reviewed the item content of the whole pool and redundant statements were excluded and replaced by new ones. The statements were administered using a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). A Spanish philologist revised the items and corrected grammar, spelling and style errors. Table 1 shows facets 1 to 6 for each domain.

Directed questions scale. A scale of 12 Likert-type items (1 = strongly disagree, 5 = strongly agree) directing participants to give specific responses (e.g., "If you are reading this question, please mark 'Disagree'") was applied to measure inattention. Scale scores were obtained by summing the correct responses.

NEO-FFI-3. The NEO-FFI-3 inventory, a 60-item version of the NEO-PI-3 (McCrae, Costa Jr, & Martin, 2005) to measure the FFM domains, was included to obtain evidences for convergent validity of the new item pool. The NEO-PI-3 is a revision of the NEO PI-R. Due to there are no Spanish versions of the NEO-PI-3 and the NEO-FFI-3 questionnaires, 59 of the 60 items of the NEO-FFI-3 were selected from the Spanish version of the NEO-PI-R (Cordero, Pamos, & Seisdedos, 2008). The remaining item was translated from the English version of the NEO-FFI-3.

Procedure

The items from the personality item pool, the Directed questions scale and the NEO-FFI-3 were used to create two booklets that were administered in two sessions in a counterbalanced order. Participants completed the items within an official system of data collection in a faculty of Psychology whose purpose is the participation of students in research projects in exchange for academic compensation.

Table 1
Five Factor model: Domains and facets

	Domain				
Facet	**Neuroticism**	**Extraversion**	**Openness**	**Agreeableness**	**Conscientiousness**
1	Anxiety	Warmth	Fantasy	Trust	Competence
2	Angry/hostility	Gregariousness	Aesthetics	Straightforwardness	Order
3	Depression	Assertiveness	Feelings	Altruism	Dutifulness
4	Self-consciousness	Activity	Actions	Compliance	Achievement striving
5	Impulsiveness	Excitement seeking	Ideas	Modesty	Self-discipline
6	Vulnerability	Positive emotions	Values	Tender-mindedness	Deliberation

Data analysis

Evidence for content validity. Evidence for content validity of the personality item pool was obtained. Thirty-six experts in personality research and psychometrics were asked to select the facet to which each item belonged. Each expert evaluated the items from two domains. The level of congruence between the experts for each item was measured as the percentage of classification agreement for its most chosen facet. After excluding the responses from experts with low reliability (i.e., percentage of congruence lower than 70% in at least one domain), items with less than 50% of classification in their corresponding theoretical facet were removed from the pool.

Personality item pool IRT calibration. Psychometric properties of the pool were analyzed by fitting the unidimensional graded IRT response model (Samejima, 1969) to each subset of items measuring the same facet. First, some indexes were examined in order to screen out data for careless, invalid or atypical responses (i.e., score below 9 points on the Directed questions scale, double responses in more than three items, more than 10 missing values on the personality items, outliers regarding the number of consecutive identical responses).

For each facet, the unidimensionality assumption was tested on the model-derivation sample by applying parallel analysis (PA) and the unidimensional factor model with the polychoric correlation matrix and the robust unweighted least squares (ULSMV) estimator. If unidimensionality was not tenable according to PA or some variables had very low factor loadings, items were iteratively removed until the unidimensionality assumption was met and all the items had factor loadings larger than .2. For purposes of achieving unidimensionality, the highest residual correlation was identified and the item with the smaller loading in this pair was deleted. At the end of the iterative process, PA and the comparative fit index (CFI) were used, as recommended in Garrido, Abad, & Ponsoda (2016) to assess the unidimensionality of facets in the cross-validation sample. The conventional cutoff values for the CFI, are .90 or greater for acceptable fit, and .95 or greater for good fit (Hu & Bentler, 1999).

The selected subset of unidimensional items of each facet was calibrated separately according to the graded IRT response model using the Metropolis-Hastings Robbins-Monro algorithm (MHRM; Cai, 2010a, 2010b) on the whole sample. Item fit was tested on the sample with complete response patterns using the polytomous variant of the S-χ^2 index (Orlando & Thissen, 2000) with the Benjamini-Hochberg adjustment to control Type I error (Benjamini & Hochberg, 1995). Finally, the IRT maximum a posteriori (MAP; Embretson & Reise, 2000) pool facet scores and the standard errors (*SE*s), indicating the precision of trait estimates (θ), were obtained for each individual in each facet. IRT marginal reliabilities for pool facet scores were also obtained (Brown & Croudace, 2015; p. 314).

Performance of the CAT. A post hoc simulation study was carried out to analyze the performance of the CATs in measuring the FFM facets. We simulated a separate CAT for each facet using the item responses obtained from the respondents. Since omissions are not allowed in CATs, the response vectors were completed using item and respondent estimated parameters obtained in the previous calibration step. The CAT algorithm started by selecting the item that maximized the Fisher information at $\theta = 0$ for all the respondents. Then, attending to a respondent answer, the MAP θ estimate was obtained. The next item selected was the one that maximized the Fisher information evaluated at the θ estimate. These steps were repeated until the algorithm stopped when four items were administered. Then, the final CAT facet score was estimated using the MAP method.

Different criteria were used to analyze the precision of the CATs. For each facet, the correlation between the CAT and the pool scores were obtained. We also obtained the empirical reliability and the median of the *SE* across examinees for each CAT score.

Evidence for internal structure and convergent validity for pool and CAT facet scores. First, evidence based on the factorial structure of the pool facet scores was obtained. PA with Pearson correlations was used to verify that the suggested number of factors was five as expected (one factor per personality domain). Next, we applied exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) with the maximum likelihood estimator. Unlike exploratory factor analysis, ESEM models can include both exploratory and confirmatory methods (e.g., correlated error terms). Using the model-derivation sample, we defined five correlated ESEM factors corresponding to the five domains. The Oblimin rotation method was used. Since modification indices suggested some correlated residuals, a new model including them was tested using the cross-validation dataset. Again, PA and the CFI were used for model evaluation. Additionally, the same ESEM factor model was used to test the internal structure of CAT facet scores. Factor congruence coefficients were obtained to study the similarities of the factorial structure obtained with pool and CAT scores.

Following the previous step, pool and CAT domain scores were obtained as an average of the correspondent six facet scores. Composite reliabilities for domain scores were estimated from the ESEM models as the squared correlation between the domain trait score and the corresponding latent factor (Raykov, 1997). Finally, evidence for convergent validity was obtained by computing the correlations between the CAT and the pool domain scores with the NEO-FFI-3 raw scores.

All the analyses were performed with M*plus* 7 (Muthén & Muthén, 1998-2012) and the R packages *psych* (Revelle, 2016), *mirt* (Chalmers, 2012), and *mirtCAT* (Chalmers, 2016).

Results

Evidence for content validity. Two experts out of 36 were excluded by their low percentage of congruence (below 70%) in the Extraversion domain. After excluding these experts, the average percentages of congruence by domain were 84% for Neuroticism, 86% for Extraversion, 93% for Openness, 89% for Agreeableness, and 86% for Conscientiousness. Twenty-five items out of 480 were removed from the item pool by their low percentage of classification in the theoretical facet (less than the 50%). After excluding these items, the average percentages of classification accuracy by domain were 89% for Neuroticism, 87% for Extraversion, 94% for Openness, 90% for Agreeableness, and 89% for Conscientiousness.

Personality item pool IRT calibration. Out of 871 participants 45 were excluded from the sample of analysis because they presented careless, invalid or atypical responses. Missing data rate for item nonresponse was very low with a maximum value of 2%.

Out of 455 items 95 were removed in order to preserve the unidimensionality of each facet. The largest number of excluded items in one facet was 7 (i.e., in the Assertiveness, Straightforwardness, and Dutifulness facets). For the retained items, the unidimensionality assumption was always tenable according to PA. The unidimensional solution showed acceptable fit according to the CFI, which was equal or above .90 in 80% of the cases and equal to or higher than .85 in the remaining facets (except for Tender-mindedness, CFI = .62). PA indicated that the 67% of the facets were unidimensional. In the remaining facets, PA suggested a two-factor solution (except for Excitement seeking that PA indicated three factors). In this cases, the scree test revealed that the second empirical eigenvalue was barely greater than the random eigenvalue. All the item factor loadings on the unidimensional solutions were statistically significant ($p<.05$), with average loadings ranging from .45 to .73.

Within the framework of the IRT, only 4 items out of 360 were identified as misfitting to the graded response model according to the S-χ^2 index. The *a-parameter* of the items showed adequate positive values ranging from 0.35 to 3.86 ($a^- = 1.51$), with 23% of them being highly discriminative (i.e., $a > 2$).

Figure 1 illustrates the information and *SE* for each θ pool facet scores. For θ between –3 and 3, the *SE*s for almost all the facets, except Compliance and Dutifulness, were lower than .5, which is approximately equivalent to a reliability coefficient of .75. This indicates that the items provide good information across the different traits levels of each facet, except for the two facets mentioned. Regarding marginal reliability, all facet scores presented values equal to or above .72. Average reliabilities for pool facet scores within a domain were .89, .90, .88, .85 and .86 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

Performance of the CAT. Correlations between each CAT facet scores and pool facet scores were high for all the facets with values ranging from .92 to .98 (r = .95). For most facets, the median of the participants' *SE* was lower than .4. Only Ideas (Mdn_{SE}=.41), Compliance (Mdn_{SE}= .48), Tender-mindedness (Mdn_{SE}= .41), and Dutifulness (Mdn_{SE}= .53) presented higher values. Regarding marginal reliability, most facet scores presented values equal or above .7, except the Dutifulness facet with a value of .68. Average reliabilities for pool facet scores within a domain were .82, .86, .81, .79 and .79 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

Evidence for internal structure and convergent validity for pool and CAT facet scores. As expected, PA based on the analysis of the pool facet scores suggested five factors. Thus, a five-factor

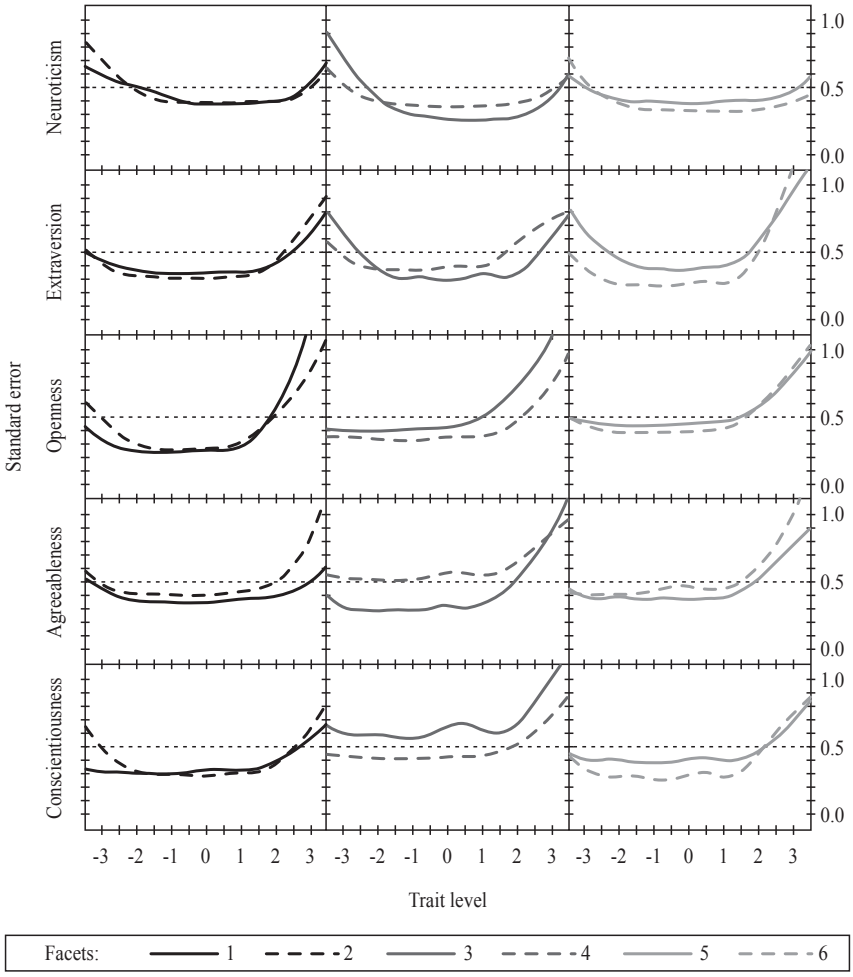

Figure 1. Standard error (*SE*) across the trait level for the facets of each domain of the FFM. *SE* equal to .50 is indicated with a dotted line. The facets 1 to 6 of each domain are specified in Table 1

exploratory model was first specified for the ESEM analyses in the model-derivation sample. This model was then modified adding six correlated residuals according to modification indexes above 40. Correlated residuals were theoretically meaningful (e.g., a negative correlation between Deliberation and Impulsiveness) and were replicated in the validation sample in which the modified model fit was acceptable: CFI was .91 and PA indicated a five-factor solution.

In the final modified model, almost all the facet scores loaded higher and significantly on its respective domain factor. These loadings were medium-high sized with values above .40 (M = .61). Only the Social anxiety and Deliberation facets presented values below .40 (.35, and .31, respectively). Regarding cross-loadings, most of them were on the Extraversion (Depression: –.33, Social anxiety: –.63, Impulsiveness: .45, Actions: .39, Trust: .35, and Deliberation: –.43), Agreeableness (Angry/hostility: –.35, Feelings: .37, Dutifulness: .35, and Deliberation: .30), and Openness (Emotions seeking: .38; Order: –.36) domains. Also Activity and Competence facets cross-loaded .33 and –.44 on Conscientiousness and Neuroticism, respectively. Average cross-loading (in absolute value) was low (.14).

The factor correlation matrix showed that Neuroticism correlated negatively with Extraversion (r = –.28; p<.001), and Conscientiousness (r = –.21; p<.001). Additionally, Extraversion also correlated, positively, with Openness (r = .24; p<.001) and Conscientiousness (r = .23; p<.001). Conscientiousness was also correlated with Openness (r = .12; p<.001) and Agreeableness (r = .10; p<.001). The remaining correlations were small ($|r|$ < .06).

When the ESEM was applied to the CAT facet scores, the results were highly similar (i.e., congruence coefficients were .99 for each of the five factors). Composite reliabilities for pool domain scores were acceptable and ranged from .75 (Agreeableness) to .87 (Extraversion). Reliabilities for CAT domain scores were inferior as expected but acceptable and ranged from .70 (Openness) to .86 (Extraversion). According to the Spearman-Brown formula and the pool composite reliabilities, it must be noted that in order to obtain these 24-item length CAT domain score reliabilities, 56 items would be required, in average, in a fixed form.

Finally, correlations between the pool domain scale scores and the NEO-FFI-3 raw scores were good. The Extraversion and Neuroticism domains presented the highest convergent validity values (r = .88 and .86, respectively). In the case of Openness and Agreeableness scales the value was similar (r =.83), and Conscientiousness presented the lowest value (r =.80). Convergent validity for the CAT domain scale scores with the NEO-FFI-3 were only slightly inferior (the largest difference, .02, was for Neuroticism).

Discussion

Recent studies in personality have investigated the possibility of obtaining accurate personality facet scores with CATs (e.g., Makransky et al., 2012). The purpose of this research was to build a new personality item pool and develop the first Spanish CAT based on the FFM facets. Analyses were performed at the facet-level. This is one of the key aspects of this study because recent research has shown that facet-level analysis increases the predictive validity of personality scores (Ashton et al., 2014).

In this study a pool of items for personality assessment is provided and efficiently administered with CAT. Although there are several commercial paper-and-pencil tests for assessing the FFM, this might be an important contribution to the evaluation of personality in applied settings where short-time assessments are required and the item content should be unknown to the examinees prior to administration.

Four main steps are distinguished in this study. First, item statements were developed and evidence for content validity was obtained via the evaluation of experts. Second, each facet was calibrated separately according to the Samejima graded response model. Unidimensionality of facets was guaranteed through a strict iterative analysis procedure and almost all the items showed adequate fit to the Samejima graded response model. In terms of precision, the facet scales showed generally good reliability with small SE over a wide range of θ. In line with previous studies (e.g., Benet-Martínez & John, 1998) and the NEO PI-R manuals, the facets of the Neuroticism, Extraversion and Openness domains were, on average, the most reliable.

Third, a CAT simulation study revealed that using separate 4-item CATs to assess the facets (i.e., with an administration of 120 items), facet scores are estimated accurately with low SEs in most cases. Finally, internal structures of the pool and the CAT were analyzed obtaining similar results: facets in both instruments measured the narrow traits of their corresponding FFM domains. Some facets loaded on more than one domain (e.g., Angry/hostility was designed to measure a subdomain of Neuroticism and was also an indicator of Agreeableness). This is consistent with previous studies that have shown that an important part of the variance of the facets scales is due to different domains (e.g., Abad, Sorrel, García, & Aluja, in press). In addition, both the item pool and CAT scores showed good convergent validity with the NEO-FFI-3 questionnaire.

One limitation of the current study is the generalizability of the results to other samples, although the intercorrelations found between the five personality factors are consistent with previous research. For example, Neuroticism correlated negatively with Extraversion and Conscientiousness, and Extraversion also correlated positively with Openness (e.g. Mount, Barrick, Scullen, & Rounds, 2005; Van der Linden, te Nijenhuis, & Bakker, 2010). Furthermore, domains such as Neuroticism and Openness showed lower correlations. However, due to the fact that the sample consisted of psychology undergraduate students, we are aware that the results may not be generalized to other sub-populations (e.g., clinical, workforce).

Recent research has suggested that multidimensional IRT models and multidimensional CATs may increase the precision of personality trait scores (e.g., Makransky et al., 2012). In this regard, future research with the presently developed item pool should be oriented toward the application of multidimensional models in the calibration and adaptive administration phases.

Acknowledgements

The research has been funded by the Ministry of Economy and Competitivity of Spain (PSI2013-44300-P), and the UAM-IIC Chair «*Psychometric Models and Applications*».

References

Abad, F. J., Sorrel, M. A., García, L. F., & Aluja, A. (in press). Modeling general, specific, and method variance in personality measures. Results for ZKA-PQ and NEO-PI-R. *Assessment*. doi: 10.1177/1073191116667547

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56*, 24-28. doi: 10.1016/j.paid.2013.08.019

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. doi: 10.1080/10705510903008204

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729-750. doi: 10.1037/0022-3514.75.3.729

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. doi: 10.2307/2346101

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 334-363). New York, NY: Routledge.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hasings Robbins-Monro algorithm. *Psychometrika, 75*, 33-57. doi: 10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335. doi: 10.3102/1076998609353115

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*, 1-39. doi: 10.18637/jss.v071.i05

Cordero, A., Pamos, A., & Seisdedos, N. (2008). NEO PI-R, Inventario de Personalidad NEO Revisado [Revised NEO Personality Inventory]. *Madrid: TEA Ediciones*.

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional*. Odessa, FL: Psychological Assessment Resources, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93. doi: 10.1037/met0000064

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe, 7*, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

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. John, R. Robins & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, Guilford.

Makransky, G., Mortensen, E. L., & Glas, C. A. (2012). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment, 20*, 3-13. doi: 10.1177/1073191112437756

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*, 261-270. doi: 10.1207/s15327752jpa8403_05

McCrae, R. R., & Costa Jr., P. T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28*, 116-128. doi: 10.1027/1614-0001.28.3.116

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology*, *58*, 447-478. doi: 10.1111/j.1744-6570.2005.00468.x

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th ed.)*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64. doi: 10.1177/01466216000241003

Pedrosa, I., Suárez-Álvarez, J., García-Cueto, E., & Muñiz, J. (2016). A computerized adaptive test for enterprising personality assessment in youth. *Psicothema, 28*, 471-478. doi: 10.7334/psicothema2016.68

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184. doi: 10.1177/01466216970212006

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364. doi: 10.1177/107319110000700404

Revelle, W. (2016). Procedures for personality and psychological research. Evanston, Il: Northwestern University.

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 334-363). New York, NY: Routledge.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. *Psychometrika*, *34 (Suppl. 1)*, 1-97. doi: 10.1007/BF02290599

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*, 463-487. doi: 10.1177/1094428112444611

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327. doi: 10.1016/j.jrp.2010.03.003