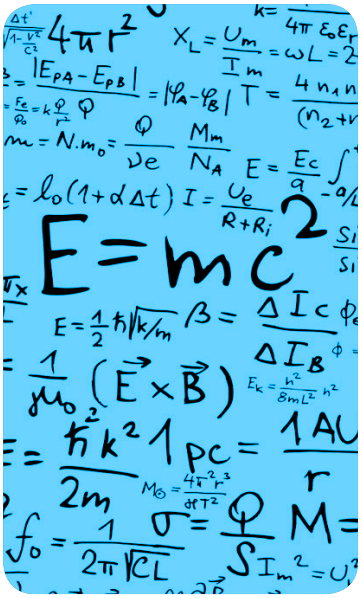
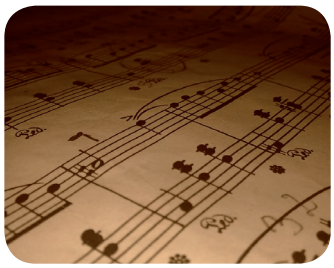




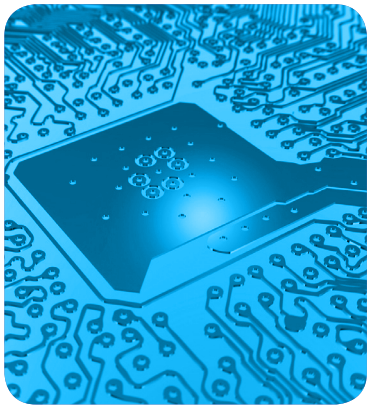
MÁSTERES de la UAM

Facultad de Psicología /13-14

Máster en Metodología de las Ciencias del Comportamiento y la Salud

UNIVERSIDAD AUTONOMA DE MADRID

excelencia Campus Internacional UAM CSIC



Spanish Validation of Short Five (S5), a comprehensive measure of personality

Regina Gárcia-Velázquez

UAM EDICIONES

Abstract

This paper describes the Spanish adaptation of the Short Five personality inventory (Konstabel, Lönnqvist, Walkowitz, Konstabel, & Verkasalo, 2012). The perspective from where it was constructed, by implementing comprehensive single items, is discussed along with the cross-cultural features of personality research. The main goal of this work was to supply evidence of equivalence with the other language versions, by applying multiple validation strategies and focusing on distinctive validity issues. Four studies compose this work. First, a differential item functioning analysis between Spanish and Finnish versions is described. A second study presents the psychometric properties of Spanish version, showing internal consistency, congruence coefficients and dimensionality of the instrument. Third, item bias concerning gender is assessed by logistic regression detection method. The fourth study provides with different measures of evidence regarding construct and predictive validity: multitrait-multimethod matrix with Spanish NEO PI-R (Costa & McCrae, 1992), interpretations of the associations of personality facets with Schwartz's motivational values (2012) and prediction of relevant behavioural criteria assessed by Paunonen's Behavioral Report Form (2003). Results support the use of Short Five specially in settings were time is limited or there is important risk of respondent's fatigue that likely introduce important bias.

Keywords: Big Five, Personality, Inventory, Validation, Spanish

Spanish Validation of Short Five (S5), a comprehensive measure of personality.

Contents

Introduction

Short Five (SF) was developed in 2012 (Konstabel, Lönnqvist, Walkowitz, Konstabel, and Verkasalo) as a new way to construct short questionnaires in personality measurement. It was designed to evaluate the facets identified by the NEO PI-R (Costa & McCrae, 1992). These facets were intended to sample both the entire domain, and single constructs of interest in their own particularity (Costa & McCrae, 1995, 2008). Although the facet system of the NEO PI-R may not be flawless, it has some advantages that advocate the decision. First, it is difficult to argue that it is the most familiar personality taxonomy among researchers. Second, it can be considered one the most tested theories in personality. This entails a large, solid, volume of literature and empirical evidence, broad enough to feed the iterative nature of theory building and revision (Cronbach & Meehl, 1955).

After few decades of production (McCrae & Costa, 1987) and despite findings being not always conclusive, strong support has been found for the Big Five theory (BF, McCrae and Costa 1996, 2010) across dozens of cultures (McCrae & Allik, 2002; Schmitt, Allik, McCrae, & Benet-Martínez, 2007; Schmitt, Realo, Voracek, & Allik, 2008), behavioural genetics studies (Briley & Tucker-Drob, 2012; Yamagata et al., 2006), development and ageing related to stability and plasticity of the traits (Marsh, Nagengast, & Morin, 2013; Soto, John, Gosling, & Potter, 2011; Specht, Egloff, & Schmukle, 2011; Terracciano, McCrae, Brant, & Costa Jr, 2005), prediction of behaviours (Paunonen, 2003) and relevant criteria about life outcomes (George, Helson, & John, 2011; Ozer & Benet-Martinez, 2006).

Thus, the BF model was chosen for constituting the structure of S5 from the facet level, since it has been demonstrated that the lower level constructs display better predictions than the broad, upper in hierarchy, domains (Paunonen, 1998; Paunonen & Ashton, 2001). This idea conveniently suits the perspective of construct disaggregation for a better understanding of the domains' inner nature. As Smith and Zapolski (2009) argue, multidimensional scores (such as the BF domain scores, accounting for multiple facets) have unclear meaning concerning (a) it is not possible to know which of those components account for how much of the test's covariance with measures of other constructs, (b) the contribution of the different components to the composite score is indistinguishable, and (c) the same composite score can reflect infinite different combinations of the construct components.

Rationale under the Short Five questionnaire

The idea was to present a form defined by being *short* and composed by *comprehensive* items. A short scale is obviously more attractive for a potential participant, since it demands less time to be completed. The consequences: an expected increase in respon-

dents' involvement, and side effects associated to long inventories are reduced. Moreover, as Lönnqvist et al. (2007) showed, the less demanding a scale is, the less biased the personality profile of the volunteers. By comprehensive items are understood those items which broadly and inclusively describe the construct measured. This strategy counteracts some of the common criticism to short scales, such as the effects of the biased response styles and limitations in content validity.

Comprehensive items elaborate on the feature measured, somewhat contributing to neutralize the fact that items in personality tend to be ambiguous and easy to generalize (Church, 2010). By sparing a test from narrow items, content validity is favoured when detailing more information regarding the attribute to rate, so respondents have a deeper insight. This point also follows the recommendation of Knowles and Condon (1999), who point that reflection-inductive items and balanced wording stand as palliatives of response style effects.

Short scales rate lower on reliability in comparison to longer ones, since measurement error is reduced by adding more items in the latter. On the other hand, as Robins, Hendin, and Trzesniewski (2001) claim, item redundancy found in longer questionnaires may frustrate, fatigue and bore participants asked to answer similar questions repeatedly. For these reasons the quality of the information provided by hundreds of questions answered in a row induces some doubts, suggesting that balance should be pursued in this point.

On the other side of the coin, scales composed by large number of items are often psychometrically evaluated at the facet level due to the big number of variables, so composites of items are used already in early stages. Parceling has been a controverted practice in measurement, although it has advantages, when validating a test, single items ought to be assessed not only for reliability, but also concerning structure and other validity evidence. When parcels of items are used, there is no chance to distinguish the functioning of the items which compose it. Additionally in personality research, it has been shown that facets with high internal consistency do not show better validity (McCrae, Kurtz, Yamagata, & Terracciano, 2011). This findings advocate the idea of revising scales at the item level as a way to approach more accurate information about their properties.

After some reflection it comes naturally that less does not necessarily mean worse in measurement. Under this perspective, the SF inventory was developed and analyzed its psychometric properties within the framework of the BF theory. Versions of several languages have been adapted. The aim of this work is to present the Spanish version of S5.

Previous results of Short Five

Four different language versions were published together in 2012: Estonian, Finnish, English and German (Konstabel et al.). Studies provided with conventional reliability

and structural evidences, as well as with different information each of them. All versions showed acceptable to good reliability ($\hat{\alpha}$ scale values from .72 to .89), and Tucker's Phi congruence coefficients suggested in all cases high similarities with normative BF structure (.90 to .97).

In the Estonian version, EPIP-NEO (Estonian version of the IPIP-NEO) and S5 were compared in predicting self-ratings of emotional experience. Also measures correlations of self-two peers ratings of S5, and other personality inventories were computed and compared, obtaining S5 good indices. The Finnish S5 was answered together with the Finnish and Swedish versions NEO PI-R (the latter by a bilingual sample), obtaining domain-level correlations between .79 (Swedish A) and .91 (Finnish N). Correlations were also computed in the same fashion, between English NEO PI-R and S5, reaching values between .78 (A) and .87 (N). Information about timing of S5 responding was provided. Finally, the German version was used for predicting behavioural criteria as assessed by the Behavior Report Form (Paunonen, 1998; 2003). R^2 of the linear models were compared with German NEO PI-R and NEO-FFI (Ostendorf & Angleitner, 2004). Results supported the use of S5 as a short measure of personality, with all measures showing very similar levels of validity when compared with the other Big Five forms.

Short Five comprises a total of 60 items, being each of the 30 facets (six facets by five personality traits) measured by two items. From these two items, one represents the facet positively keyed, and the other one negatively. By this strategy a balanced set of items measures each construct in the questionnaire, as it has been recommended for decreasing biased response style risk (Knowles & Condon, 1999).

Cross-cultural research in Personality Psychology

It is well known that cross-cultural test adaptations need for a rigorous and thorough procedure. There is consensus about the taxonomy of equivalence and bias proposed by van de Vijver and colleges (van de Vijver & Leung, 1997; Van de Vijver & Poortinga, 1997), who claim that validity of inferences should be the main goal pursued, given the fact the cross-cultural research does not easily allow for experimental manipulations or other high control conditions.

Van de Vijver and Leung (2011) understand *equivalence* as "the level of comparability of measurement outcomes" (pp.19). They distinguish between the procedures aiming to enhance equivalence from those which test it. On the other hand, the authors point out the importance of minimizing *bias* as a systematic distortion of the outcome of a test, being this an obstacle for comparing different scores. For van de Vijver and Leung bias and equivalence are "two sides of the same coin". Our work endeavors to optimize validity based on this perspective. The strategy will be explained in the following paragraphs.

Three types usually comprise the taxonomy of bias in cross-cultural contexts. As treatment for *construct bias* in S5, it can be noted that item redaction aimed to cover,

comprehensively, each one of the facets composing the BF model, being the cross-cultural applicability of this theory well documented, especially in occidental cultures. Regarding the cultural relevance of the content chosen, no item depicts any particular situation susceptible of differing across cultures, as it is not compatible with the idea of comprehensibility and also typical for personality inventories, where items tend to show minimal specification in order to be generally applicable (Church, 2010).

Method bias covers different topics. First, sample equivalence was assured by using as similar sample in the different instrument applications. For this purpose, mainly university students and some snowball-gathered participants were administered the test, so the procedure would be akin to the other countries. Concerning administration and instrument bias, these were neutralized by applying identical format of the questionnaire, as well as displaying both online and paper forms. For instance, administrators had native language skills and gave the same instructions. Thus, conditions were not different to the application of the other S5 versions.

The last category, avoidance of *item bias*, has experimented improvement since the statistical packages provide with methods for assessing *differential item functioning* (Holland & Wainer, 1993; Teresi, 2001). It is understood as a difference in the probability of answering to a given item, due to other reasons than the construct measured. Thus, variables such as gender, language or country have been typically applied in order to detect whether they influence the respondents' answers. Van de Vijver and Leung (2011) propose two ways to deal with item bias. In the Spanish S5 validation, *judgmental methods* to avoid DIF were carried out in the adaptation procedure, which is detailed in the next section. Once gathered the data, DIF was consequently examined through *psychometrical methods* (a) concerning differences in language versions (Finnish and Spanish) with bilingual participants in a randomized design, and (b) controlling for gender in the Spanish version.

Concerning equivalence issues, first it can be argued that the BF model has been replicated across numerous cultures (McCrae, 2002; Schmitt et al., 2007), bringing a solid status of *construct equivalence* to the the measures embodying this system. As it has been mentioned, broad but comprehensive items compose S5 as an endeavour to ensure validity of construct through a thorough coverage. In our work, a posteriori analysis are linked to validity studies, such as a CFA-MTMM with NEO PI-R (Costa & McCrae, 1992), predictive power of behavioural criteria and expected convergence and discrimination of personality facets and motivational values.

Second, *functional equivalence* is referred to the relationship of the construct measured with others, depicting the idea of nomological networks (Cronbach & Meehl, 1955) mainly examined by the construct validity procedures commented just above, also useful for construct equivalence inspection.

One of the main goals of Spanish S5 validation is providing evidence about *struc-*

tural equivalence. This is materialized by examining the dimensionality of the instrument. Structural equivalence will be checked (a) as commonly, towards the normative BF model structure (Costa & McCrae, 1992), and (b) with respect to the rest of the S5 language versions in order to obtain a measure of structural similarity. Congruence coefficients utilized for this purpose are Tucker's Phi for proportionality, the most spread one (Lorenzo-Seva & Ten Berge, 2006), and the identity coefficient, which is the most stringent of this type (Fischer & Fontaine, 2011).

In personality research it has been shown that Exploratory Factor Analysis (EFA) with Procrustes targeted rotation provides more accurate information about the structure of the measure than other confirmatory techniques (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). This is mainly due to the excessive restrictions, which do not suit actual personality structure. As it was demonstrated by Marsh (1991a; 1991b), Confirmatory Factor Analysis (CFA) structures do not show an adequate fit to the data due to the fact that many items hold minor cross-loadings to other factors. More recently, Marsh, Hau, and Grayson (2005) claimed that "it is almost impossible to get an acceptable fit for even "good" multifactor rating instruments when analysis are done at the item level and there are multiple factors (e.g., 5-10), each measured with a reasonable number of items (e.g., at least 5-10/per scale) so that the are at least 50 items overall" (p. 325). This claim was placed on an electronic network devoted to SEM (SEMNET) by the year 2000, inviting the members (over 2000) to provide counter-examples. No responses were offered, leading some authors to conclude that the independent cluster model inherent from CFA (ICM-CFA; Morin, Marsh, and Nagengast 2013) may be too restrictive, suffering from inflated factor correlations (Marsh et al., 2010) and either forcing to the inclusion of *post-hoc* cross-loadings or the "ill-advised" use of item parcels (Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013). These practices have been criticised for biasing construct validity estimates, decreasing in the generalization and replicability of the models tested, and camouflaging model misfit and misspecification.

It is out of doubt that "pure" items are an advantage, but even the traditional concept of simple structure does not force all cross-loadings to zero (Carroll, 1953); neither it does the BF theory.

As an attempt to conciliate both benefits from EFA and CFA, Asparouhov and Muthén (2009; see also Marsh, Liem, Martin, Morin, and Nagengast 2011) developed Exploratory Structural Equation Modeling (ESEM) as an integration of EFA within global Structural Equation Modeling (SEM) framework. Morin et al. (2013) underline that ESEM fulfils the need for a flexible approach while supplying with all prototypical parameters from the SEM statistical advances: standard errors, goodness of fit indices, model comparisons, inclusion of correlated uniqueness, specification of both CFA and EFA factors, estimation of method effects and tests of multiple group.

ESEM strategy has been successfully applied to the BF model, offering good results

and wide applicability in recent studies (Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013). It has also been compared to CFA in personality measurement, obtaining better fit statistics and advantages in accuracy of construct validity (Herrmann & Pfister, 2013)

Therefore, there are reasons to consider ESEM as a suitable approach for personality research, especially when targeted rotation is applied. We believe that congruence coefficients and ESEM targeted-solution are conclusive tools for testing the structural equivalence of cross-cultural measures. For further elaboration on ESEM modelling and formulae, see the specific bibliography (Asparouhov & Muthén, 2009; Marsh, Morin, Parker, & Kaur, 2014; Morin et al., 2013)

Adaptation procedure

The translation method of the Spanish version of SF followed a double back-translation strategy (Sireci, 2005). First, the English version was translated into Spanish by two native speakers, proficient in English. This target language version was translated back independently by bilinguals, native speakers one of Finnish and other English who were blind to the source original versions of the test. The two back-translated drafts (English and Finnish) were reviewed by the authors of SF, all of them native or skilled speakers in both Finnish and English languages (see Figure 1). The differences in nuance were discussed and iteratively modified if necessary, reaching consensus for each item.

The translation was evaluated following the recommendations by Hambleton and Zenisky (2011). Given the case of comprehensive items such as S5 ones, some aspects must be carefully minded due to the additional cognitive complexity of the elaborated sentences (Yan & Tourangeau, 2008). The decisions made emphasized (a) conceptual rather than literal equivalence, as well as (b) aimed to use natural and clear language.

As the first criterion argues, similarity in meaning is essential to compare between cultures. Does the question ask the same thing? Is the connotation similar? Nuance of meaning can lead to different response patterns among cultures that could be avoided by looking deep into the accurate meaning of the words used. As van de Vijver and Poortinga (1997) discuss, linguistic equivalence does not hold for psychological equivalence. We agree with van de Vijver and Poortinga (2005) that, when translating tests across cultures, both linguistic and specially psychological considerations should be made. The so-called *application option* (van de Vijver & Leung, 1997) will only focus on semantic and style aspects by literal transformation, hence it is not sufficient to assure psychological adequacy. A truthful *adaptation* process also goes through modifying some features in order to harmonize for different cultural settings. In personality research this is particularly intricate, as the words used in different languages to describe personality traits and emotional states exhibit a broad range of connotations (Rogler, 1999; Wierzbicka, 1994)

Adopted this perspective, item wording was slightly adjusted so it would follow some

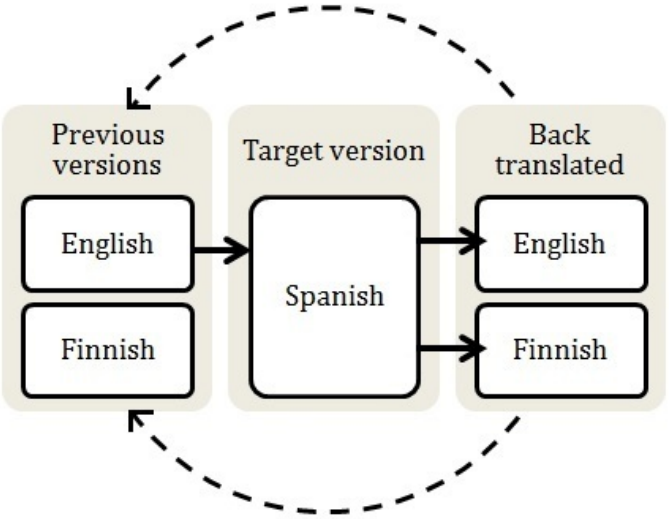


Figure 1. Chart of double translation-back-translation of the Spanish S5.

principles. Items too difficult to endorse or too general in Spanish culture where discussed until expected discrimination was considered to be consistent according to all S5 versions. As an example, the adjective "frío" regarding a person in Spanish culture pictures so deep devaluation of oneself, that it would be too unusually endorsed. Consequently, a culturally equivalent attribute such as "seco" was chosen.

The second criterion was considered as well of special relevance: translated items should be comparable in terms of difficulty and commonality with respect to the words in the item from other source language versions. Thus, most common words and expressions should be included. On the other hand, when idioms and colloquialisms were used in some items, it was thoroughly discussed their concrete meaning and whether including them would result in a more natural expression. Moreover, as mentioned above, the use of clear and natural wording is justified by the fact that comprehensive items are more demanding in cognitive terms (Yan & Tourangeau, 2008).

More structural features such as the fact that versions did not show any changes in the text, item format, physical layout, and answer choices were also taken into account. It was also taken into account to use as similar grammatical phrasing as possible.

Finally, the questionnaire was revised and modified in order to suit the Guidelines for Translating and Adapting Test, by the International Test Commission (ITC, 2010).

Study 1

In this preliminary phase translation and evaluation of item bias were conducted.

Method

Participants.

The link to the online form was spread across multicultural organizations, hispanic-finnish associations and students' e-mailing lists. 46 persons fulfilled the online form. The final sample N=42 was composed by the persons who registered proficiency in both languages, as reported by a short vocabulary test. Finnish was the mother tongue for 36 of them. The average age of the sample was 32.3 years, ranking from 20 to 57. From the 42 definite participants, 36 were women.

Measures.

- *Language vocabulary test.* Short multiple-choice test composed by 16 items including relevant words present in the S5 items. This measure was used as criterion of bilingual proficiency.
- *S5.* Both Spanish and Finnish versions, administered by halves depending on the randomly assigned group condition.

Procedure.

Once the preliminary version was ready, native speakers of Finnish and Spanish who were bilingual for both languages were administered SF for examining the comparability of both language versions. A four-group design was conducted (Sireci, 2005). Actually, despite of having implemented such design the little representation of native Spanish speakers did not allow for further analysis regarding mother tongue. Hence, DIF was examined for language versions and no trait comparisons among Spanish-Finnish native speakers could be made.

In practice, each participant was randomly assigned to one condition. Each respondent fulfilled only one form of the questionnaire, which was composed by ten identical anchor items (five traits by two languages) and 50 different language-items (25 in each language). The difference between forms is given by the language in which every half was written. The concrete design is presented in Figure 2 for a more intuitive understanding.

With the use of such design the potential practice effects are avoided and, because groups are randomly equivalent, there should not be any group effects (Sireci, 2005). The anchor items are examined to assess whether the assumption of randomly equivalent groups is met. Another example of this type of mixed-language administration can be found in Sireci and Berberoglu (2000).

Data was dichotomized as a function of endorsement of the item (response categories 1 to 3) or not (from -3 to 0). This measure was undertaken in order to ease DIF analysis, and due to the small sample size. Also for this reason, the *modified delta plot* detection method was chosen (Magis & Facon, 2012). The improvement introduced by a modification on the estimation of the threshold parameter of the previous *delta plot* method has shown to be the most appropriate algorithm for small sample sizes, even resulting more accurate

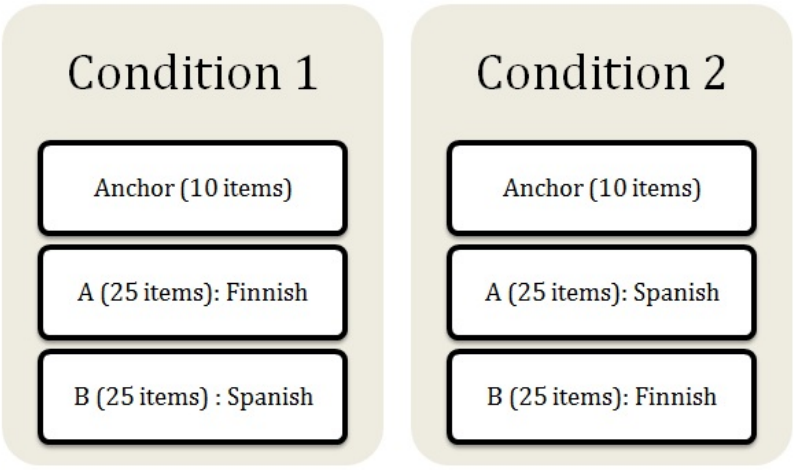


Figure 2. Double-group bilingual randomized design.

than the Mantel-Haenszel method. Among its features, the *modified delta plot* has been found efficient in terms of Type I error control and power to detect DIF items (Magis & Facon, 2012, 2013b). Analysis were performed using the package *deltaPlotR* (Magis & Facon, 2013a) in *R* free-software (R Core Team, 2013).

Results and discussion

A first inspection on the descriptive statistics did not reveal large dissimilarities between language versions of the items.

The different purification methods conducted yielded the same results, under a threshold of $T_\alpha = .711$. There was only one item flagged as DIF, SF49, which belongs to Straightforwardness (A2-). Figure 3 displays the Delta plot, where such item can be clearly identified as the only one rounded by a circle. Black triangles indicate that more than one item share those coordinates, while white triangles correspond to single items.

A thorough inspection at the items reveals that the differential functioning can be explained by different wording between both languages:

- **Finnish.** "Uskon, että pelkällä rehellisyydellä ei pääse elämässä kovin pitkälle. Saatan joskus petkuttaa ja käyttää toisia hyväkseni."
- **Spanish.**"Creo que la honestidad no lleva a la gente muy lejos en la vida. Cuando es necesario, intento tomar ventaja de otros."

Further revision upon the rest of the versions show that, in fact, the Finnish wording differs from the others:

- **English.**"I believe that honesty does not take one very far in life; when necessary, I try to take advantage of others."
- **German.** "Ich glaube, dass Ehrlichkeit einen nicht sehr weit bringt im Leben; ich nutze andere Leute aus, falls nötig."

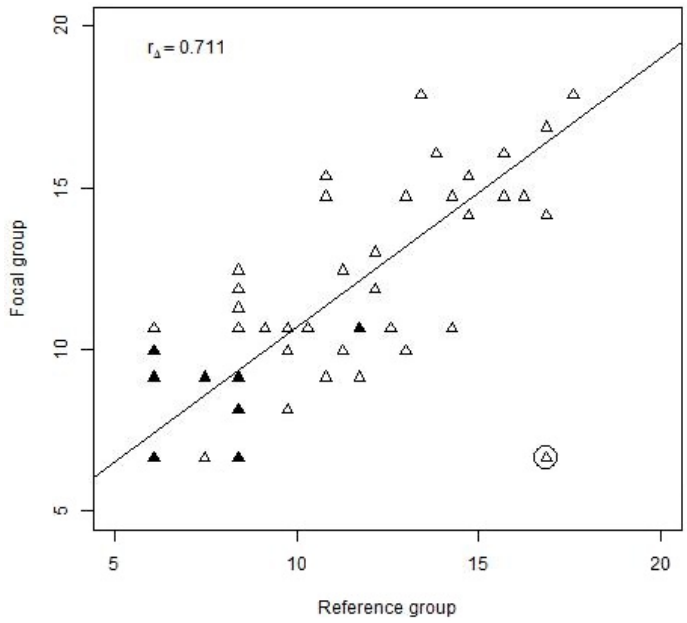


Figure 3. Delta plot from the bilingual administration of S5 (N=42).

• **Estonian.** "Ma usun, et aususega ei jõua elus kuigi kaugele. Vajadusel püüan ma teisi enda huvides ära kasutada."

The findings suggest that Finnish version of SF49 should be reviewed, given the apparent equality of all the rest language versions.

Despite of so little evidence of DIF-flagged items, it should be taken into account that Delta method is only able to detect uniformly biased items. As mentioned, due to the small amount of respondents it was not possible to perform more sophisticated DIF detection methods. In this sense, the present study holds the limitation of not addressing fully the issue, and leaves a more definite conclusion to further research.

Additionally, we agree with Sireci (2011) in that bilinguals may not be representative of the groups to which the test results could be generalized –monolingual test takers. In this sense bilingual examinees are not a sure solution for approaching the comparability of test versions (Hambleton & Kanjee, 1995; Sireci, 1997). However, their skills stand as an unique manner to identify problematic items, since language-group effects are discarded. Concerning the differences in psychological meaning, Church (2010) notes that bilingual respondents "will endorse items in the direction valued by their native culture to a greater extent", (p.155) as predicted by the *cross-cultural accommodation* hypothesis. The multiple group randomized design ensures that such bias is neutralized. Besides, as Sireci (2005) points out, if an item functions differently when administered to bilinguals, its psychological meaning is not likely to be equivalent across languages.

Gender	Secondary	Post-secondary	Graduate	Postgraduate	Total
Male	3.5%	5.3%	15.0%	3.1%	27.0%
Female	9.7%	8.4%	48.7%	6.2%	73.0%
Total	13.3%	13.7%	63.7%	9.3%	100.0%

Table 1

Distribution of the sample gathered online by gender and educational background.

Study 2

The Spanish S5 was administered to Spanish native speakers to examine its psychometric properties. Four new items for Altruism (A5) facet were added. Earlier studies showed that both items did not function correctly in terms of response pattern and social desirability rates (unpublished). Thus, the best fitting pair of items for A5 was chosen to replace the former. Then, traditional *Classical Test Theory* (CTT) analyses were performed and the instrument structure was assessed.

Method

Participants.

A pilot study was conducted as pretesting step. N=23 persons answered to the questionnaire online, and reported their impressions. The remarks were taken into account and made small corrections when necessary.

The final sample, after selecting only Spanish native speakers with complete data responses, consisted of 478 persons. Participants were gathered in a mixed-mode data collection manner. 257 students from the University of Huelva answered to a paper form including SF and other scales. The rest of the participants' data were collected by spreading the link of an online form among some universities in Spain (Sevilla, Madrid), as well as social networks as an snowball sampling process (Cohen & Arieli, 2011). These participants were given a feedback in response to their participation. In Table 1 is presented the distribution of the online sample classified by gender and educational level. Although the individuals present some variability in background, 73% were or had been university students, so it is still possible to consider it comparable to the samples from the other S5 versions, composed mainly by university students as well.

In the total dataset, the age ranged from 17 to 49, with a mean of 21.86 years. It was mainly comprised by young people, as the 95% was aged between 17 and 31. From the them, 120 were males. This features also go along with the characteristics of the other S5 samples.

Procedure.

The anonymity of the participants was guaranteed, also as an strategy for addressing social desirably response tendency (Sireci, 2011; Tourangeau & Smith, 1996). Both

sources of data had identical format and instructions.

In the paper-form administration, the scales were handed out collectively to the attendants of different lessons in first year courses of Psychology. All the inventories with missing data were rejected, as well as those with more than one answer by item.

There were no missing data in the online-form, because the respondents were forced to not to leave any unanswered item in order to be given feedback, their personal profile of values based on normative scores (ESS6, 2013; Verkasalo, Lönnqvist, Lipsanen, & Helkama, 2009). The time spent to fill out the scales was examined following Bassili and Fletcher (1991) indications concerning survey response times for different types of questions. Participants' data who completed the form in implausible timing were rejected.

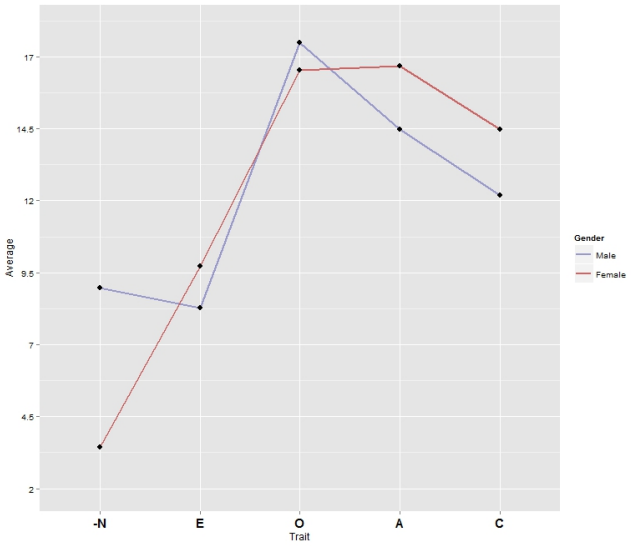


Figure 4. Average domain scores by gender (N=478).

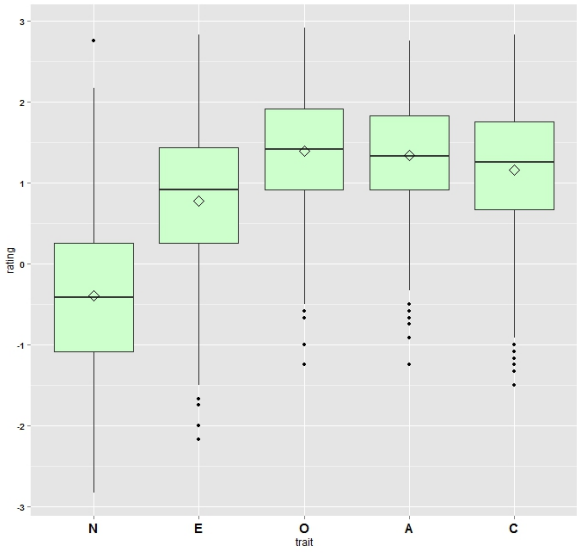


Figure 5. Box plot of response pattern (Axe Y) by S5 trait (Axe X) (N=478).

	Towards Big Five structure		Towards other S5 versions' structure	
	Proportionality (Tucker's Phi)	Identity	Proportionality (Tucker's Phi)	Identity
Neuroticism	.98	.98	.99	.99
Extraversion	.90	.89	.98	.98
Openness	.99	.99	.97	.97
Agreeableness	.92	.92	.96	.96
Consciousness	1.00	1.00	.99	.99

Table 2

Factorial congruence coefficients applied to Spanish S5 (=221).

Aiming to select the most appropriate items for Altruism (A5), the following criteria were followed: (a) highest inter-correlations with Agreeableness-trait score, (b) highest intra-facet correlation (between the two items), (c) improvement of internal consistency of the scale.

A new set of items was subsequently compiled by replacing old A5 items for a better fitting pair. All further analyses were made based on this new set, starting from descriptive and usual CTT analysis. Reliability coefficients were obtained and compared by applying Feldt's test of equal Cronbach's alpha (Feldt, Woodruff, & Salih, 1987). The comparisons supplied evidence of which alpha coefficients do actually differ statistically, taking into account sample size and number of items in each of the scales. The routine was conducted within the *R* environment (R Core Team, 2013). Functions included in the packages *CTT* (Willse, 2014), *psych* (Revelle, 2014) and *cocron* (Diedenhofen, 2013) were used.

Structure was assessed by means of a two-step procedure. First, the dataset was randomly split into two parts by 50%. Exploratory Factor Analysis (EFA) was conducted at item level on the training sample, being the orthogonal solution rotated according to the American Big Five normative structure (Costa & McCrae, 1992; Costa, McCrae, & Dye, 1991) by using targeted Procrustes rotation (McCrae et al., 1996; Paunonen, 1997). Congruence coefficients were computed for assessing structural similarity towards the standard normative structure and also concerning the rest of the S5 language versions.

In the pursuit of reaching more valid conclusions, ESEM approach was implemented for both the validation set and upon the entire dataset, in order to check for goodness-of-fit (GOF) variation according to sample. We believe that congruence coefficients and ESEM targeted-solution are conclusive tools for testing the structural equivalence of cross-cultural measures, because they give complementary, but yet different information. Procrustes-rotated EFA loadings have been conventionally utilized for calculating congruence coefficients, which gauge the degree of structural similarity -in this case we use them for assessing how it suits to the expected Big Five structure, and also for evaluat-

ing the agreement with the rest of the S5 data from the other versions. Second, ESEM modelling provide with GOF indices and other features from the SEM framework, such as modelling facets (by correlating uniqueness) or obtaining error estimators, that are not available in EFA methods. In this paper, both approaches are applied in order to maximize the information about the data gathered.

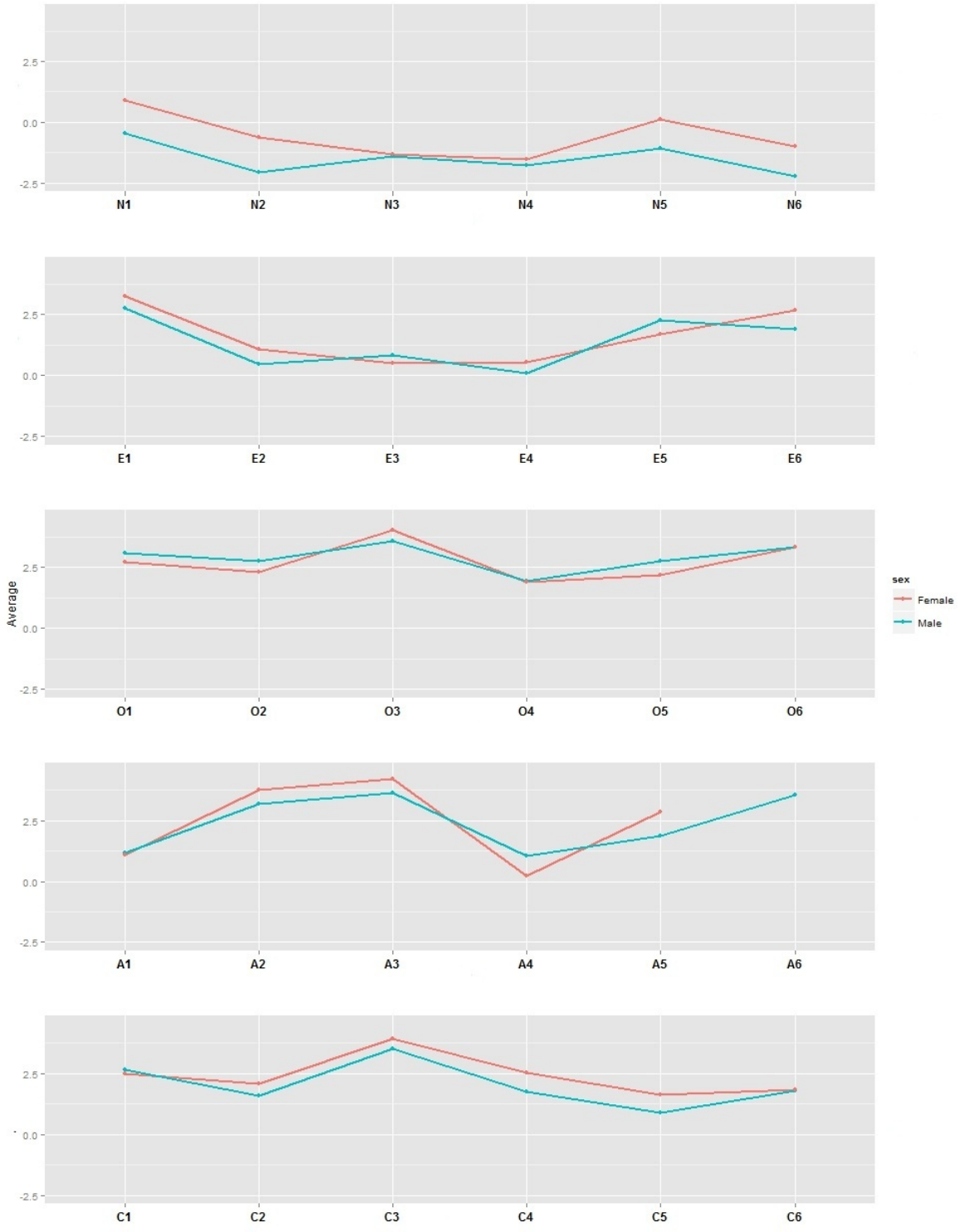


Figure 6. Average facet scores by gender (N=478).

ESEM was performed by Mplus 7 software (Muthén & Muthén, 2013). The script was generated based on a routine provided by Marsh et al. (2014) for *R* environment (R Core Team, 2013), which can be downloaded freely (Parker, 2014). Due to the a priori intended factor structure, target rotation was likewise implemented. Thus, a targeted Big Five ESEM model was conducted.

Trait	$\hat{\alpha}$	Feldt's $\chi^2(gl)$	p. value
Neuroticism	Est: .87 Eng: .87 Fin: .89 Ger: .85 Spa: .82	13.107(4)	.0108*
Extraversion	Est: .87 Eng: .74 Fin: .85 Ger: .86 Spa: .81	31.792(4)	<.001**
Openness	Est: .76 Eng: .80 Fin: .77 Ger: .81 Spa: .75	5.424(4)	.274
Agreeableness	Est: .74 Eng: .74 Fin: .72 Ger: .74 Spa: .68	4.972(4)	.290
Consciousness	Est: .85 Eng: .80 Fin: .81 Ger: .82 Spa: .81	5.207(4)	.267

Table 3

Summary of Feldt't test of equal Cronbach's alpha for the S5 versions.

Note. *Est*: Estonian, *Eng*: English, *Fin*: Finnish, *Ger*: German, *Spa*: Spanish. *p<.05. **p<.001.

Results and discussion

First, all A5 items were examined. The items with highest correlations to the trait score (all the other Agreeableness facets) were SF62 (r=.203) and SF63 (r=.287). These values also overtook the former items-trait correlations. These both items were too, as reasonably, the ones with strongest associations with the rest of the trait at the item level. The difference in internal consistency when replacing the former items by this pair was meaningful, increasing from $\hat{\alpha} = .61$ to $\hat{\alpha} = .68$. Item statistics were revised, comparing them to the others in the scale. Means and standard deviations were in harmony, and the distribution of the response displayed an homogeneous pattern.

Fortunately the pair of items still respect the structure, as one expresses A5 positively (SF62: *"Nadie me describiría como arrogante o engreído. No alardeo sobre mí o mis logros."*) and the other, absence of modesty (SF63:*"Creo que soy en varios aspectos mejor que otros, y por ello merezco más atención."*). Hence, this pair of items was selected to constitute A5. Descriptive information about S5 scores at both facet and domain level, split by gender, is shown in Figures 4 and 6. In Figure 5, a box plot pictures the distribution of the response patterns by domain.

Targeted Procrustes-rotated loadings of the EFA are provided in Table 12. The targeted five-factor model was able to explain 35% of the common variance. Tucker's factor congruence coefficients showed good values (Lorenzo-Seva & Ten Berge, 2006). The identity coefficients, as the most stringent type, exhibited also fair to good indices (all congruence coefficients in Table 2). Six items had only their second highest loading in the intended factor.

The congruence results for equivalence towards the other S5 version's structure, all of them above .95, suggesting that the measure can be considered structurally very similar.

Cronbach's alpha (Table 3; Cronbach, 1951) registered acceptable to good values in all scales. Although Agreeableness reliability rates as acceptable, it is systematically the lowest domain in terms of internal consistency for all S5 language versions (Konstabel et al., 2012). This same tendency has been reported other Big Five instruments around the world (Schmitt et al., 2007). Facet internal consistency is also shown in Table 6. Note that every facet is composed by only two items.

Performed the Feldt's test of significance between alpha coefficients, the null hypothesis of equality was retained for all S5 versions (Estonian, Finnish, English, German and Spanish) in the case of Openness, Agreeableness and Consciousness (Table 3). For these traits reliability can be considered as equal.

Concerning Neuroticism, Spanish Cronbach's alpha was equal to all versions excepting the Finnish. It was only possible to retain the null hypothesis of all versions as equivalent when one of the extreme coefficients (Finnish or Spanish) was excluded from the computation. The same case can be observed with Extraversion reliability, where the

	N=239	N=478
χ^2/df	2339.229/1450	2773.979/1450
RMSEA *90% C.I.*	[.047,.054]	[.041,.046]
CFI	.812	.853
RMSR	.050	.045

Table 4
Summary of Goodness-of-fit statistics for ESEM model of S5 according to the normative American structure of the Big Five.

German scale showed a considerably smaller $\hat{\alpha}$. When all but this value are included in the Feldt's test, null hypothesis is retained, thus Extraversion Cronbach's alpha of the Estonian, English, Finnish and Spanish versions can be considered as equal.

The ESEM model was run with 50% validation subset N=239, as well as the whole N=478 dataset. The reason is, as it can be seen from Table 4, that the fit indices varied with the sample size, which was the only difference in between both ESEM model estimations. It is not surprising that the χ^2 value, in both sample sizes, does not give support for perfect-fit model, given the characteristics of the size and estimations. Nevertheless, RMSEA provides evidence for a close-fit model and RMSR registered good values in both N=239 and N=478.

The poor CFI statistic value could be explained as affected by two phenomena which actually define the present data: (a) CFI penalizes every parameter estimated (440 free parameters in this case) and (b) CFI may be underestimated, as it is affected by weak correlations in the data (Kenny, 2014). Average absolute value of the correlation matrix in this case is r=.127. Moreover, around 95% of the correlations had an absolute value equal or smaller than 0.35.

It should be born in mind the recentness of ESEM as SEM-framework approach. At this respect, Morin et al. (2013) point out that "[...] the total number of parameter estimates in ESEM applications can be massively more than in CFA. This feature might make problematic any index that does not control for parsimony (due to capitalization on chance), and yet might call into question the appropriateness of controls for parsimony in indices that do. In the meantime, we suggest that applied researchers use a multifaceted approach based on the integration of a variety of different indices, detailed evaluations of the actual parameters estimates in relation to theory, a priori predictions, and common sense" (p.405). This is the case of CFI index.

Standardized ESEM loadings are presented in Table 13. All targeted loadings are displayed, and cross-loadings are shown when being higher than the target loadings, or above .4. The model showed substantial cross-loadings (>.4) in the case of six items. SF52 (E6+, Positive emotions) and SF60 (C1-, Competence) had a noteworthy load on N (both negatively related). Both items of the facet Self-consciousness (SF26 and SF31)

registered a meaningful relation to E, as well as one Openness to actions item (SF33+). In Openness, SF42 (Excitement seeking, positive item) registered a relevant association, although still smaller than the loading to the targeted trait Extraversion. The average R^2 was .321.

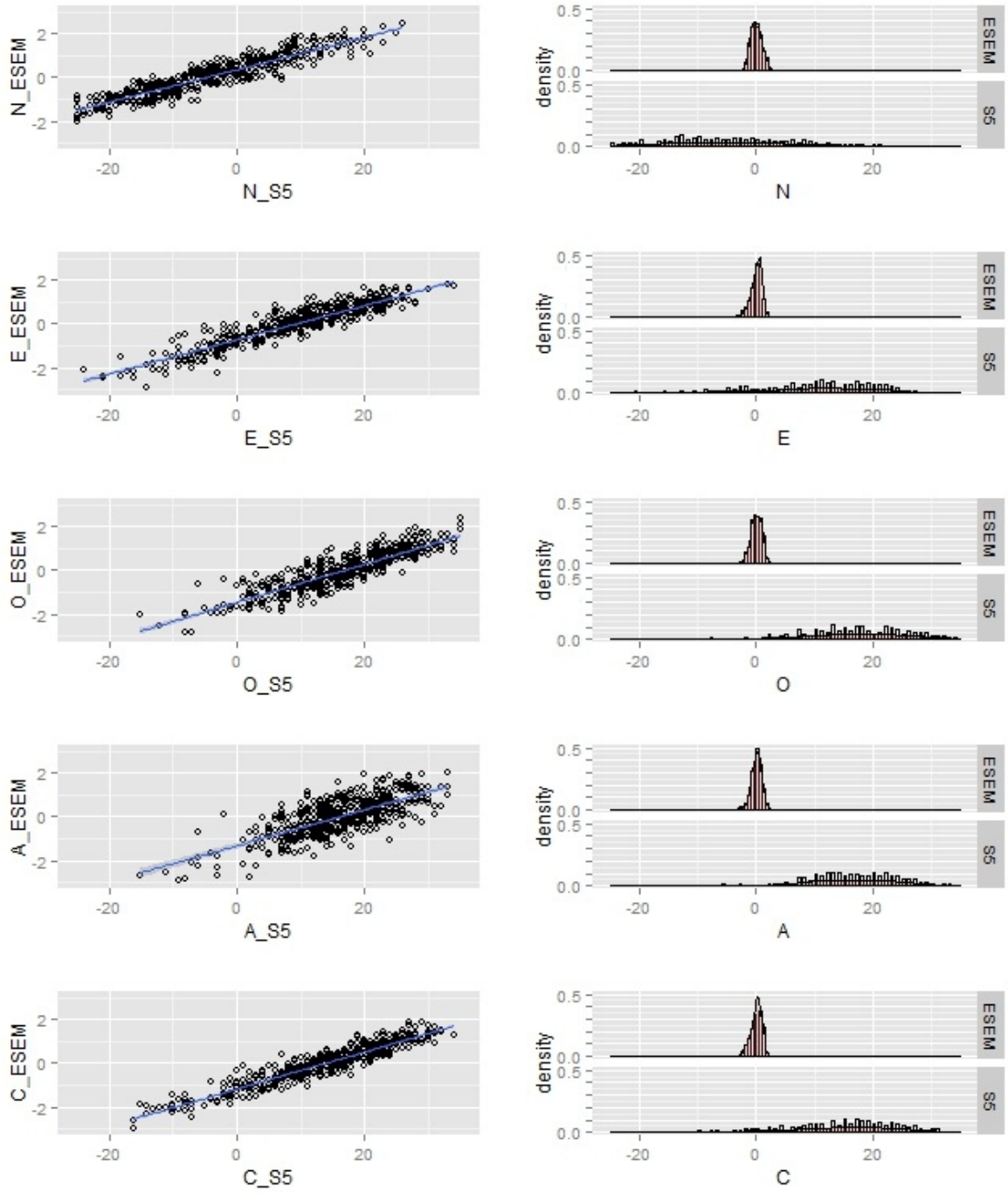


Figure 7. Left panels: Correlation between S5 scores and S5 ESEM-estimated scores. Right panels: Distributions of S5 scores and S5 ESEM-estimated scores (N=478).

We acknowledge that, in both ESEM and EFA solutions, Agreeableness domain exhibits unacceptably poor loading values. It should also be noted that A5 items hold the

highest weights. In this sense, it seems that the rest of the domain should be revised likewise and selected new set of items in order to improve this insufficient performance, as done with A5. However, introducing such changes is rather complex, because the process should be conducted for all existing versions together and total agreement ought to be reached in all together.

Correlations between ESEM trait scores and S5 scores were also computed. Herrmann and Pfister (2013) underlined the potential bias introduced in personality test scoring when complex SEM models are applied to fit the data. Although the proposed model is quite simple, ESEM estimation requirements make it more complex than a CFA with the same specifications. By estimating score correlations obtained by different methods it can be observed how much the model estimations respect the rank order of the original measure, although the ESEM scores are normally distributed. Hence, it is also detected up to what extent the complexity of the model affects the conventional score ordering. In this case, correlations between S5 score computation (sum of facet scores) and ESEM score estimations where: 0.93, .92, .89, .75 and .94 for N, E, O, A, and C respectively (depicted in Figure 7). Domain scores where uncorrelated, since factors are specified as orthogonal in BF model.

Correlation values highly respect the ranking with the exception of A, which has already shown slightly poorer results in comparison to the rest of the scales. Worth to note that ESEM estimated scores follow normal distribution $N(0, 1)$, allowing for more robust estimations and easier to meet assumptions. As can be examined in Figure 7 the distribution of the scores presents some potential advantages without distorting the score ordering.

Study 3

Analysis of item bias regarding gender are carried out upon the Spanish S5 dataset.

Method

Participants.

Same dataset as in Study 2.

Procedure.

DIF analyses were conducted through logistic regression procedure (Swaminathan & Rogers, 1990), by utilizing the *difR* package (Magis, Beland, & Raiche, 2013), designed for *R* environment (R Core Team, 2013). For this purpose, items were dichotomized depending of endorsement (response categories 1 to 3) or not (from -3 to 0). DIF analyses were carried out separately for each of the five scales, being assured unidimensionality of the domain-test score. The formulation of the model as follows:

$$P(x_j = 1|X = x, G = g) = \frac{1}{1 + e^{(-\beta_0+\beta_1 x+\beta_2 g+\beta_3 xg)}} \quad (1)$$

where the conditional probability of endorsing the j item, depends of three terms: the x score in the test, the g group to which the respondent belongs (gender, in this case) and the interaction between these two terms (xg).

When assessing DIF, a series of comparisons of nested models are conducted to test whether the different terms (β_0,β_1, β_2 and β_3) statistically contribute to predict the response to a given item. Likelihood Ratio Test (LRT) are conducted first upon the most saturated model (the one including the interaction between scores and group, β_3) in order to check out for *non-uniform* DIF. The most parsimonious model will always be selected by means of LRT. If discarded the complex one, a model composed by both terms group and scores (β_1 and β_2) is compared with the model where only the scores term (β_1) is left along predicting the item endorsement, with β_0, present in all models as the intersect. This second step will define the presence of *uniform* DIF. R^2 difference between nested models in interpreted as a size effect measure.

Results and discussion

Items with DIF are reported in Table 5. From the eight items, only one displayed *non-uniform* DIF, showing the rest *uniform* differences under $\alpha = .01$. The R^2 values are interpreted in the aforementioned table following the criteria proposed by Jodoin and Gierl (2001). The thresholds are: negligible effect for $R^2 < .035$, moderate DIF for R^2 between .035 and .070, and large effect when obtaining $R^2 > .07$

The DIF plot of SF49 can be observed in Figure 8. As it shows, males where more likely than females to endorse Straightforwardness when rating low on other Agreeableness features. Nevertheless, when scoring high on the A trait, the chance of males does not increase along with the endorsement of the other Agreeableness items. This is consistent with a significantly lower score in A2 for males in comparison to females by a $t(167) = -3.24, p = 0.002$.

Item		Higher endorsement	β_1	β_2	Nagelkerke's R^2 size	Effect
SF04	Trust (A1)	Males		9.708	.040	Moderate
SF21	Depression (N3)	Males		9.834	.032	Negligible
SF22	Assertiveness (E3)	Males		18.707*	.105	Large
SF34	Compliance (A4)	Males		20.569	.082	Large
SF43	Op. to ideas (O5)	Males		10.252	.018	Negligible
SF46	Angry-hostility (N2)	Females		11.425	.067	Moderate
SF49	Straightforwardness (A2)	Females	6.875		.099	Large
SF56	Anxiety (N1)	Females		20.651*	.123	Large

Note. All β coefficients were significant at $\alpha = .01$ *p<.001.

Table 5

Summary of S5 DIF item analysis by Logistic Regression procedure.

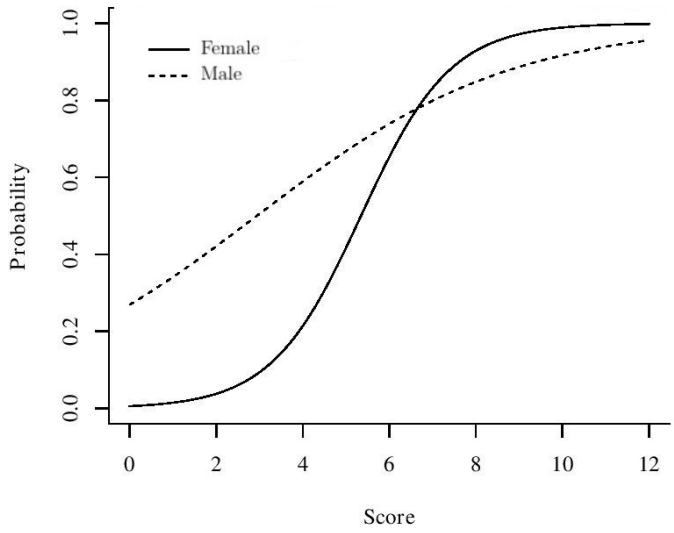


Figure 8. *Non uniform* DIF by gender, exhibited by SF49 (A2 Straightforwardness).

All items flagged with DIF will be examined in further administrations. In concrete, item SF49 will be more carefully reviewed and compared DIF with other S5 versions. Possible modifications on the item wording will be considered.

Study 4

Here evidence of construct validity is provided through the examination of additional measures.

Method

Participants.

As mentioned, two sub-samples compose the whole dataset. Due to time restrictions, the participants who answered to the paper version (N=257) were administered S5 and the Spanish validated version of NEO PI-R (Pando, Pamos, Cubero, & Costa, 1999). The ones who responded online (N=221) completed S5, the Behavioural Report Form (BRF; Paunonen 2003) and the Portrait Values Questionnaire (PVQ).

Measures.

- *S5.*
- *NEO-PI-R.* Spanish validated version of the self-reported, 240-items NEO inventory. Normative scores for youth and general population where utilized.
- *BRQ.* Self-report measure designed to evaluate several complex behaviours of personal, social and cultural relevance, such as drinking and smoking customs, participation in sports or fraternity belonging.

• *PVQ*. This measure evaluates motivational values as the taxonomy by Schwartz (1992; 2012). The 21 items describe different people whose goals, aspirations and wishes are characterized. The response options rank from 1 to 7 in a Likert-type agreement scale, choosing how much the description could be applied to the respondent. This 21-items version is a short one used in the European Social Survey, based on the PVQ-40 developed by Schwartz (2012).

Procedure.

The correlation matrices of both S5 and NEO-PI-R were analyzed through the network approach, in order to have a first insight of the structure of both questionnaires. The routine was carried out by using the *qgraph* package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012), implemented in the *R* environment.

In order to examine construct validity, multitrait-multimethod matrix method (MTMM; Campbell and Fiske 1959) was carried out by dint of evaluating the BF traits (N, E, O, A, C) scores through both S5 and NEO PI-R as method measures. To this end, a CFA was modelled by Mplus 7 (Muthén & Muthén, 2013) were five trait factors were correlated to each other, and two method factors were orthogonal to all the rest. The *equal loadings model* (Kenny & Kashy, 1992) was specified, whose general feature is that loadings on each trait-factor are constrained to be equal, although loadings on the method factors are estimated freely. This feature reduces the convergence difficulties by fixing parameter values, and so it is also a rather parsimonious model with more degrees of freedom, that assure identification. Fixed-factor method was chosen so the variances of latent variables are fixed to 1. Two models were proposed differing on whether trait factors were allowed to correlate or not, in order to revise the associations among the Big Five traits. Worth to mention that a model with freely estimated factor loadings was also tested, and no convergence was reached. The *equal loadings model* was chosen as more parsimonious, and even though being more stringent, GOF showed good values. No estimation problems were encountered in this case.

To sum, degree of convergence with NEO PI-R is studied by two different means. First, as typically conducted, MTMM matrix method implemented by CFA supplies statistical indices of model fit and informs concerning convergent and discriminant validity of the personality traits, which is free of method effects. On the other hand, we consider that the network approach offers a different insight regarding the inner configuration of the instruments. This procedure was administered as a way to compare graphically how facet in both questionnaires related to the rest of the components.

Paunonen's version of BRF (2003) was administered for gathering evidence concerning criterion validity. Some small changes were introduced. First, as done previously in the other S5 validations, a single variable composite was created as self-enhancement, collecting scores from self-ratings of attractiveness, intelligence and popularity. Other self-perceptions not related to behaviour were excluded (such as religiosity, femininity or

honesty). Third, and only in the case of the Spanish version, variables related to working behaviour (to have or want a part-time or full-time job) were discarded, as the situation of the youth in the country is strongly affected by the economic circumstances, most likely biasing the personality-based inferences regarding work.

One multiple regression model was specified for predicting each of the behavioural indicators. The scores of all five traits were included as predictors. This decision was made according to previous estimations of this measure, most commonly made at the trait level. R^2 was used as indicator, so comparisons with the other S5 versions can be made. No significant violations of assumptions, such as multi-collinearity or homocedasticity, were encountered.

Additional measures of construct validity were gathered through PVQ. Values have been a central concept in the social sciences, since they have been utilized to characterize cultural groups, societies and individuals, to track change over time, and to explain the motivational bases of attitudes and behaviour (Schwartz, 2012). Schwartz's theory comprises ten different motivational values, and postulates the dynamic relations among them in terms of compatibility or conflict (for instance, benevolence and power). Among their features, it can be said that values are beliefs which refer to desirable goals that motivate actions, transcending specific actions and situations. There has been found support to consider values and their structure universal (Bilsky, Janik, & Schwartz, 2011; Davidov, Schmidt, & Schwartz, 2008; Schwartz, 2006).

In words of Schwartz (2012), "traits are tendencies to show consistent patterns of thought, feelings and actions across time and situations. Values are said to be a central component of our self and personality, distinct and related to attitudes, beliefs, norms and traits" (p.16). The main difference between the concepts of value and trait is the scale to measure them: values vary on importance at leading to evaluation, and traits vary in frequency and intensity of exhibition. In spite of the fact that the relationship between values and personality is complex and difficult to disentangle, it seems justified to consider that both might interact on predicting behaviour (Parks & Guay, 2009). Moreover, Locke (2006) considers both personality and values as same type of components in his integrated model of work motivation.

The core idea about their connection is that traits describe what people are like, and values what people consider important. As Olver and Mooradian (2003) argue, personality may have an relevant role on the development on one's values, as it seems reasonable that, for instance, a person who scores high in Agreeableness might decide that a value such as *benevolence* is more important than *power*. Under this perspective, low to moderate associations are expected between personality traits and values. In case of confirming the expectations, the findings may also be in consonance with previous studies (Aluja & García, 2004; McCrae & Sutin, 2009; Parks & Guay, 2009; Roccas, Sagiv, Schwartz, & Knafo, 2002).

Associations among self-reported personality and values were addressed by means of partial correlation controlling for sex, as it has been suggested that it may introduce bias (Schwartz & Rubel, 2005; Struch, Schwartz, & Van Der Kloot, 2002). Estimations were computed for both trait and facet level. Trait scores are the usual unit for this purpose. In this case, facet scores are also applied so that more accurate information is obtained about the components of the traits which relate the most to certain values.

Results and discussion

Reliability information and comparisons, as well as facet-level correlations between S5 and NEO PI-R are reported in Table 6.

Some things should be born in mind when interpreting Figure 9. The representation is based on the correlations at the facet level, each one of these facets being a node. Nodes are connected by weighted-thick lines depending on the magnitude of their association.

Only correlations higher than $|r| > .1$ are displayed. The spring-based algorithm locates towards the centre, and closer together, the most tighly linked nodes (Fruchterman & Reingold, 1991). From this point of view, the more distal a node (facet) is, the less related it is to other features of personality, as it posits the notion of central versus peripheral components (Cramer et al., 2012).

As it can be seen from Figure 9, all traits are well grouped together in both questionnaires, with the exception of Trust (A1) and Compliance (A4) in Agreeableness domain. The strength of the (inverse) association towards Neuroticsm facets, and the closeness to Extraversion seem to be an explanation for their location. It can also be observed that Extraversion plays a central role in both S5 and NEO PI-R, and most of the peripheral components concur for both questionnaires. Worth to mention in both cases that Impulsiveness (N5) seems to mediate the association between Consciousness and Neuroticism, specially through the strong correlation with Deliberation (C6), quite a peripheral component otherwise. A difference between both questionnaires, on the other hand, and obvious difference is the amount and intensity of the links (lines), that can be explained by the variability in the questionnaires, that differs noticeably: while NEO-PI-R facet values (composed by eight items each) have a rank average 25.67, S5 facet values (two items) can only range from -6 o +6, and registered a rank mean on 11.67. This fact automatically increases the strength of correlations.

Concerning S5 structure, the network representation (NR) bears obvious resemblance to the EFA-ESEM results. The benefit of NR is their contribution to disentangle the nature of a factorial solution. For instance, Positive emotions (E6) loads in N as it is strongly (and negatively) related to facets such as Depression (N3) and Self-Consciousness (N4). Such a phenomenon occurs also the other way round: N4 heavily loads on E as it is fully linked to many E facets. Openness to actions (O4) plays an interesting rol as the most nuclear node of the Openness domain, by defining strong connections with E and N,

all of them fairly reasonable. Worth to mention how Excitement Seeking (E5) relates to Openness facets, as it loads in the ESEM solution. Apart from A4, Openness to Feelings (O3) is also centrally widely connected to Extraversion and Agreeableness, although not intensively.

Facet	$\hat{\alpha}^{S5}$	r_{S5-NEO}	$\hat{\alpha}^{NEO}$	Feldt's $\chi^2(1)$	p. value
N1 Anxiety	.64	.68	.74	4.694	.0303*
N2 Angry-Hostiliy	.61	.607	.74	7.289	<.001***
N3 Depression	.78	.69	.82	1.782	.182
N4 Self-conciousness	.65	.62	.69	.651	.4198
N5 Impulsiviness	.65	.61	.62	.297	.586
N6 Vulnerability	.63	.72	.79	14.202	<.0002***
E1 Warmth	.76	.61	.72	1.039	.308
E2 Gregariousness	.47	.74	.74	22.380	<.0001***
E3 Assertiveness	.64	.60	.71	2.070	.150
E4 Activity	.42	.58	.71	21.21	<.0001***
E5 Excitement seeking	.81	.54	.59	24.683	<.0001***
E6 Positive emotions	.68	.67	.82	14.653	.0001****
O1 Fantasy	.65	.63	.80	13.866	.0002***
O2 Aesthetics	.83	.70	.78	2.891	.0891
O3 Feelings	.67	.43	.72	1.194	.2745
O4 Actions	.59	.59	.64	.748	.3873
O5 Ideas	.52	,64	.80	33.613	<.0001***
O6 Values	.32	.34	.57	9.313	.0023***
A1 Trust	.67	.69	.72	1.194	.2745
A2 Straightforwardness	.40	.47	.69	19.268	<.0001***
A3 Altruism	.41	.38	.71	22.2458	<.0001***
A4 Compliance	.63	.61	.72	3.442	.0636
A5 Modesty	.50	.57	.73	16.793	<.0001***
A6 Tender-mindness	.51	.31	.59	1.406	.2358
C1 Competence	.60	.48	.61	.082	.8665
C2 Order	.79	.76	.80	.105	.7459
C3 Dutifulness	.63	.45	.63	.000	1.0000
C4 Achievement striving	.62	.62	.78	13.228	.0003***
C5 Self-discipline	.74	.71	.84	10.446	.0012**
C6 Deliberation	.76	.73	.81	2.417	.1200

Table 6

Summary of facet reliabilities and correlations of Spanish S5 and NEO-PI-R.

*p<.05. **p<.01. ***p<.001.

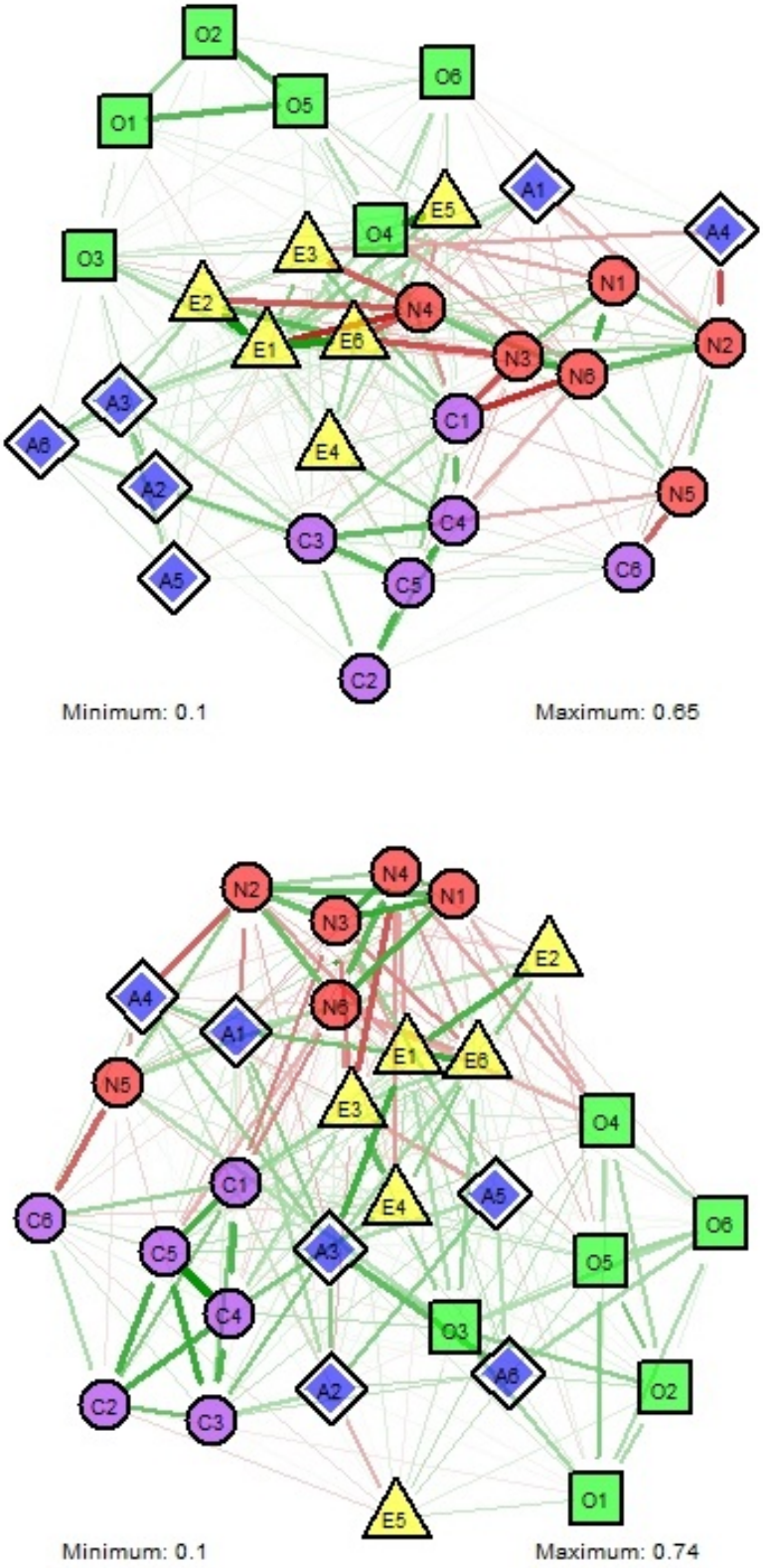


Figure 9. Network representation of the thirty S5 (Up panel) and NEO-PI-R (Down panel) facets.

		NEO-PI-R					Short Five				
		N	E	O	A	C	N	E	O	A	C
NEO-PI-R	N	**(.91)**									
	E	-0.36	**(.87)**								
	O	-0.14	0.27	**(.88)**							
	A	-0.09	0.11	0.05	**(.86)**						
	C	-0.23	0.2	0.08	0.22	**(.92)**					
Short Five	N	0.83	-0.32	-0.13	-0.13	-0.19	**(.82)**				
	E	-0.34	0.76	0.16	0.02	0.07	-0.38	**(.81)**			
	O	-0.13	0.22	0.7	-0.06	-0.03	-0.16	0.32	**(.75)**		
	A	-0.2	0.06	-0.07	0.64	0.04	-0.3	0.22	0.12	**(.68)**	
	C	-0.34	0.21	0.01	0.1	0.78	-0.38	0.27	0.14	0.21	**(.81)**

Table 7

Multimethod-Multitrait matrix of Spanish S5 and NEO-PI-R (N=257).

Unfortunately, it is not possible to examine further replicability, since there is no other available network representation of a Big Five-inspired instrument at the facet level.

Although there is a fruitful controversy about the implications of the network approach in what comes to modelling personality (Asendorpf, 2012; Ashton & Lee, 2012; Franìc, Borsboom, Dolan, & Boomsma, 2013; Rothmund, Baumert, & Schmitt, 2012; Schimmack & Gere, 2012), those arguments will not be discussed in here as it is not the purpose of this study. Nevertheless, it is hard to deny that network representations provide an intuitive and useful insight when picturing associations among the components of a correlation matrix. Moreover, it cannot be neglected that it is more frequent than desirable the need for allowing cross-loadings in confirmatory models of Big Five-based instruments. We have the impression that the network approach could assist in the task of disclosing which of these associations are substantially relevant and normative, based on empirical findings.

MTMM correlation matrix is shown in Table 7. Note that in monomethod S5 cell values correspond to the table entire sample (N=478). The correlation matrix from where MTMM-CFA model was estimated was N=255, although it has been considered that the association among S5 traits was better to observe based on full data information. Fit indices for the model were traits were constrained to be orthogonal (no correlations) were: $\chi^2(30, N = 255) = 110.347$, p<.001, RMSEA (90% confidence interval) =[0.082, 0.123], CFI=.943, SRMR=.0145. Fit indices for correlated-traits model registered the following values: $\chi^2(20, N = 255) = 33.313$, p<.001, RMSEA (90% confidence interval) =[0.015, 0.080], CFI=.991, SRMR=.032. Despite the value of the correlations among MTMM traits being from low to moderate values (highest value -.414), model fit results suggest that associations among traits are significantly important, because GOF values for the parsimonious model were unacceptable.

Estimated parameters for the MTMM-CFA in Table 8. Loadings on method factors (average .304) are systemically lower than the coefficients for trait-factors (average of .863), substantiating convergent validity. Average of residual variances with value .149 supports convergence as well. In terms of discriminant validity, low to moderate correlations among traits hold (Table 9, average of absolute values .192). In this sense, as claimed by several authors, the magnitude of the associations between traits which has been estimated free from method factors offers more reliable information about the real correlations that the ones calculated upon one single measure tool (Anusic, Schimmack, Pinkus, & Lockwood, 2009; Schimmack, Schupp, & Wagner, 2008).

Values in the MTMM matrix provide further information. In favour of convergent validity, heteromethod-monotrait correlations are highest values apart from the reliability

Parameter	Estimation	S.E.	p. value
Neuroticism			
NEO-PI-R	0.903	0.044	<0.001**
S5	0.903	0.044	<0.001**
Extraversion			
NEO-PI-R	0.870	0.045	<0.001**
S5	0.870	0.045	<0.001**
Openness			
NEO-PI-R	0.846	0.045	<0.001**
S5	0.846	0.045	<0.001**
Agreeableness			
NEO-PI-R	0.798	0.046	<0.001**
S5	0.798	0.046	<0.001**
Consciousness			
NEO-PI-R	0.897	0.045	<0.001**
S5	0.897	0.045	<0.001**
NEO-PI-R Method			
N	0.065	0.061	0.284
E	0.202	0.058	0.001*
O	0.284	0.065	<0.001**
A	0.469	0.082	<0.001**
C	0.282	0.057	<0.001**
Short-Five Method			
N	-0.217	0.050	<0.001**
E	0.356	0.050	<0.001**
O	0.383	0.055	<0.001**
A	0.412	0.059	<0.001**
C	0.372	0.047	<0.001**

Table 8

Multimethod-Multitrait Confirmatory Factor Analysis estimates (N=257).

*p<.01. **p<.001.

	N	E	O	A	C	NEO-PI-R Method	S5 Method
N	-					0	0
E	-.414	-				0	0
O	-.165	.272	-			0	0
A	-.238	.071	-.079	-		0	0
C	-.358	.198	.020	.109	-	0	0

Table 9

Correlations among Multimethod-Multitrait factors (N=257).

	R^2	Significant predictors ($\alpha = .05$)
Participant sex	.178*	N A C
Alcohol consumption	.059*	C
Driving fast	.071**	E O A
Self enhancement	.255***	N E A
High school GPA	.055*	A C
Traffic violations	.048	E
Routinely exercises	.074	E
Participation in sports	.054*	E C
Tobacco consumption	.084**	A C
Fraternity interest	.071**	O
Parties attended	.135***	E C
Last year GPA	.118***	A C
Plays musical instrument	.047	E O
Blood donations	.030	
Dieting behavior	.012	
Buys lottery tickets	.157*	E O
Medication usage	.021	

Table 10

Prediction of BRF criteria by S5 domain scores (N=221)

*$p<.05$. **$p<.01$. ***$p<.001$.

estimates (diagonal line). Along with it, monomethod-heterotrait correlations registered low associations. Heteromethod-heterotrait correlations were, as expected, lowest of all values obtaining in most of the cases non-significant estimates.

Although the topic is out of the purposes of this study, it is worth to note that, in terms of the long discussed Alpha-Beta Model (Digman, 1997) results do not support the structure predicted as correlations between N, A anc C on one hand, and E-O on the other, do not reflect distinctive associations.

R^2 values of the predicted BRF variables are presented in Table 10. Values are, on average, lesser than the ones registered in the German validation (Konstabel et al., 2012), although there is strong inter-version agreement in terms of which models are significant, with the exception of parties attended, fraternity interest and last year GPA,

which are only significantly predicted in the Spanish S5. Trait scores predicted on average 11,25% of the variance in the behaviours contained in BRF. Statistically significant predictors mostly concurred with those involved in significant partial correlations reported by Paunonen (2003).

Partial correlations among Big Five traits and Schwartz's values are displayed in Table 11. The interpretation of the statistically significant associations makes sense for all traits, for instance Extraversion is substantially related to *self-direction* and *stimulation*, and Agreeableness is connected distinctively to *benevolence*, *universalism* and against *power*. Moreover, as broadly found Neuroticism is weakly related to any values, and so also happens to a lesser extent with Consciousness. A thorough examination at the values suggests that results are compatible with content validity expectations.

A further step focuses on which concrete components of personality relate to values. Due to the big size of the matrix of correlations (10 values times 30 facets of estimates), only the most relevant findings will be commented. First, as foreseen no relevant associations with Neuroticism (35% of all possible 60 significant correlations). Only the link between Vulnerability (N6) and *self-direction*, value r= -.304 was over .30. In the case of Consciousness (30%), no facets reached this threshold. Contrary was the case of Extraversion, from where *stimulation* attained r=.76 with Excitement Seeking (E5), as well as Warmth (E1; r=.373), Gregariousness (E2, r=.338) and Positive emotions (E6; r=.318). Assertiveness (E3) was also positively related to *achievement* (r=.402) and negatively to *traditionalism* (r=-.391). Apart from the aforementioned relationship to *stimulation*, E5 was found to be negatively related to *security* and *conformity* (r=-.387 and r=-.342, respectively).

Regarding Openness, although many significant associations were registered (58%) only few met the .3 edge. The most substantial value, r=.608, was obtained between *stim-*

Schwartz's value	N	E	O	A	C
Benevolence				.446**	
Universalism			.236**	.379**	
Self-direction	-.197*		.358**		
Stimulation	-.205*	.527**	.432**		
Hedonism		.174			-.218*
Achievement				-.323**	
Power	.146		-.334**	-.422**	
Security		-.209	-.259**	-.146*	
Conformity	.164	-.299**	-.35**		.175*
Tradition		-.322**	-.273**	.215*	

Table 11

Partial correlations controlling for sex between S5 traits and Schwartz's motivational values (N=221).

Note. Only significant correlations ($\alpha = .05$) are reported. *p<.01. **p<.001.

ulation and Openness to Actions (O4). Yet it was related negatively to *security* (r=-.316). Openness to ideas (O5) registered a correlation of ,344 with *self-direction*. Openness to values (O6) attained relations to *universalism* (r=.367), *stimulation* (r=.329), and *power* (r=-.317).

Agreeableness acquired 40% of statistically significant correlations, most of the substantial ones related to *benevolence* (Altruism, r=.452; Modesty, r=.390; Tender-mindness, r=.340). Also along with previous findings, *power* showed to be negatively linked to Altruism (A3; r= -.323) and Modesty (A5; r=-.375). Interestingly, Modesty was additionally associated to negative scores in *achievement* (r=.493) and as expected, to *universalism* (r=.346).

The aim of this comprehensive examination was, along with contributing to obtain content validity evidence, attempting to disentangle the affinity between values and personality in a more fine detailed way. In this sense, there are couple points worth to be made. First, the size of the correlations with the traits does not differ much from the facet-value correlations (significant absolute values ranking from .135 to .511), even presenting higher associations than the trait level-value such as the cases of stimulation with E5 and O4. This point supports the idea that both construct maintain an indisputable relationship. In what comes to S5, results are consistent with previous studies that administered different measure tools, contributing to bolster it as an appropriate short personality measure.

General discussion

The present work introduced the Spanish adaptation of S5 as an instance of the fact that comprehensive single items can be successfully utilized. It is acknowledged that short questionnaires hold some shortcomings in comparison to longer ones. S5 rationale provides an example of the advantages offered by shorter measures, and how balance at this point should be pursued according to the assessment goals. In research contexts and especially when additional measures will be gathered along with personality data, there is a substantial risk of time consuming and exhausting evaluation, which likely leads to biased results. S5 questionnaire stands as a good choice for these kind of settings.

Apart from practical reasons, more theoretical ones could be considered. For instance, measuring personality with broad, comprehensive items has particular implications. A first impression suggests that dimensionality is not so accurately defined and the items contain more error term, because less of their variance can be explained by the factors. On the other hand, comprehensive items are genuinely lifelike, as personality is not a set of separate boxes, but a tendency that is expressed upon different, more or less ambiguous, situations through behaviour. Short Five, as a comprehensive tool, provides an interesting point of view to study personality itself and how components relate to each other from a new, global perspective. For example, Self-consciousness (N4) has turned

out to be in Spanish culture, as measured by S5, strongly linked to Extraversion instead of Neuroticism. This could suggest that, for a collectivist culture given an ambiguous and broad statement (item), a social set up is more likely to be identified to Self-consciousness rather than emotional stability. This point hints that personality configuration could be still studied as modulated by the context, still under the scope of the Big Five traits, and still going along with more or less strict theoretical settings.

Also by means of the "ambiguous" nature of comprehensive items could be explained some of the low results shown by Spanish S5. As Church (2010) notes, personality inventories tend to give minimal specification of situational context, in order to be more generalizable to different cultures. The inconvenience is, more concrete items are less representative and contextual, with the chance of validity loss. Broad items (those from S5) call for an intuitive aggregation of one's behaviour across different situations, that respondents from collectivist cultures may find troublesome, since greater emphasis is placed on situational factors that influence feelings and behaviour (Marsella, Dubanoski, Hamada, & Morse, 2000). A classical example to illustrate this idea is offered by Marsella and Leong (1995), quoting a respondent from Philippines: "Sir, sometimes true and sometimes false. I cannot tell you true or false all the time" (p.208). Hence, this could lead to the slightly lower validity and reliability measures shown by the Spanish S5 in comparison to the other versions. For instance, if less consistent responses are given to the items composing a personality facet, less internal reliability will be obtained and the increased chance of diffused dimensional structure.

In this sense, one of the potential limitations of S5 items' rationale is a more vague dimensionality. This has been reflected in factor loadings, but also in low internal consistency of Agreeableness domain, for instance. Despite of some blunt factor loadings, S5 studies have provided with substantial evidence of construct validity. MTMM displayed solid convergence and discrimination validity values for personality domain, and so have shown theoretically expected associations between Schwartz's motivational values and BF facets. Network representations, as a descriptive tool, have offered an interesting insight of how personality correlation patterns distribute, bringing out a noticeable underlying structure of the Spanish Short Five which is highly consistent with BF principles. The prediction of behavioural criteria through Paunonen's BRF demonstrates both the relevance of the information provided by S5, as well as the comparability of S5 versions in terms of prediction.

Precisely, as an essential part of this job, it was aimed to ensure an check equivalence of Spanish S5 (a) towards the BF model, and (b) regarding the rest of the versions. This has been successfully proved by a cross-validation procedure in which the most stringent congruence coefficients suggest solid similarities. DIF analysis revealed a remarkably equivalent functioning of the test, clearly outperforming other alike questionnaires (Balluerka, Gorostiaga, Alonso-Arbiol, & Haranburu, 2007; Huang, Church, & Katigbak,

1997). As a first application of the Spanish version, it could be concluded that results convey rigorous evidence and potential applicability. We consider some modifications should still be made in order to improve several aspects.

Limitations and potential research on S5

In first place, Agreeableness has often been found as a weak domain. Low internal consistency, poor factor loadings and even the only non-uniform DIF item detected are some examples, although modifications in one facet (A5 Modesty) were meant to improve its properties to some extent. We consider revealing the fact that both S5 and NEO PI-R display same disaggregated correlation patterns in the networks representation -with Trust and Compliance tightly related to Neuroticism. Further research on this point might be convenient, and so it will be carried out in further applications of this inventory. For instance, a new pool of items will be analyzed and selected, and SF49 will be studied for being flagged with DIF.

Some facets should be reviewed due to strong cross-loadings. Self-Consciousness (N4), Excitement seeking (E5), Positive emotions (E6) and Openness to actions (O4) hold the highest associations with unintended traits (E, O, N and E respectively). Although the theoretical explanation for them is indubitably reasonable, the presence of cross-loadings reduces the generalizability of the model. Hence, further efforts should aim to fit the target structure.

The network approach has revealed rather interesting information about the associations among personality facets. The concept of central-peripheral nodes unveils which facets bear a nuclear role, materializing the nature of some cross-loadings (such as the ones mentioned above). It should be noticed that, although the inter-domain associations are high for central nodes (hence, cross-loadings are), still the distribution notably embodies the Big Five standards. Unfortunately, there are no studies published yet to compare how invariant these central position-nodes are among instruments and populations.

All indicators have disclosed satisfactory levels of construct and structural equivalence, with the exception of CFI fit index in ESEM. As discussed above, this finding alone does not affect the overall interpretation, however it is worth some though. Due to the differences as a function of the sample size in S5, we consider a thorough simulation study ought to be conducted regarding the behaviour of CFI in ESEM settings. Moreover, Morin et al. (2013) already point out that these kind of effects could be encountered as a consequence of the ESEM estimation idiosyncrasy.

Further potential studies focused on S5 undoubtedly go through increasing the evidence for comparability of the different language versions. In first place, it is a priority to provide with proof of scalar equivalence of S5 in order to make possible comparisons across the different versions. This could be conveniently examined by applying multigroup ESEM. Additionally, many scholars call attention to the differential response styles across

cultures (Church, 2010; T. Johnson, Kulesa, Cho, Shavitt, et al., 2005; T. P. Johnson, Shavitt, & Holbrook, 2011). It is well known that validity of personality measurement can be threatened by social desirability, extreme response style and acquiescence tendencies. Studies on this topic could be conducted upon the S5 versions by a bi-factor ESEM model, where response style is modeled as an orthogonal factor to whom items from different scales relate (Cheung & Rensvold, 2000; Wiesner & Schanding, 2013). Although these studies could have been already conducted in the present paper, we remind the reader that the CFI index has displayed inconclusive information about the model, thus it cannot be used for the iterative steps involved in invariance studies and model selection. In this sense, a deeper knowledge about CFI performance conditions further decisions to be made about S5 structure.

We acknowledge that the modifications introduced in the Spanish S5 upon the Modesty (A5) facet make a difference with the rest of the versions. Although equivalence is therefore reduced, it should be understood that Short Five is a recent measure susceptible to be reduced and improved. As the change was made in the Spanish version first, it will be applied and tested in subsequent administrations of the instrument.

Still in line with generalizability of S5 is the target population. Items in S5 have been argued to be cognitively complex, and for this reason university samples have been utilized to study its properties. A next step on this topic could be gathering data from a more varied target population, still will high educational background, in order to find out further about the functioning. A second point regarding the sample is the lack of normative scales for interpretation. During next administrations of S5 these could be elaborated, preferably with cross-cultural keys for comparisons by scalar invariance evaluation or linking and equating strategies.

Our goal with this paper was to report evidence about the Spanish adaptation of an inventory constructed from an alternative perspective. Although there are several aspects left to improve, we believe that certain reasons have been supplied to make Short Five a choice to consider in personality research.

Acknowledgements

This project was partially supported by La Caixa, through a grant awarded to Regina García Velázquez. I would like to thank Vicente Ponsoda Gil and Markku Verkasalo for their advice and supervision. I also thank Féliz Arbinaga Ibarzábal, Carmen García García, Leena Honkasalo, Kenn Konstabel, María José López López, Jan-Erik Lönnqvist, Julio Olea Díaz and Essi Pakarinen (in alphabetical order) for supporting and contributing in different manners to this project.

			N	E	O	A	C
Neuroticism	N1: Anxiety	SF01	.62				
		SF56	.55				
	N2: Angry-hostility	SF11	.59				
		SF46	.55				
	N3: Depression	SF21	.69				
		SF36	.57				
	N4: Self-consciousness	SF31	.36	-.50			
		SF26	.31				
	N5: Impulsiveness	SF41	.55				
		SF16	.31				
	N6: Vulnerability	SF51	.71				
		SF06	.57				
Extraversion	E1: Warmth	SF02		.71			
		SF57		.68			
	E2: Gregariousness	SF12		.62			
		SF47		.42			
	E3: Assertiveness	SF22		.35			
		SF37		.44			
	E4: Activity	SF32		.40			
		SF27		.26			
	E5: Excitement-seeking	SF42		.47	.43		
		SF17		.46			
	E6: Positive emotions	SF52	-.42	.47			
		SF07		.67			
Openness	O1: Fantasy	SF03			.63		
		SF58			.65		
	O2: Aesthetics	SF13			.74		
		SF48			.67		
	O3: Feelings	SF23			.36		
		SF38			.30		
	O4: Actions	SF33		.47	.29		
		SF28			.31		
	O5: Ideas	SF43			.57		
		SF18			.57		
	O6: Values	SF53	*-.16*		*.15*		
		SF08			.28		
Agreeableness	A1: Trust	SF04				.26	
		SF59				.39	
	A2: Straightforwardness	SF14		*.20*		*.15*	
		SF49				.40	
	A3: Altruism	SF24				.34	
		SF39				.54	
	A4: Compliance	SF34	-.28			**.19**	
		SF29	-.39			.34	
	A5: Modesty	SF44				.39	
		SF19				.56	
	A6: Tender-mindness	SF54				.28	
		SF09				.35	
Consciousness	C1: Competence	SF05					.35
		SF60					.37
	C2: Order	SF15					.51
		SF50					.64
	C3: Dutifulness	SF25					.44
		SF40					.49
	C4: Achievement striving	SF35					.67
		SF30					.52
	C5: Self-discipline	SF45					.70
		SF20					.59
	C6: Deliberation	SF55					.48
		SF10					.39
Cumulative proportion of variance explained			.09	.18	.24	.28	.35

Table 12

EFA Procrustes-rotated loadings for Short Five according to the Big Five model (N=239).

Note. Values lower than .25 are displayed in italics.

			N	E	O	A	C
Neuroticism	N1: Anxiety	SF01	.63				
		SF56	.48				
	N2: Angry-hostility	SF11	.60				
		SF46	.43				
	N3: Depression	SF21	.65				
		SF36	.54				
	N4: Self-consciousness	SF31	.37	-.55			
		SF26	.30	-.51			
	N5: Impulsiveness	SF41	.46				
		SF16	*.20*				
	N6: Vulnerability	SF51	.70				
		SF06	.56				
Extraversion	E1: Warmth	SF02		.74			
		SF57		.74			
	E2: Gregariousness	SF12		.66			
		SF47		.42			
	E3: Assertiveness	SF22		.33			
		SF37		.46			
	E4: Activity	SF32		.42			
		SF27		*.20*			
	E5: Excitement-seeking	SF42		.45	.43		
		SF17		.39			
	E6: Positive emotions	SF52	-.44	.54			
		SF07		.62			
Openness	O1: Fantasy	SF03			.50		
		SF58			.46		
	O2: Aesthetics	SF13			.58		
		SF48			.50		
	O3: Feelings	SF23			.30		
		SF38			*.15*		
	O4: Actions	SF33		.41	.41		
		SF28			.30		
	O5: Ideas	SF43			.70		
		SF18			.47		
	O6: Values	SF53			*.24*		
		SF08			*.24*		
Agreeableness	A1: Trust	SF04				*.11*	
		SF59				.30	
	A2: Straightforwardness	SF14				*.12*	
		SF49				.31	
	A3: Altruism	SF24				*.21*	
		SF39				.50	
	A4: Compliance	SF34				*.14*	
		SF29				.35	
	A5: Modesty	SF44				.40	
		SF19				.55	
	A6: Tender-mindness	SF54				.31	
		SF09				.34	
Consciousness	C1: Competence	SF05					.36
		SF60	-.41				.39
	C2: Order	SF15					.49
		SF50					.57
	C3: Dutifulness	SF25					.52
		SF40					.57
	C4: Achievement striving	SF35					.68
		SF30					.51
	C5: Self-discipline	SF45					.73
		SF20					.57
	C6: Deliberation	SF55					.37
		SF10					*.20*

Table 13

ESEM standardized targeted loadings for Short Five according to the Big Five model (N=239).

Note. Targeted loadings lower than .25 are displayed in italics.

References

Aluja, A., & García, L. F. (2004). Relationships between Big Five personality factors and values. *Social Behavior and Personality: an international journal*, *32*(7), 619–625.

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The nature and structure of correlations among Big Five ratings: The halo-alpha-beta model. *Journal of Personality and Social Psychology*, *97*(6), 1142.

Asendorpf, J. B. (2012). What do the items and their associations refer to in a Network Approach to Personality? *European Journal of Personality*, *26*(4), 432–433. Retrieved from `http://dx.doi.org/10.1002/per.1867`

Ashton, M. C., & Lee, K. (2012). On models of personality structure. *European Journal of Personality*, *26*(4), 433–434. Retrieved from `http://dx.doi.org/10.1002/per.1868`

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438.

Balluerka, N., Gorostiaga, A., Alonso-Arbiol, I., & Haranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica. *Psicothema*, *19*(1).

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346.

Bilsky, W., Janik, M., & Schwartz, S. H. (2011). The structural organization of human values -Evidence from three rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology*, *42*(5), 759–776.

Briley, D. A., & Tucker-Drob, E. M. (2012). Broad bandwidth or high fidelity? Evidence from the structure of genetic and environmental effects on the facets of the Five Factor Model. *Behavior genetics*, *42*(5), 743–763.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, *18*(1), 23–38.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212.

Church, A. T. (2010). Measurement issues in cross-cultural research. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement.* Thousand Oaks, California: Sage Publications.

Cieciuch, J., & Schwartz, S. H. (2012). The number of distinct basic values and their structure assessed by PVQ–40. *Journal of personality assessment*, *94*(3), 321–328.

Cohen, N., & Arieli, T. (2011). Field research in conflict environments: Methodological challenges and snowball sampling. *Journal of Peace Research*, *48*(4), 423–435.

Costa, P. T., & McCrae, R. R. (1992). *NEO Personality Inventory-Revised (NEO PI-R) and NEO-Five Factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of personality assessment*, *64*(1), 21–50.

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO PI-R). *The SAGE handbook of personality theory and assessment*, *2*, 179–198.

Costa, P. T., McCrae, R. R., & Dye, D. A. (1991). Facet scales for Agreeableness and Conscientiousness: a revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*(9), 887–898.

Cramer, A. O., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in the adequacy of the European Social Survey to measure values in 20 countries. *Public opinion quarterly*, *72*(3), 420–445.

Diedenhofen, B. (2013). cocron: Statistical comparisons of two or more alpha coefficients [Computer software manual]. Retrieved from `http://r.birkdiedenhofen.de/pckg/cocron/` ((Version 1.0-0))

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of personality and social psychology*, *73*(6), 1246.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. Retrieved from `http://www.jstatsoft.org/v48/i04/`

European Social Survey. (2013). *ESS-6 2012 documentation Report. Edition 1.3.* Bergen, Norway: Norwegian Social Science Data Services.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103.

Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods*

in psychology. New York: Cambridge University Press.

Franìc, S., Borsboom, D., Dolan, C. V., & Boomsma, D. I. (2013). The Big Five personality traits: Psychological entities or statistical constructs? *Behavior Genetics*, 1-14. doi: 10.1007/s10519-013-9625-7

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

George, L. G., Helson, R., & John, O. P. (2011). The "CEO" of women's work lives: How Big Five Conscientiousness, Extraversion, and Openness predict 50 years of work experiences in a changing sociocultural context. *Journal of personality and social psychology*, *101*(4), 812.

Hambleton, R., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, *11*(3), 147.

Hambleton, R., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. *Cross-cultural research methods in psychology*, 46–70.

Herrmann, A., & Pfister, H.-R. (2013). Simple measures and complex structures: Is it worth employing a more complex model of personality in Big Five inventories? *Journal of Research in Personality*, *47*(5), 599–608.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Mahwah, New Yersey: Lawrence Erlbaum Associates.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192–218.

International Test Commission. (2010). *International Test Commission Guidelines for translating and adapting tests.* Vienna, Austria. Retrieved from `http://www.intestcom.org/`

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type i error and power rates using an effect size measure with the logistic regression procedure for dif detection. *Applied Measurement in Education*, *14*(4), 329–349.

Johnson, T., Kulesa, P., Cho, Y. I., Shavitt, S., et al. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-cultural psychology*, *36*(2), 264–277.

Johnson, T. P., Shavitt, S., & Holbrook, A. L. (2011). Survey response style across cultures. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology.* New York: Cambridge University Press.

Kenny, D. A. (2014). *Measuring model fit.* Retrieved from `http://davidakenny.net/cm/fit.htm`

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*(1), 165.

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379.

Konstabel, K., Lönnqvist, J.-E., Walkowitz, G., Konstabel, K., & Verkasalo, M. (2012). The 'Short Five' (S5): Measuring personality traits using comprehensive single items. *European Journal of Personality*, *26*(1), 13–29.

Locke, E. A. (2006). The motivation to work: What we know. In M. L. Maehr & P. R. Pintrick (Eds.), *Advances in motivation and achievement* (Vol. 10). Greenwich: JAI Press Inc.

Lönnqvist, J.-E., Paunonen, S., Verkasalo, M., Leikas, S., Tuulio-Henriksson, A., & Lönnqvist, J. (2007). Personality characteristics of research volunteers. *European Journal of Personality*, *21*(8), 1017–1030.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 57–64.

Magis, D., Beland, S., & Raiche, G. (2013). difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics [Computer software manual]. (R package version 4.5.)

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302–321.

Magis, D., & Facon, B. (2013a). deltaplotr: Identification of dichotomous differential item functioning (DIF) using Angoff's Delta Plot method [Computer software manual]. (R package version 1.3.)

Magis, D., & Facon, B. (2013b). Item purification does not always improve DIF detection. a counterexample with Angoff's Delta Plot. *Educational and Psychological Measurement*, *73*(2), 293–311.

Marsella, A. J., Dubanoski, J., Hamada, W. C., & Morse, H. (2000). The measurement of personality across cultures historical, conceptual, and methodological issues and considerations. *American Behavioral Scientist*, *44*(1), 41–62.

Marsella, A. J., & Leong, F. T. (1995). Cross-cultural issues in personality and career assessment. *Journal of Career Assessment*, *3*(2), 202–218.

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and D'Apollonia (1991). *Journal of Educational Psychology*, *83*, 416–421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285–296.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Psychometrics. a*

festschrift to Roderick P. McDonald. Hillsdale, NJ: Erlbaum Associates.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J., & Nagengast, B. (2011). Methodological measurement fruitfulness of Exploratory Structural Equation Modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29*(4), 322–346.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through Exploratory Structural Equation Modeling. *Psychological Assessment*, *22*(3), 471.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right —Camouflaging misspecification with item parcels in CFA models. *Psychological methods*, *18*(3), 257.

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An integration of the best features of Exploratory and Confirmatory Factor Analysis. *Annual review of clinical psychology*, *3*(10), 1–26.

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental psychology*, *49*(6), 1194.

McCrae, R. R. (2002). NEO PI-R data from 36 cultures. In *The five-factor model of personality across cultures* (pp. 105–125). New York: Springer.

McCrae, R. R., & Allik, I. U. (2002). *The Five-Factor model of personality across cultures.* New York: Springer.

McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

McCrae, R. R., & Costa, P. T. (1996). Toward a new generation of personality theories: Theoretical context for the Five-Factor model. In J. S. Wiggins (Ed.), . New York: Guilford Press.

McCrae, R. R., & Costa, P. T. (2010). The Five-Factor Theory of Personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), . New York: Guilford Press.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50.

McCrae, R. R., & Sutin, A. R. (2009). Openness to Experience. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior.* Ney York: Guilford.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and*

Social Psychology, *70*(3), 552.

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course.* Charlotte, NC: Information Age Publishing.

Muthén, L., & Muthén, B. (2013). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Olver, J. M., & Mooradian, T. A. (2003). Personality traits and personal values: a conceptual and empirical integration. *Personality and individual differences*, *35*(1), 109–125.

Ostendorf, F., & Angleitner, A. (2004). NEO PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae. *Hogrefe-Verlag GmbH Co. KG, Göttingen*.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, *57*, 401–421.

Pando, A. C., Pamos, A., Cubero, N. S., & Costa, P. T. (1999). *Inventario de personalidad neo revisado (NEO PI-R), inventario NEO reducido de cinco factores (NEO-FFI): manual profesional.* Madrid, Spain: Tea Ediciones.

Parker, P. D. (2014). *ESEM Invariance Script Writter.* Retrieved from `https://github.com/pdparker/ESEM`

Parks, L., & Guay, R. P. (2009). Personality, values, and motivation. *Personality and Individual Differences*, *47*(7), 675–684.

Paunonen, S. V. (1997). On chance and factor congruence following orthogonal Procrustes rotation. *Educational and Psychological Measurement*, *57*(1), 33–59.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology*, *74*(2), 538.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of personality and social psychology*, *84*(2), 411.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, *81*(3), 524.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Revelle, W. (2014). psych: Procedures for Psychological, Psychometric, and Personality Research [Computer software manual]. Evanston, Illinois. Retrieved from `http://CRAN.R-project.org/package=psych` (R package version 1.4.2)

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and social psychology bulletin*, *27*(2), 151–161.

Roccas, S., Sagiv, L., Schwartz, S. H., & Knafo, A. (2002). The Big Five personality factors and personal values. *Personality and social psychology bulletin*, *28*(6), 789–

801.

Rogler, L. H. (1999). Methodological sources of cultural insensitivity in mental health research. *American Psychologist*, *54*(6), 424.

Rothmund, T., Baumert, A., & Schmitt, M. (2012). Can network models represent personality structure and processes better than trait models do? *European Journal of Personality*, *26*(4), 444–445.

Schimmack, U., & Gere, J. (2012). The utility of Network Analysis for personality psychology. *European Journal of Personality*, *26*(4), 446–447.

Schimmack, U., Schupp, J., & Wagner, G. G. (2008). The influence of environment and personality on the affective and cognitive component of subjective well-being. *Social Indicators Research*, *89*(1), 41–60.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of personality and social psychology*, *94*(1), 168.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*, *25*(1), 1–65.

Schwartz, S. H. (2006). Value orientations: Measurement, antecedents and consequences across nations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally - lessons from the European Social Survey.* London: Sage.

Schwartz, S. H. (2012). An overview of the Schwartz's theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 11.

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: cross-cultural and multimethod studies. *Journal of personality and social psychology*, *89*(6), 1010.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, New Yersey: Lawrence Erlbaum Associates.

Sireci, S. G. (2011). Survey response style across cultures. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology.* New York: Cambridge University Press.

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, *13*(3), 229–248.

Smith, G. T., & Zapolski, T. C. B. (2009). Construct validation of personality measures. In J. N. Butcher (Ed.), . Oxford: Oxford University Press.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of personality and social psychology*, *100*(2), 330.

Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of personality and social psychology*, *101*(4), 862.

Struch, N., Schwartz, S. H., & Van Der Kloot, W. A. (2002). Meanings of basic values for women and men: A cross-cultural analysis. *Personality and Social Psychology Bulletin*, *28*(1), 16–28.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370.

Teresi, J. A. (2001). Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. *Journal of Mental Health and Aging*.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa Jr, P. T. (2005). Hierarchical linear modeling analyses of the NEO PI-R scales in the baltimore longitudinal study of aging. *Psychology and aging*, *20*(3), 493.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions the impact of data collection mode, question format, and question context. *Public opinion quarterly*, *60*(2), 275–304.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis of cross-cultural research.* Newbury Park, CA: SAGE Publications.

Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29.

Van de Vijver, F. J., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), . Mahwah, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology.* New York: Cambridge University Press.

Verkasalo, M., Lönnqvist, J.-E., Lipsanen, J., & Helkama, K. (2009). European norms and equations for a two dimensional presentation of values as measured with Schwartz's 21-item portrait values questionnaire. *European Journal of Social Psychology*, *39*(5), 780–792.

Wierzbicka, A. (1994). Emotion, language, and cultural scripts. In S. E. Kitayama & H. R. M. Markus (Eds.), . Washington, DC: American Psychological Association.

Wiesner, M., & Schanding, G. T. (2013). Exploratory structural equation modeling, bifactor models, and standard confirmatory factor analysis models: Application to the BASC-2 Behavioral and Emotional Screening System Teacher Form. *Journal of school psychology*, *51*(6), 751–763.

Willse, J. T. (2014). CTT: Classical Test Theory Functions [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=CTT` (R package version 2.0)

Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., ... others (2006). Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *Journal of personality and social psychology*, *90*(6), 987.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68.