# Two Mini-IPIP versions, balance and Positive Wording: validity and cross-cultural invariance

Agustín Martínez-Molina [Corresp., 1] , Víctor B Arias [2]

[1] Departamento de Psicología y Sociología, Universidad de Zaragoza, Teruel, Spain

[2] Facultad de Psicología, Universidad de Salamanca, Spain

Corresponding Author: Agustín Martínez-Molina
Email address: agustin@unizar.es

**Background**. The Mini-IPIP scales (Donellan et al., 2006) is possibly one of the most commonly used short inventories for measuring the Big Five Factors of personality. In this study, we aimed to investigate the psychometric properties of two Mini-IPIP Spanish short forms, one balance and one Positive Wording (PW). **Method.** Two samples, one from native Spanish speakers and another from native English speakers, made up a total of 940 participants in this study. Both samples were screened with an instructional manipulation check. The short forms were translated and adapted based on international guidelines. Reliability (internal and composite) and validity analyses (construct ESEM, concurrent, predictive and cross-cultural invariance through multi-group factorial models) were performed. **Results.** For both the balanced scale and the PW one, modeling a method factor was not relevant. The reliability and validity indices of both forms, were according to theory and prior studies' findings: (a) personality factors were medium-high related with affective factors; (b) personality factors were less related with life satisfaction than affective factors; (c) life satisfaction was medium-high related with affective factors; (d) Neuroticism appeared mainly related with all criteria variables; and (e) an acceptable level of invariance was achieved with regard to the English version. **Discussion.** This study contributes to research on personality assessment by providing the first evidence regarding the psychometric properties of a PW short measure. These results suggest that PW short scales of personality used after data screening techniques may be appropriate for future studies (e.g., cross-cultural, content validity).

1 **Two Mini-IPIP versions, balance and Positive Wording:**
2 **validity and cross-cultural invariance**

4 Agustín Martínez-Molina[1], Víctor B. Arias[2]

5 [1] Departamento de Psicología y Sociología, Universidad de Zaragoza, Teruel, España

6 [2] Facultad de Psicología, Universidad de Salamanca, España

9 Corresponding Author:

10 Agustín Martínez-Molina[1]

11 Ciudad Escolar s/n, Teruel, 44003, España

12 Email address: agustin@unizar.es

28                                    **Abstract**

29    **Background**. The Mini-IPIP scales (Donellan et al., 2006) is possibly one of the most

30    commonly used short inventories for measuring the Big Five Factors of personality. In this study,

31    we aimed to investigate the psychometric properties of two Mini-IPIP Spanish short forms, one

32    balance and one Positive Wording (PW).

33    **Method.** Two samples, one from native Spanish speakers and another from native English

34    speakers, made up a total of 940 participants in this study. Both samples were screened with an

35    instructional manipulation check. The short forms were translated and adapted based on

36    international guidelines. Reliability (internal and composite) and validity analyses (construct

37    ESEM, concurrent, predictive and cross-cultural invariance through multi-group factorial

38    models) were performed.

39    **Results.** For both the balanced scale and the PW one, modeling a method factor was not

40    relevant. The reliability and validity indices of both forms, were according to theory and prior

41    studies' findings: (a) personality factors were medium-high related with affective factors; (b)

42    personality factors were less related with life satisfaction than affective factors; (c) life

43    satisfaction was medium-high related with affective factors; (d) Neuroticism appeared mainly

44    related with all criteria variables; and (e) an acceptable level of invariance was achieved with

45    regard to the English version.

46    **Discussion.** This study contributes to research on personality assessment by providing the first

47    evidence regarding the psychometric properties of a PW short measure. These results suggest

48    that PW short scales of personality used after data screening techniques may be appropriate for

49    future studies (e.g., cross-cultural, content validity).

50 **Introduction**

51        Long testing sessions can lead to exhaustion, irritation, and demotivation in the

52     respondent, particularly in low-stakes contexts (Wise & DeMars, 2013). Such conditions

53     increase the probability of diminished attention to the task and produce responses that are not

54     based on the item content (e.g., careless responding; Meade & Craig, 2012), with a consequent

55     deterioration of the validity and a need to implement control measures that further prolong the

56     testing session (DeSimone, Harms, & DeSimone, 2015; Maniaci & Rogge, 2014; Oppenheimer,

57     Meyvis, & Davidenko, 2009; Weijters, Baumgartner, & Schillewaert, 2013).

58        As a consequence, possessing short-form measures that can deliver an acceptable proxy

59     for a person's position on broad constructs is clearly of interest in terms of both research and

60     applied assessment (Credé, Harms, Niehorster, & Gaye-Valentine, 2011). This is particularly

61     true in the case of multidimensional and multifaceted constructs such as human personality, the

62     assessment of which has traditionally required extensive scales. In recent years, substantial

63     efforts have been made to develop more efficient measures of basic personality traits. A

64     representative example can be found in the diverse levels of reduction of scales evaluating the

65     Big Five based on the International Personality Item Pool (IPIP; Goldberg, 1999): 300 items

66     (IPIP-NEO; Goldberg, 1999), 120 items (IPIP NEO 120; Johnson, 2014), 50 items (IPIP Big

67     Five Factor markers; Goldberg, 1992), 32 items (IPIP-IPC; Markey & Markey, 2009), and 20

68     items (Mini-IPIP; Donellan, Oswald, Baird, & Lucas, 2006).

69        The Mini-IPIP is possibly one of the most commonly used short scales based on the Big

70     Five model. Donellan et al., (2006) developed the Mini-IPIP with five main objectives: (a) to

71     achieve adequate balance between brevity and sound psychometric properties, (b) to ensure a

72     sufficient number of items per factor in order to permit their non-problematic application in

73  structural equation modelling, (c) to maximize the empirical independence between the five

74  basic factors (i.e., correlations between factors close to zero), (d) to maximize the discriminant

75  validity of the items (i.e., that they would be nearly pure indicators of their theoretical factor),

76  and (e) to minimize losses in reliability and validity with regard to longer versions of the scale.

77  The results of the five studies described by Donellan et al., (2006) suggest that the objectives

78  above were met adequately. The Mini-IPIP showed to be a reasonable proxy for evaluating the

79  Big Five factors and offering an adequate trade-off between conciseness, reliability, and both

80  construct and criterion validity. However, the authors did not obtain adequate fit in the

81  confirmatory factor models, a problem that is common in complex multidimensional models of

82  personality (Hopwood & Donellan, 2010), possibly due to the excessive restrictiveness of the

83  independent cluster model of confirmatory factor analysis (Marsh, Lüdtke, Muthén, Asparouhov,

84  Morin, Trautwein, & Nagengast, 2010).

85          Although the Mini-IPIP has been widely used in research since 2006, few studies have

86  been explicitly dedicated to evaluating its psychometric properties (Oliveira, 2017). This is

87  important considering the essential role of an independent replication of results for guaranteeing

88  the validity of short-forms (Smith, McCarthy, & Anderson, 2000). Cooper, Smillie, & Corr

89  (2010) found poor fit with a five correlated factors confirmatory model, as well as a model with

90  two superordinate factors. However, they obtained acceptable internal consistency coefficients

91  (Cronbach's alpha), and exploratory factor analysis (EFA) supported the empirical independence

92  of the dimensions (i.e., low correlations) and the discriminant validity of the items (i.e., the

93  primary loadings were substantially higher than the cross-loadings). Baldasaro, Shanahan, &

94  Bauer (2012) found similar results with a sub-optimal fit in the CFA models, with several high

95  modification indices suggesting possible violations of local independence between pairs of items.

96   They also tested measurement invariance (sex and ethnicity) without exceeding metric

97   invariance in any of the sub-scales. Cross-cultural differences in self-reported items were usually

98   analyzed through expanding an invariance test to recognize possible cultural variance in

99   language, that is, analyze if the answers to the same item although in different languages is

100  understood in an equal or similar manner. Laverdière, Morin, & St-Hilaire (2013) found poor fit

101  with the confirmatory model, requiring the freeing of correlations between three pairs of

102  residuals because of semantic similarity of items belonging to the same facet. They also found

103  reasonable evidence of measurement invariance to a level of variances/covariances

104  (undergraduates vs. employees, age, and gender).

105       Personality scales have been balanced traditionally as an intent to cancel out the effects of

106  the agreeing tendency (Couch & Keniston, 1960). Half of the items are worded positively and

107  the other half are worded negatively. Balance keying as a method control for acquiescence can

108  improve the psychometric properties of the personality measures (Kostabel et al., 2017;

109  Rammstedt et al., 2013). However, method effects associated with worded items may emerge

110  producing stable response-style factors (Kam, & Chan, 2018; Marsh, Scalas, & Nagengast,

111  2010), i.e., systematic artifacts with undesirable consequences for measurement produced by the

112  inclusion of a mix of positively and negatively keyed items.

113       Wording effects occur when we assume that some direct and inverse items are equivalent

114  in their ability to reflect a construct, and then response patterns appear in the positive or negative

115  wording of the items instead of in their content (Podsakoff and MacKenzie, 2003). These

116  response patterns are not random and can lead both to the artificial inflation of the correlations

117  between latent variables, and to the deflation of the correlation between the items of the same

118  factor (Huang, Liu & Bowling, 2015). The result can lead to a multidimensionality scenario that

119 frequently requires extra method factors (Arias & Arias, 2017; Eid, 2000), or the implementation

120 of various cleaning data procedures (Meade & Craig, 2012).

121     On personality short scales, the wording effect is even less known. Wording effects have

122 been considered as one of the possible variables that causes detriments of the psychometric

123 scales' properties (e.g., overestimated test reliability, misfit validity; Eys et al., 2007; Lai, 1994;

124 Wang et al., 2015). That being said, it should be noted that over half of the original Mini-IPIP

125 items use negative wording (i.e., due to the use of words with inverse semantic polarity to the

126 measured trait or by the use of the adverb "not"). Based on previous research we hypothesized

127 that the use of negative items is not relevant in these brief scales, that is, without validity and

128 reliability consequences.

129     Although Spanish adaptations of the IPIP Big Five markers do exist (Cupani, 2009;

130 Cupani & Lorenzo-Seva, 2016; Goldberg, 1992) and could be used to obtain translations of the

131 Mini-IPIP items, to our knowledge, no study to date has assessed the psychometric properties of

132 the Mini-IPIP in Spanish nor investigated its measurement equivalency with regard to the

133 original English version. This point is particularly relevant, given that, unless it is demonstrated

134 that the scale is measuring the same construct in the same manner, it is not possible to guarantee

135 that the translated version is truly equivalent to the original measure as designed by its creators

136 (Wu, Li, & Zumbo, 2007).

137     The purpose of this study was to investigate the psychometric properties of two translated

138 and adapted Mini-IPIP Spanish short forms, one balance and one Positive Wording (PW). To this

139 end, we: (a) adapted the Mini-IPIP scales from the English to Spanish following internationally

140 recognized quality standards (Muñiz, Elosua, & Hambleton, 2013); (b) proposed a parallel

141 positive-wording version of these scales; (c) verified, in both versions, that the dimensionality,

142 internal structure, and other evidence of validity and reliability correspond to expectations based

143 on theory and prior studies (i.e., concurrent and predictive indexes with emotions and life

144 satisfaction); and (d) investigated measurement invariance with regard to the English version

145 through multi-group factorial models.

**Method**

147 **Participants**

148     Two samples were used. Sample 1 (native Spanish speakers) was collected to assess the

149 validity and reliability of the proposed scales. Sample 2 (native English speakers) was obtained

150 to analyze the cross-cultural invariance.

151     Sample 1 comprised 560 students enrolled in five different faculties at two Chilean

152 universities. The evaluation was completely anonymous, and consent was obtained from all

153 participants for their responses to be used as part of the research. Sample 1 scales were

154 computer-lab administered in groups of 15 participants where one of the authors of this work

155 was always present.

156     Sample 2 consisted of 380 native English speakers of U.S. nationality with diverse levels

157 of educational attainment (no formal qualification: 5%; secondary school: 19%; college: 28%;

158 undergraduate degree: 37.5%; graduate degree: 8.5%; doctoral degree: 2%). Sample 2 data were

159 gathered through Prolific Academic, a service supported by Oxford University that specializes in

160 online data gathering using panels of participants defined in advance by the researcher.

161 **Procedure**

162     Before starting the study, ethical approval was obtained from the Bioethics Committee of

163 Universidad de Talca (projects nº 1151271, nº 11140524). In both applications, (a) the

164 participants received monetary compensation equivalent to 2 USD, (b) the items were presented

165 with the same visual arrangement and order, (c) the acceptance rate was 100% (no missing data

166 was observed), and (d) to identify the participants that responded inattentively, both samples

167 underwent an instructional manipulation check (IMC; Oppenheimer et al., 2009; Weijters et al.,

168 2013). The IMC consisted of an item, which displayed identical rating scale as the rest of the

169 items but contained a specific instruction ("For this statement, please do not check a response

170 option"). The participants who, despite the special instruction, responded to the item were

171 considered careless responders and thus were eliminated from the analysis.

172      A total of 45 students did not meet the criterion of attentional control in this study (n = 32

173 in Sample 1, n = 13 in Sample 2). The final sampling sizes were, Sample 1, n = 518 (age range =

174 18-34, M = 21.6, SD = 2.3; 70.7% women), and Sample 2, n = 367 (age range = 18-72, M =

175 34.6, SD = 12.7; 48.5% women).

176      For efficiency reasons and validity purposes a fraction of Sample 1 (50%, n = 280) was

177 randomly selected to complete three more short-forms scales. One out two assessment groups of

178 participants were asked to complete two extra scales (Sample 1 after attentional control, n =

179 278).

180 **Measures**

181      *The Mini-IPIP Scales* (Donellan et al., 2006):  This is the abbreviated version of the IPIP

182 Big Five factor markers (Goldberg, 1992), which consists of 20 items (4 per dimension):

183 Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness

184 (O). Eleven items are reverse-scored. Each item is answered by the participant on a five-point

185 accuracy scale (1 = Not at all, 5 = Completely) in accordance with the level to which each

186 statement is applicable to their habitual behavior. As reported in four studies by Donellan et al.,

187 (2006), the average reliabilities indices (Cronbach's α) of this version are .81 (E), .73 (A), C

188 (.70), N (.74) and O (.69). Correlation and fit indices also supported the construct, convergent

189 and discriminant validity according to other Big Five measures.

190    *The Mini-IPIP Scales Spanish version*: This form was created on the base of the

191 International Test Commission guidelines for translating and adapting tests (ITC, 2015; Muñiz,

192 Elosúa & Hambleton, 2013). Three native-Spanish speakers with advanced English proficiency

193 and academic expertise in the field (i.e., proficient level C in English according to the Common

194 European Framework of Reference for Languages and doctoral studies in measurement or

195 personality assessment) considered linguistic and cultural factors in the translation and

196 adaptation of the original Mini-IPIP items (Donellan et al., 2006). An iterative debug translating

197 procedure was performed. The experts began with independent translations that they iteratively

198 shared until they rationally agreed upon a final debugged version. The translations were highly

199 concordant in their content in the first iteration. Items that maintained the original meaning and

200 contained only standard Spanish expressions were selected. The experts did not reach sufficient

201 agreement on two items in the second iteration (Item 18 "make a mess of things" and item 19

202 "seldom feel blue"). A pilot test on 73 volunteers (convenience sampling) with socio-

203 demographic characteristics of age and sex similar to those of the final participants (sample 1)

204 was conducted to provide empirical evidence to resolve the discrepancies in the mentioned items

205 (18 and 19). For these items, we selected the wording (from two alternatives for each item) that

206 showed the higher factorial loading in their main factor and lower cross-loadings, according to

207 an EFA based on a polychoric correlations matrix with unweighted least squares estimator and

208 promax rotation. Appendix A and B shows the final versions of the Mini-IPIP with the basic

209 descriptive statistics obtained from the two main samples of this study and instructions that the

210 participants received.

211  *The Mini-IPIP Scales Positive Wording (PW) Spanish version*: Following the same

212 adaptation method described above, a positive wording set of items was composed (see

213 Appendix B). In order to have a complete positive short-form of the scales we made a parallel

214 version from the original reverse 11 items (the rest of items were already PW in the original

215 version). Of these items, those of N were elaborated with a positive semantic polarity towards

216 "Emotional Stability" instead of "Neuroticism".

217  *The PANAS-C10 Spanish version*: In accordance with the framework of this study, a

218 short-form was also used to measure positive affect (PA) and negative affect (NA) factors for

219 convergent validity purposes. The original ten-item version (Damasio et al., 2013) is a brief form

220 of the PANAS (Carvalho et al., 2013; Watson et al., 1988). Participants were asked to rate two

221 self-report semantic mood scales (five-point ordinal agreement alternatives; 1 = Not at all, 5 =

222 Completely). The PANAS-C10 English version reported that the reliability indices (Cronbach's

223 α) were .81 (PA) and .82 (NA). Fit indices also supported the unidimensionality (62% explained

224 variance, KMO = .82; Barlett (45) = 494.4 p ≤ .001; CFI =.97; CFI =.97; RMSR = .06; RMSEA

225 = .04; only one averted dimension with parallel analysis). The PANAS-C10 Spanish version was

226 translated and adapted following the same procedure described for the Mini-IPIP Spanish

227 version. Given the semantic simplicity of this measure (only ten common words, see Appendix

228 C), it was not surprising that there where almost no disagreement between experts in the item

229 creation process. One item (number 4 "Fun") did not retain its original meaning in its translated

230 version. In Spanish, it is not common to refer "fun" as a personal state of an emotion or feelings.

231 This word is commonly used to refer circumstances or people who cause you fun. The synonym

232 "Entretenido" was agreed as an alternative of "Fun", which is also naturally understood in both

233 versions of the verb "to be" in Spanish (i.e., "ser" and "estar"). PANAS was related to different

234    personality scales for convergent validity support (Watson, & Clark, 1999). Two Principal

235    Component Analysis grouped these measures accordingly (NA was substantially related with N,

236    and PA to E). Similar results were showed by Bruck & Allen (2003); only N (.69), E (−.16), and

237    A (−.27) reported significant relations with NA.

238         *The SWLS Spanish version* (Moyano-Díaz et al., 2014): The Satisfaction with Life Scale

239    was translated and adapted from Diener et al., (1985). Five items with a response scale of 5

240    levels of satisfaction (1 = Not at all, 5 = Completely). The reliability index (Cronbach's α) of this

241    version ranged for .82 to .87. Fit indices also supported the unidimensionality (64% explained

242    variance, KMO = .84; Barlett (10) = 576.0 p ≤ .001; GFI = 1.0. and RMSR = .035; only one

243    averted dimension with parallel analysis). This short-form was used in this study for predictive

244    validity purposes. Hayes & Joseph (2003), provided evidence that C, N and, E, were related to

245    this subjective measure of well-being (only N and C predicted significantly). Other authors

246    (Chico, 2006; Joshanloo & Afshari, 2011), found that N and E were related strongly to the

247    SWLS (N accounted most of the variance in both studies).

248    **Data Analysis**

249         **Measurement Models.**

250         First, the data from Sample 1 were fitted to the Big Five model (Costa & McCrae, 1992)

251    through exploratory structural equation modelling (ESEM; Asparouhov & Muthén, 2009) with

252    oblique target rotation. The ESEM is a general technique for factorial analysis that permits the

253    estimation of all possible cross-loadings. ESEM offers more precise estimators of the factor

254    loadings and correlations between factors than CFA. ESEM has been shown to be more efficient

255    than CFA in the estimation of complex models with interstitial relationships between items

256    belonging to different dimensions, and furthermore offers the main advantages of confirmatory

257     analysis while maintaining the flexibility of unrestricted factor analysis (Garrido et al., 2018;

258     González-Arias et al., 2018; Morin, Arens, & Marsh, 2016; Marsh, Morin, Parker, & Kaur,

259     2014). Target rotation was used in all the models, permitting the specification of a matrix of

260     primary loadings, enabling the use of ESEM in a confirmatory manner (Asparouhov & Muthen,

261     2009).

262        Four models were estimated for each version of the scales (i.e., Mini-IPIP Spanish and

263     Mini-IPIP Spanish PW). Model 1 (M1) and M1-PW replicated the basic theoretical structure of

264     the scale through the specification of five correlated factors.

265        M2 and M2-PW specified a series of correlated residuals or Correlated Uniqueness (CU),

266     in order to prevent the estimation of the substantive loadings from being biased by the presence

267     of spurious variance due to semantic similarity between certain pairs of items (Cole, Ciesa &

268     Steiger, 2007). CU were freed between residuals of items that simultaneously (a) pertained to the

269     same facet, (b) demonstrated strong similarity of wording, and (c) demonstrated extreme

270     modification indices in M1. In M2 two pairs of items met the three conditions ("Am not

271     interested in abstract ideas" / "Have difficulty understanding abstract ideas," and "Sympathize

272     with others' feelings" / "Feel others' emotions"). M2-PW contained the same CU specifications

273     as M2.

274        In M3 and M3-PW, a random intercept confirmatory factor was also tested (as method

275     factor). Random intercept confirmatory factor analysis (RI-IFA; Billiet & McLendon, 2000;

276     Maydeu-Olivares & Coffman, 2006; Aichholzer, 2014) consists of the inclusion of a factor

277     common to all items. The RI factor is orthogonal to the substantive dimensions, and its loadings

278     are fixed to equality (as a consequence, the RI factor occupies a single degree of freedom,

279     corresponding to its variance). The RI factor imposes an artificial relationship between the items

280   with different wording polarity, capturing and isolating the variance associated with response

281   artefacts such as acquiescence (Maydeu-Olivares & Coffman, 2006).

282       M4 and M4-PW added CU and RI specifications of previous models. Finally, Model 5

283   (M5), based on the English-speaking sample, was structurally identical to M4.

284       **Cross-Cultural Invariance Analysis.**

285       Second, for the selected model in the previous phase, a series of multi-group ESEM

286   models nested in successive levels of increasing restriction were estimated (Meredith, 1993;

287   Millsap & Yu-Tein, 2004). The tested models were, in the following order, Configural, Strong,

288   and Strict. Each of these models tests a series of hypotheses regarding the equivalence of the

289   parameters between the groups compared. Configural test assumes that for both groups (Chilean

290   and US samples) has the same number of dimensions and the same configuration of factorial

291   loadings, that is, in both groups, a qualitatively similar construct is being measured.

292       Strong invariance requires that the factorial loadings as well as the intercepts of the items

293   (the thresholds, in this case) be equal between groups. Achieving strong invariance implies that

294   group differences in the latent scores are due to differences in the trait (not dependent on the

295   scale), that therefore, the scores are not biased against any of the groups. As a result, if strong

296   invariance is achieved, the changes observed in the latent measures of the factor can be

297   interpreted as a function of the change in the latent construct (Marsh, Lüdtke, Muthén,

298   Asparouhov, Morin, Trautwein, & Nagengast, 2010).

299       Strict invariance entails that the residual variances of the items must be equivalent

300   between groups. Failure to achieve the strict level can be interpreted as differences in the

301   reliability of the raw scores. As a consequence, achieving this level is not as relevant as fulfilling

302   the rest of the invariance conditions, at least when one is working with latent variables where the

303   measurement error is controlled (Byrne, 2013). However, the above is only true when the

304   residual variances meet the requirement of conditional independence, that is, they must be

305   authentically random (Wu et al., 2007; Deshon, 2004; Vanderberg & Lance, 2000). As a result,

306   given the logical possibility that part of the error would be systematic in nature (which is

307   plausible given that the Mini-IPIP, like other personality scales, contains sub-groups of items

308   pertaining to the same facet), the Strict test acquires importance (a) to guarantee that the results

309   of the scalar invariance are not biased and (b) to guarantee the unbiased comparability of the raw

310   scores and factor scores.

311          The analyses described above yielded the estimation of six multi-group models (M6 to

312   M8-PW). The PW-Models were tested using the positive wording items in the Spanish sample

313   *vs.* the balance English version in the English sample.

314          All of the models used the Weighted Least Squares and adjusted Mean and Variance

315   (WLSMV) estimator, given the ordered-categorical nature of the raw data (Beauducel &

316   Herzberg, 2006). Goodness of fit was evaluated using the comparative fit index (CFI), the

317   Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA). CFI and

318   TLI values $\geq .95$ are considered adequate, as are RMSEA values $< .05$ (Schreiber, 2017). To

319   compare the fit of nested models in the multi-group analysis, the recommendations of Chen

320   (2007) and Cheung & Rensvold (2002) were followed ($\Delta\text{CFI} \geq -.01$ supplemented by a

321   $\Delta\text{RMSEA} \geq .015$ suggest noninvariance). The analyses were performed using MPlus v. 7.4

322   (Muthén & Muthén, 2014) and Winsteps 4.1.0 (Linacre, 2018).

323                                          **Results**

324   **Estimated Models**

325   The fit indices for all the models are shown in Table 1. The ESEM analysis produced a

326 sub-optimal fit in the case of M1 and M1-PW (e.g., RMSEA > .08, TLI < .90). As expected, the

327 modification indices suggested freeing the correlations between the residuals of the commented

328 items. M2 and M2-PW (CU allowed) showed an acceptable fit ($\chi 2$ (98) = 240 and 171, RMSEA

329 = .05; CFI > .97; TLI > .94).

330   M3 and M3-PW (response artefacts controlled) presented no fit improvement over M2

331 and M2-PW ($\chi 2$ (99) = 337 and 178, RMSEA < .07; CFI > .95; TLI > .91), suggesting that the

332 wording effect, while present to a certain extent, were concentrated in the commented items and

333 were not a relevant problem in this sample.

334   M4, M4-PW and M5 (U.S. sample) that added CU and RI specifications achieved

335 adequate levels of fit ($\chi 2$ (97) = 235, 172 and 241, RMSEA < .06; CFI > .97; TLI > .95). The

336 estimated factor loading for the RI method factor in M4 and M4-PW were .141 and .085. One of

337 the advantages of modeling the method factor with RI is that the loss of degrees of freedom is

338 minimal (i.e., 1 degree). When the method factor is not relevant, as it was in this case, the fit

339 indices of these models were almost identical.

340         (please insert table 1 here)

341   Among the models tested, the most parsimonious and with the best fit was the M2 and

342 M2-PW. The parameters obtained from this model are shown in Table 2. All the primary

343 loadings were significant (M = .69; DT = 0.13) and their standard errors reasonably reduced (M

344 = .03, SD = 0.01). All the cross-loadings were small (absolute mean = .02; SD = 0.005), and

345 mainly non-significant (p > .01). Finally, the reliability of each substantive factor was calculated

346 using the coefficient of composite reliability (Raykov, 1997). The five factors acquired good

347 composite reliability (range = .90 to .94). As expected based on the theoretical model (Costa and

348   McCrae, 1992) and previous research (Donellan et al., 2006; Baldasaro et al., 2013), the

349   correlations between factors were low (ranged from –.01 to .27).

350                          (please insert table 2 here)

**Cross-Cultural Invariance Analysis**

352         **Test of configural invariance.**

353         The fit indices for the invariance models are shown in Table 1. Configural structure M6

354   and M6-PW, based on M2 and M2-PW (ESEM + CU), were the base models (i.e., with which

355   the rest of the invariance models were compared). With the exception of the RMSEA that

356   slightly exceeds the recommended value, M6 and M6-PW produced an acceptable fit (RMSEA <

357   .06; CFI ≥ .95; TLI ≥ .95).

358         **Test of strong invariance.**

359         M7 strong invariance test was executed imposing equality restrictions on the 80

360   thresholds corresponding to the five categories of each item (i.e., scalar invariance). Between

361   samples (Spanish-English) the balance version of the scales showed that the change in the

362   RMSEA satisfied the criterion of invariance while the CFI did not (M7: $\Delta\chi2 = 516$; $\Delta df = 130$;

363   $\Delta$RMSEA = .011; $\Delta$CFI = –.036; $\Delta$TLI = –.022). The invariance analysis showed local misfit

364   (thresholds with standardized expected parameter changes > .20 and modification indices > 10;

365   Garrido et al., 2018; Saris et al., 2009; Whittaker, 2012). A partial invariance model (M7p) was

366   executed following a stepwise implementation. One at a time, the threshold with the highest

367   outlier value was constricted. Even though the CFI-threshold criterion was not reached in M7p,

368   the $\Delta$RMSEA did (M7p: $\Delta\chi2 = 318$; $\Delta df = 114$; $\Delta$RMSEA = .003; $\Delta$CFI = –.019; $\Delta$TLI =   –

369   .005).

370      As we pointed out before, the PW-Models were tested using the positive wording items in

371   the Spanish sample *vs.* the balance English version in the English sample. M7-PW strong

372   invariance test was modeled by imposing the same initial constraints on the thresholds as M7. In

373   this invariance test the change of the two fit indices did not meet the recommended criteria (M7-

374   PW: $\Delta\chi 2 = 732$; $\Delta df = 130$; $\Delta RMSEA = .030$; $\Delta CFI = -.059$; $\Delta TLI = -.053$). The M7-PWp was

375   also tested following the same constriction process as M7p. The change in RMSEA meet the

376   invariance threshold and the change in CFI was very close to achieving it (M7-PWp: $\Delta\chi 2 = 212$;

377   $\Delta df = 90$; $\Delta RMSEA = .002$; $\Delta CFI = -.012$; $\Delta TLI = -.004$).

378      **Test of strict invariance.**

379      From the strong invariance model, the residual variances were fixed to equality. As

380   showed in Table 1, the RMSEA satisfied the criterion of invariance while the CFI did not (M8:

381   $\Delta\chi 2 = 601$; $\Delta df = 150$; $\Delta RMSEA = .012$; $\Delta CFI = -.042$; $\Delta TLI = -.024$). Like M7-PW, M8-PW

382   did not obtain an adequate fit either (M8-PW: $\Delta\chi 2 = 855$; $\Delta df = 150$; $\Delta RMSEA = .033$; $\Delta CFI = -$

383   $.068$; $\Delta TLI = -.058$). A partial invariance test where executed after observing local misfit. As

384   happened in the strong invariance tests, the RMSEA met the criteria while the CFI did not (M8p:

385   $\Delta\chi 2 = 359$; $\Delta df = 130$; $\Delta RMSEA = .003$; $\Delta CFI = -.022$; $\Delta TLI = -.005$; M8-PWp: $\Delta\chi 2 = 348$;

386   $\Delta df = 90$; $\Delta RMSEA = .012$; $\Delta CFI = -.024$; $\Delta TLI = -.018$).

387      **Invariance tests with method factor.**

388      The invariance of the M4 and M4-PW models was also tested, that is, models with

389   method factor. For the same reasons we pointed out above about the loss of degrees of freedom,

390   the fit indices of these models were almost identical to those obtained in the invariance of M2

391   and M2-PW.

392   **Convergent and predictive validity of the scales**

393    Table 3 shows the bivariate correlations between the study scales. SWLS, PA, NA, E and

394    N described among them most of the substantive correlations of this study. Firstly PA, NA and

395    SWLS correlated with a similar magnitude (.5. –.40 and .41). Second, we found the same pattern

396    of Pearson correlations between the adapted short-form scales compared to previous studies with

397    the longer original ones: (a) PA with E (.47) and N (–.40); (b) NA with N (.48) and E (–.19); (c)

398    SWLS with N (–.25), E (.23), and C (.22). C also described a significant but smaller correlation

399    than N and E with NA (–.16). The lowest significant correlation was between PA and O (.13).

400    We found the same pattern of correlations described above between the Mini-IPIP

401    Spanish PW and the rest of scales: (a) PA with E (.47) and N (–.37); (b) NA with N (.45) and E

402    (–.17); (c) SWLS with N (–.23), E (.25), and C (.24). C also described a significant but smaller

403    correlation than N and E with NA (–.13). And again, the lowest significant correlation was

404    between PA and O (.13). There were no statistical differences between correlations using Fisher's

405    z' transformation at p < .05 (SD$\Delta r_i$ = 0.03).

406    The summary of the factor relations of this study was fully illustrated in the SEM model

407    of Figure 1 (WLSMV and mentioned CU; $\chi 2$ (529) = 947.65, RMSEA = .053; CFI = .939; TLI =

408    .931). The variances of PA and NA were explained in a medium to high extent by the personality

409    factors ($R^2_{PA}$ = .54, $R^2_{NA}$ = .62). And the Life Satisfaction variance was explained medially high

410    mainly by affective factors ($R^2_{LS}$ = .47). Again, the same pattern of substantive correlations, fit

411    information and magnitude of variance explained on the criteria variables was found with the

412    Mini-IPIP PW ($\chi 2$ (529) = 994, RMSEA = .056; CFI = .936; TLI = .928, $R^2_{PA}$ = .53, $R^2_{NA}$ = .56,

413    $R^2_{LS}$ = .47).

414    Two more reliability indices are shown in the same table; all were good (from .78 to .86)

415    except N (as with composite reliability) which showed a tight but acceptable magnitude of

416    reliability for a short-form.

417                                    (please insert Figure 1 here)

418                                              **Discussion**

419    This study aimed to translate and adapt two versions (balance and positive wording) of

420    the Mini-IPIP scales to Spanish (Donellan et al., 2006), and examine aspects related to its

421    validity (internal structure, presence of response artefacts, and cross-cultural invariance) by

422    estimating a series of exploratory structural equation models (ESEM) on samples from Chile and

423    the United States.

424    Our results suggest the following: First, both Mini-IPIP versions (balance and positive

425    wording) showed congruent properties with theoretical expectations in the form of five clearly

426    identifiable and separable factors. The items demonstrated high convergent validity, as their

427    primary loadings were generally high (> .70) and the cross-loadings were generally low or not

428    significantly different from zero (high discriminant validity). The reliability of the factors was

429    adequate, particularly considering the low number of items per factor.

430    Second, after using data screening techniques, modeling a method factor was not

431    relevant. These results are especially interesting because support the use of positive worded

432    items for personality assessment purposes (simpler for the respondent) supplemented by data

433    screening techniques to control the undesired effect of other sources of variability.

434    Third, both versions (balance and positive wording) were close to achieve strong and

435    strict level of invariance with regard to the original version (in English). As in the measurement

436    invariance study performed by Zemojtel-Piotrowska et al., (2017), although the CFI difference

437 criterion was not reached, the RMSEA's did. It is necessary to point out that for measurement

438 invariance analysis Chen (2007) recommend using the thresholds criteria in a supplemented way

439 (not as added conditions). So, what is the reason for the RMSEA-CFI discrepancy in this

440 invariance tests? It is appropriate to consider that (a) the cut-off values for noninvariance are no

441 "golden rules" and may not be in all generalizable conditions. Most of the fit indices may be

442 sensitive to some conditions such as sample size, factor structure (e.g., one factor, bifactor),

443 factor relations (e.g., orthogonal, oblique), factorial approach (e.g., confirmatory, exploratory),

444 data nature (e.g., continuous, ordinal), estimation methods (e.g., ML, WLSMV) or correlated

445 errors (Fan et al., 2007; Greiff, & Scherer, 2018; Greiff & Heene, 2017; Heene et al., 2012; Sass,

446 Schmitt & Marsh , 2014); (b) RMSEA and CFI assess the adjustment from different perspectives

447 (Lai & Green, 2016; McNeish, An, & Hancock, 2018).

448       Judging that the invariant tests of this study were conducted from an ESEM approach,

449 with little related factors, ordinal data, correlated errors and WLSMV estimation, we consider

450 acceptable the strong partial and strict partial invariance level for both Mini-IPIP versions. This

451 indicates that language and wording had a little effect on the way in which the scale measures the

452 five personality factors, allowing unbiased cross-cultural comparisons (at least between the

453 Spanish and English versions), both in the context of latent variables as well as raw scores. In

454 other words, the differences in the Big Five observed scores between the Spanish-speaking and

455 English-speaking samples were due exclusively to differences in the measured traits, rather than

456 differences in language.

457       This is a remarkable achievement considering that we have tested the invariance between

458 items with two different wordings and two different languages (i.e., balance version with English

459 speakers *vs.* positive wording version with Spanish speakers). Given these results, it is certain

460   that the positive version of the Mini-IPIP would also reach a better level of invariance if

461   compared to a positive version in English.

462         An exception was the item "seldom feel blue," which demonstrated uniform differential

463   functioning. However, given that there is (apparently) no substantive reason to expect differential

464   functioning for this item, it is difficult to understand the nature of the discrepancy with any

465   certainty without gathering additional data. Nonetheless, at least three hypotheses can be

466   proposed: (a) the differential item functioning obeyed random sample fluctuations that in broader

467   samples would tend to disappear; (b) the item was understood differently by both samples and

468   hence would have to be reformulated; or (c) the DIF was not related to language but rather to

469   other sociodemographic characteristics that were not modeled (for example, differences in the

470   distribution of age or educational level). On the other hand, this item was also problematic in the

471   studies of Donellan et al. (2006) and Baldasaro et al. (2013). Future research should consider

472   replacing it with another item with more stable properties.

473         Third, the other short scales in Spanish used in this study, the adapted *PANAS-C10* and

474   the *SWLS*, showed good psychometric properties (reliability and validity) and support the

475   magnitudes of the relationships previously described in the scientific literature among the 5

476   personality factors, emotions, and life satisfaction. The relationships between some of the Big

477   Five scales (N and E) and self-reported affective states was medium in magnitude. It should be

478   noted the role of N, which appeared in this and previous studies related consistently,

479   significantly, and mainly with all criteria variables. Moreover, the relationship of these

480   personality factors with life satisfaction was low. The affective states factors had more relation

481   with life satisfaction than the personality factors. In addition, an ESEM model of two correlated

482   affective factors (positive and negative) was performed. This model showed similar fit indices

483    with or without a method factor ($\Delta\chi2 = -11.95$; $\Delta df = -1$; $\Delta$RMSEA $= -.007$; $\Delta$CFI $= .004$; $\Delta$TLI

484    $= .006$).

485          Finally, our results supported both Spanish versions (balance and positive wording). This

486    study provides the first evidence regarding the psychometric properties of a positive wording

487    Big-Five short measure. The PW version showed the best fit indices of the estimated models

488    (Table 1, M2-PW). Because they are simpler to understand and improve the results, we

489    recommend the use of PW items for research purposes if there is no exist other justified reason.

490    It is desirable to include items that improve the psychometric properties regardless of its

491    semantic polarity (e.g., content validity), and not only because they are negative. It is also

492    important to point out the possible confusion between trait-variance, acquiescence-variance and

493    social desirability. In the case of the PW version this may be relevant because all items were

494    written in the most socially accepted or natural way. These possible sources of variability may

495    require modeling more than one method factor.

496                                      **Conclusion**

497          This study has contributed to research on personality measurement by providing the first

498    psychometric properties of a short positive wording inventory in Spanish based on the Big Five

499    personality traits (the Mini-IPIP). One completely positive worded version of the Mini-IPIP and

500    other balanced (i.e., balance keying), showed reliability and validity indices according to theory

501    and prior studies' findings.

502          Using an exploratory structural factor approach (ESEM), a five-factor structure was

503    modelled after controlling acquiescence by data screening items. In those assessment conditions,

504    (a) the Big Five-dimensional structure was modeled satisfactorily in two different samples

505    (Spanish and English speakers), and (b) modeling a method factor did not improve fit indices.

506   Our results support the recommendation of Kam & Chan (2018) about the use of data screening

507   techniques to identify properly careless respondents. The use of negative items was not

508   psychometrically relevant in our study. In general, the PW version showed slightly better

509   psychometric indices than the balance one. This result is consistent with the study of Gnambs

510   and Schroeders (2017), where cognitive abilities can explain wording effects (i.e., worded

511   negative items add an extra difficulty for the participant to correctly understand the content).

512   Therefore, PW short versions of personality used after data screening techniques may be

513   appropriate for future studies.

514        The demonstration of sufficient measurement invariance with regard to the English

515   version of the Mini-IPIP is particularly interesting for future cross-cultural studies. It is even

516   more interesting to highlight the good psychometric performance shown by the scales regardless

517   of their wording (i.e., the invariance level achieved between the positive version in Spanish vs.,

518   the English balance version). We still have to think about what the causes of the method factors

519   are. In this study, this causes do not appear to be a main cause of misfit.

520        **Limitations.**

521        Some limitations of the study should be considered. Perhaps the most relevant of these is

522   the absence of direct evaluations of criterion validity through the inclusion of the Mini-IPIP in a

523   broader nomological network than the one deduced from the original version. Second, it would

524   be desirable in future studies to contrast the functioning of the Mini-IPIP in Spanish with other

525   broader personality measures in order to verify that its results converge sufficiently. Finally, we

526   would like to point out some considerations made by a reviewer of this manuscript: (a) the

527   Spanish-speaking sample was composed exclusively of university students of a limited age

528   range. The Mini-IPIP-PW may be less appropriate for non-university participants, where

529    individual differences in acquiescent responding would be more-pronounced (Rammstedt &

530    Farmer, 2013). However, the equivalency of the Spanish version with regard to the English-

531    speaking sample with more varied sociodemographic characteristics provides reasonable

532    evidence of the fact that our results are not biased due to the origin of the sample; (b) although

533    the results of the invariance analysis support the equivalence of the studied scales, the process of

534    reducing the English IPIP version to the Mini-IPIP does not have to converge in the same set of

535    items than those that could have resulted from abbreviating the scales with data from participants

536    in Spanish. It would be interesting to carry out studies to analyze if the same set of items

537    converge from different languages; (c) if data control techniques are not taken (e.g., bogus or

538    instructed items), the use of fully imbalanced scales such the Mini-IPIP-PW (i.e., only positively

539    worded items) can confound substantive variance with other sources of uncontrolled variance

540    (e.g., the acquiescent response style).

541        **Prospective.**

542        As noted by Baldasaro et al., (2013), the strategy to select the items of the Mini-IPIP

543    (those with the largest loading in their theoretical factor) possibly implies that the items represent

544    their respective constructs in a quite narrow way (i.e., lack of content validity). This is supported

545    by the need to relax correlations between residuals of items that are practically equal in wording

546    and content (especially in the Intellect factor, where the two pairs of items are clearly redundant).

547        In order to continue the validity exercise, we suggest include a larger samples and larger

548    variable sets (e.g., academic performance, depression). In future versions of the Mini-IPIP it

549    would be recommended to test new items with greater diversity of content and positive and

550    negative wording, taking advantage of the fact that recent analytical techniques (like ESEM)

551    facilitate the modelling of complex latent structures.

552        Finally, it is necessary perform simulations to examine the sensitivity of goodness of fit

553    indexes to lack of measurement invariance considering other features which are becoming more

554    frequent (e.g., ESEM, WLSMV, ordinal data).

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577                                    References

578    Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in*

579         *Personality*, 53:1-4. DOI: 10.1016/j.jrp.2014.07.001.

580    Arias, V. B., & Arias, B. (2017). The negative wording factor of Core Self-Evaluations Scale

581         (CSES): Methodological artifact, or substantive specific variance? *Personality and*

582         *Individual Differences*, 109:28-34. DOI: 10.1016/j.paid.2016.12.038.

583    Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural*

584         *Equation Modeling: A Multidisciplinary Journal*, 16:397-438. DOI:

585         10.1080/10705510903008204.

586    Baldasaro, R. E., Shanahan, M. J., & Bauer, D. J. (2013). Psychometric properties of the Mini-

587         IPIP in a large, nationally representative sample of young adults. *Journal of Personality*

588         *Assessment,* 95:74-84. DOI: 10.1080/00223891.2012.700466.

589    Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus

590         means and variance adjusted weighted least squares estimation in CFA. *Structural*

591         *Equation Modeling,* 13:186-203. DOI: 10.1207/s15328007sem1302_2.

592    Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for

593         two balanced sets of items. *Structural Equation Modeling,* 7:608–628. DOI:

594         10.1207/S15328007SEM0704_5.

595    Bruck, C. S., & Allen, T. D. (2003). The relationship between big five personality traits, negative

596         affectivity, type A behavior, and work–family conflict. *Journal of Vocational*

597         *Behavior*, 63:457-472. DOI: 10.1016/S0001-8791(02)00040-4.

598   Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications,*

599        *and programming*. Routledge.

600   Carvalho, H. W. D., Andreoli, S. B., Lara, D. R., Patrick, C. J., Quintana, M. I., Bressan, R. A.,

601        ... & Jorge, M. R. (2013). Structural validity and reliability of the Positive and Negative

602        Affect Schedule (PANAS): evidence from a large Brazilian community sample. *Revista*

603        *Brasileira de Psiquiatria*, 35:169-172. DOI: 10.1590/1516-4446-2012-0957.

604   Chen, F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance.

605        *Structural Equation Modeling,* 14:464-504. DOI: 10.1080/10705510701301834.

606   Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing

607        measurement invariance. *Structural Equation Modeling,* 9:233-255. DOI:

608        10.1207/S15328007SEM0902_5.

609   Chico, E. (2006). Personality dimensions and subjective well-being. *The Spanish Journal of*

610        *Psychology*, 9:38-44.

611   Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include

612        design-driven correlated residuals in latent-variable covariance structure analysis.

613        *Psychological methods*, 12:381-398. DOI: 10.1037/1082-989X.12.4.381.

614   Cooper, A. J., Smillie, L. D., & Corr, P. J. (2010). A confirmatory factor analysis of the Mini-

615        IPIP five-factor model personality scale. *Personality and Individual Differences*, 48:688-

616        691. DOI: 10.1016/j.paid.2010.01.004.

617   Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and*

618        *Individual Differences,* 13:653-665. DOI: 10.1016/0191-8869(92)90236-I.

619    Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a

620          personality variable. *The Journal of Abnormal and Social Psychology*, 60:151-174. DOI:

621          10.1037/h0040372.

622    Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the

623          consequences of using short measures of the Big Five personality traits. *Journal of*

624          *Personality and Social Psychology*, 102:874-888. DOI: 10.1037/a0027403.

625    Cupani, M. (2009). El cuestionario de personalidad ipip-ffm: resultados preliminares de una

626          adaptación en una muestra de preadolescentes argentinos [The IPIP-FFM personality

627          questionnaire: preliminary results of an adaptation in a sample of Argentine pre-teens].

628          *Perspectivas en Psicología*, 6:51-58.

629    Cupani, M., & Lorenzo-Seva, U. (2016). The development of an alternative IPIP inventory

630          measuring the Big-Five factor markers in an Argentine sample. *Personality and*

631          *Individual Differences*, 91:40-46. DOI: 10.1016/j.paid.2015.11.051.

632    Damásio, B. F., Pacico, J. C., Poletto, M., & Koller, S. H. (2013). Refinement and psychometric

633          properties of the eight-item Brazilian Positive and Negative Affective Schedule for

634          Children (PANAS-C8). *Journal of Happiness Studies*, 14:1363-1378.

635    DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for

636          data screening. *Journal of Organizational Behavior*, 36:171-181.

637    Deshon, R. P. (2004). Measures are not invariant across groups with error variance homogeneity.

638          *Psychology Science,* 46:137-49.

639    Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life

640          scale. *Journal of personality assessment*, 49:71-75. DOI:

641          http://10.1207/s15327752jpa4901_13.

642 Donellan, M., Oswald, F., Baird, B. & Lucas, R. (2006). The mini-IPIP scales: tiny-yet-effective

643      measures of the Big Five factors of personality. *Psychological Assessment, 18*(2), 192-

644      203. DOI: 10.1037/1040-3590.18.2.192.

645 Eid, M. (2000). A multitrait-multimethod model with minimal assumptions.

646      *Psychometrika*, 65:241-261.

647 Eys, M. A., Carron, A. V., Bray, S. R., & Brawley, L. R. (2007). Item wording and internal

648      consistency of a measure of cohesion: The Group Environment Questionnaire. *Journal of*

649      *Sport and Exercise Psychology*, 29:395-402. DOI: 10.1123/jsep.29.3.395.

650 Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model

651      types. *Multivariate Behavioral Research*, *42*(3), 509-529. DOI:

652      10.1080/00273170701382864.

653 Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H.,

654      Legaz, E., Ferrís, G., & Rojo-Moreno, L. (2018). Is small still beautiful for the strengths

655      and difficulties questionnaire? Novel findings using exploratory structural equation

656      modeling. *Assessment*. DOI: 10.1177/1073191118780461.

657 Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the

658      Rosenberg Self-Esteem Scale. *Assessment*. DOI: 10.1177/1073191117746503.

659 Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure.

660      *Psychological Assessment,* 4:26-42.

661 Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the

662      lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. de Fruyt,

663      and F. Ostendorf (Eds.). *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg:

664      Tilburg University Press.

665    González-Arias, M., Martínez-Molina, A., Galdames, S., & Urzúa, A. (2018). Psychometric

666        properties of the 20-item Toronto Alexithymia Scale in the Chilean population. *Frontiers*

667        *in psychology*, 9:963. DOI: 10.3389/fpsyg.2018.00963.

668    Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the

669        dimensionality of the Rosenberg Self-Esteem Scale: do they matter?. *Personality and*

670        *individual differences*, 35(6), 1241-1254. DOI: 10.1016/S0191-8869(02)00331-8.

671    Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about

672        model fit. *European Journal of Psychological Assessment*, 33:313–317. DOI:

673        10.1027/1015-5759/a000450.

674    Greiff, S., & Scherer, R. (2018). Still Comparing Apples With Oranges? *European Journal of*

675        *Psychological Assessment*, 34:141–144. DOI: 10.1027/1015-5759/a000487.

676    Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit

677        indexes with respect to violations of uncorrelated errors. *Structural equation modeling: a*

678        *multidisciplinary journal*, 19:36-50. DOI: 10.1080/10705511.2012.634710.

679    Hayes, N., & Joseph, S. (2003). Big 5 correlates of three measures of subjective well-

680        being. *Personality and Individual differences*, 34:723-727. DOI: 10.1016/S0191-

681        8869(02)00057-0.

682    Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality

683        inventories be evaluated? *Personality and Social Psychology Review*, 14:332-346. DOI:

684        10.1177/1088868310361240.

685    Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an

686        insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828. DOI:

687        10.1037/a0038510.

688    International Test Commission (2005). International Guidelines on Test Adaptation. Retrieved

689         from https://www.intestcom.org/files/guideline_test_adaptation.pdf

690    Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public

691         domain inventory: Development of the IPIP-NEO-120. *Journal of Research in*

692         *Personality*, 51:78-89. DOI: 10.1016/j.jrp.2014.05.003.

693    Joshanloo, M., & Afshari, S. (2011). Big five personality traits and self-esteem as predictors of

694         life satisfaction in Iranian Muslim university students. *Journal of Happiness*

695         *Studies*, 12:105-113. DOI: 10.1007/s10902-009-9177-y.

696    Kam, C. C. S., & Chan, G. H. H. (2018). Examination of the validity of instructed response items

697         in identifying careless respondents. *Personality and Individual Differences*, 129:83-87.

698         DOI: 10.1016/j.paid.2018.03.022.

699    Konstabel, K., Lönnqvist, J. E., Leikas, S., Velázquez, R. G., Qin, H., Verkasalo, M., &

700         Walkowitz, G. (2017). Measuring single constructs by single items: Constructing an even

701         shorter version of the "Short Five" personality inventory. *PloS one*, 12:e0182714. DOI:

702         10.1371/journal.pone.0182714.

703    Lai, J. C. (1994). Differential predictive power of the positively versus the negatively worded

704         items of the Life Orientation Test. *Psychological Reports*, 75:1507-1515.

705         DOI.10.2466/pr0.1994.75.3f.1507.

706    Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when

707         RMSEA and CFI disagree. *Multivariate behavioral research*, 51:220-239. DOI:

708         10.1080/00273171.2015.1134306.

709    Laverdière, O., Morin, A. J. S., & St-Hilaire, F. (2013). Factor structure and measurement

710         invariance of a short measure of the Big Five personality traits. *Personality and*

711         *Individual Differences,* 55:739-743. DOI: 10.1016/j.paid.2013.06.008.

712    Linacre, J.M. (2018). Winsteps® (Version 4.1.0) [Computer Software]. Beaverton, Oregon:

713         Winsteps.com. Retrieved January 1, 2018. Available from https://www.winsteps.com/

714    Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its

715         effects on research. *Journal of Research in Personality*, 48:61-83. DOI:

716         10.1016/j.jrp.2013.09.008.

717    Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-

718         Haenszel procedure. *Journal of the American Statistical Association*, 58:690-700.

719    Markey, P. M., & Markey, C. N. (2009). A brief assessment of the interpersonal circumplex: The

720         IPIP-IPC. *Assessment,* 16:352-361. DOI: 10.1177/1073191109340382.

721    Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., &

722         Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory

723         structural equation modeling. *Psychological assessment*, 22:471-491. DOI:

724         10.1037/a0019227.

725    Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation

726         modeling: An integration of the best features of exploratory and confirmatory factor

727         analysis. *Annual Review of Clinical Psychology*, 1:85-110. DOI: 10.1146/annurev-

728         clinpsy-032813-153700.

729    Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor

730         structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable

731         response styles. *Psychological assessment*, 22:366-381. DOI 10.1037/a0019225.

732  Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis.

733     *Psychological Methods*, 11:344-362. DOI: 10.1037/1082-989X.11.4.344.

734  McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement

735     quality and fit index cutoffs in latent variable models. *Journal of personality*

736     *assessment*, 100:43-52. DOI: 10.1080/00223891.2017.1281286.

737  Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data.

738     *Psychological Methods*, 17:437-455. DOI: 10.1037/a0028085.

739  Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance.

740     *Psychometrika,* 58:525-543.

741  Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical

742     measures. *Multivariate Behavioral Research,* 39:479-515. DOI: 10.1207/

743     S15327906MBR3903_4.

744  Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation

745     modeling framework for the identification of distinct sources of construct-relevant

746     psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary*

747     *Journal,* 23:116-139. DOI: 10.1080/10705511.2014.961800.

748  Moyano-Díaz, E., Martínez-Molina, A., & Ponce, F. P. (2014). The price of gaining:

749     maximization in decision-making, regret and life satisfaction. *Judgment and Decision*

750     *Making*, 9:500-509.

751  Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de

752     los tests: segunda edición [Guidelines for translating and adapting tests: second edition].

753     *Psicothema*, 25:151-157.

754  Muthén, L. K., & Muthén, B. O. (2014). Mplus 7.2. Los Angeles: Muthén and Muthén.

755 Oliveira, J. P. (2017). Psychometric Properties of the Portuguese Version of the Mini-IPIP five-
756       Factor Model Personality Scale. *Current Psychology*, 1-8. DOI: 10.1007/s12144-017-
757       9625-5.

758 Oppenheimer, D. A., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks:
759       Detecting satisficing to increase statistical power. *Journal of Experimental Social*
760       *Psychology,* 45:867–872. DOI: 10.1016/j.jesp.2009.03.009.

761 Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method
762       biases in behavioral research: A critical review of the literature and recommended
763       remedies. *Journal of Applied Psychology*, 88 (5), 879-903.

764 Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of
765       personality structure. *Psychological assessment*, *25*(4), 1137. DOI: 10.1037/a0033323.

766 Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five personality measurements
767       for acquiescence: An 18-country cross-cultural study. *European Journal of*
768       *Personality*, 27:71-81. DOI: 10.1002/per.1894.

769 Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and
770       negative item wording and its influence on the assessment of callous-unemotional
771       traits. *Psychological Assessment*, 28:394-404. DOI: 10.1037/pas0000183.

772 Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied*
773       *Psychological Measurement,* 21:173-184.

774 Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered
775       categorical data within a measurement invariance framework: A comparison of
776       estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21:167-180.
777       DOI: 10.1080/10705511.2014.882658.

778    Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or

779        detection of misspecifications?. *Structural Equation Modeling*, 16:561-582. DOI:

780        10.1080/10705510903203433.

781    Schreiber, J. B. (2017). Update to core reporting practices in structural equation

782        modeling. *Research in Social and Administrative Pharmacy*, 13:634-643. DOI:

783        10.1016/j.sapharm.2016.06.006.

784    Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form

785        development. *Psychological Assessment,* 12:102–111. DOI: 10.1037/1040-

786        3590.12.1.102.

787    Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance

788        literature: Suggestions, practices, and recommendations for organizational research.

789        *Organizational Research Methods,* 3:4-70. DOI:

790    Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects

791        in mixed-format scales. *Educational and Psychological Measurement*, 75:157-178. DOI:

792    Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the positive and negative affect*

793        *schedule-expanded form*. Iowa: The university of Iowa.

794    Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures

795        of positive and negative affect: the PANAS scales. *Journal of personality and social*

796        *psychology*, 54:1063-1070.

797    Whittaker, T. A. (2012). Using the modification index and standardized expected parameter

798        change for model modification. *The Journal of Experimental Education*, 80:26-44. DOI:

799        10.1080/00220973.2010.531299.

800     Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative

801         model. *Psychological Methods*, 18:320-334. DOI: 10.1037/a0032121.

802     Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems

803         and potential solutions. *Educational assessment*, 10:1-17. DOI:

804         10.1207/s15326977ea1001_1.

805     Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and

806         updating the practice of multi-group confirmatory factor analysis: A demonstration with

807         TIMSS data. *Practical Assessment, Research and Evaluation*, 12:1-26.

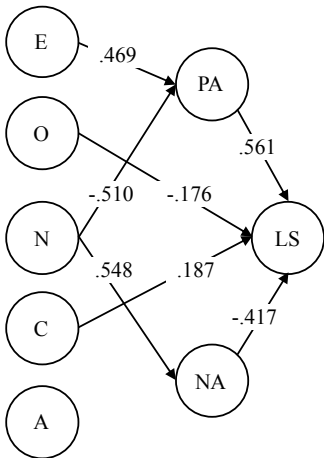
808     Żemojtel-Piotrowska, M., Piotrowski, J. P., Cieciuch, J., Adams, B. G., Osin, E. N., Ardi, R., ...

809         & Esteves, C. (2017). Measurement invariance of Personal Well-being Index (PWI-8)

810         across 26 countries. *Journal of Happiness Studies*, 18:1697-1711. DOI:

811         http://10.1007/s10902-016-9795-0.

Estimated SEM model for validity purposes.

Two Mini-IPIP versions, regular and Positive Wording: validity and cross-cultural invariance E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; LS = The Satisfaction with Life Scale; PA = Positive Affective; NA = Negative Affective; only significant parameters are shown; $R^2PA= .54$, $R^2NA = .62$, $R^2LS = .47$.

Fit indices of the estimated models

*Note.* RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; ESEM = Exploratory Structural Equation Model; CU = Correlated Uniqueness; RI = Random Intercept factor; PW = Positive Wording; us = United States sample; Bold measurement models were selected for the invariance tests; Invariance models by ΔRMSEA ≤ .015 are in bold; *In italic the invariance results of the PW-Models using the positive wording items in the Spanish sample vs. the regular English version (balanced) in the English sample.*

1 **Table 1.** Fit indices of the estimated models

|  | Model | Type | RMSEA | CFI | TLI | χ2 | df | ΔRMSEA | ΔCFI | ΔTLI | Δχ2 | Δdf |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measurement | M1 | ESEM | .084 | .929 | .864 | 469 | 100 | | | | | |
| | M1-PW | ESEM | .082 | .96 | .924 | 288 | 100 | | | | | |
| | **M2** | **ESEM+CU** | **.053** | **.973** | **.947** | **240** | **98** | | | | | |
| | **M2-PW** | **ESEM+CU** | **.052** | **.984** | **.970** | **171** | **98** | | | | | |
| | M3 | ESEM+RI | .068 | .954 | .912 | 337 | 99 | | | | | |
| | M3-PW | ESEM+RI | .054 | .983 | .968 | 178 | 99 | | | | | |
| | M4 | ESEM+CU+RI | .052 | .973 | .948 | 235 | 97 | | | | | |
| | M4-PW | ESEM+CU+RI | .053 | .984 | .969 | 172 | 97 | | | | | |
| | M5 | ESEM+CU+RI (us) | .064 | .974 | .949 | 241 | 97 | | | | | |
| Invariance | *Language, Spanish-English* | | | | | | | | | | | |
| | M6 | Configural | .057 | .974 | .949 | 478 | 196 | | | | | |
| | **M7** | **Strong** | **.068** | **.938** | **.927** | **994** | **326** | **.011** | **-.036** | **-.022** | **516** | **130** |
| | **M8** | **Strict** | **.069** | **.932** | **.925** | **1079** | **346** | **.012** | **-.042** | **-.024** | **601** | **150** |
| | **M7p** | **Strong (partial)** | **.060** | **.955** | **.944** | **796** | **310** | **.003** | **-.019** | **-.005** | **318** | **114** |
| | **M8p** | **Strict (partial)** | **.060** | **.952** | **.944** | **837** | **326** | **.003** | **-.022** | **-.005** | **359** | **130** |
| | *M6-PW* | *Configural* | *.058* | *.979* | *.960* | *409* | *196* | | | | | |
| | *M7-PW* | *Strong* | *.088* | *.920* | *.907* | *1141* | *326* | *.030* | *-.059* | *-.053* | *732* | *130* |
| | *M8-PW* | *Strict* | *.091* | *.911* | *.902* | *1264* | *346* | *.033* | *-.068* | *-.058* | *855* | *150* |
| | ***M7-PWp*** | ***Strong (partial)*** | ***.060*** | ***.967*** | ***.956*** | ***621*** | ***286*** | ***.002*** | ***-.012*** | ***-.004*** | ***212*** | ***90*** |
| | ***M8-PWp*** | ***Strict (partial)*** | ***.070*** | ***.955*** | ***.942*** | ***757*** | ***294*** | ***.012*** | ***-.024*** | ***-.018*** | ***348*** | ***98*** |

2 *Note.* RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis
3 Index; ESEM = Exploratory Structural Equation Model; CU = Correlated Uniqueness; RI = Random Intercept
4 factor; PW = Positive Wording; us = United States sample; Bold measurement models were selected for the
5 invariance tests; Invariance models by ΔRMSEA ≤ .015 are in bold; *In italic the invariance results of the PW-*
6 *Models using the positive wording items in the Spanish sample vs. the regular English version (balanced) in the*
7 *English sample.*

Factor loadings, correlations and reliability of M2 and M2-PW (ESEM + CU)

*Note.* E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; # = Item administration order; R = Regular version; PW = Positive wording version; CR = Composite reliability; Loadings > .20 and p < 0.01 are shown; Main loadings are in bold; Factor correlations p < 0.01 are in bold.

1    **Table 2**. Factor loadings, correlations and reliability of M2 and M2-PW (ESEM + CU)

| Factor | # | F1 R | F1 PW | F2 R | F2 PW | F3 R | F3 PW | F4 R | F4 PW | F5 R | F5 PW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| E | 1 | **.824** | **.802** | - | - | - | - | - | - | - | - |
|  | 6 | **.730** | **.756** | - | - | - | - | - | - | - | - |
|  | 11 | **.784** | **.794** | - | - | - | - | - | - | - | - |
|  | 16 | **.718** | **.739** | - | - | - | - | - | - | - | - |
| N | 4 | - | - | **.732** | **.740** | - | - | - | - | - | - |
|  | 9 | - | - | **.568** | **.466** | - | - | .258 | - | - | - |
|  | 14 | - | - | **.634** | **.568** | - | - | - | - | - | - |
|  | 19 | - | - | **.543** | **.667** | - | - | - | - | - | - |
| A | 2 | - | - | - | - | **.598** | **.791** | - | - | - | - |
|  | 7 | - | - | - | - | **.800** | **.913** | - | - | - | - |
|  | 12 | - | - | - | - | **.500** | **.836** | - | - | - | - |
|  | 17 | - | - | - | - | **.806** | **.845** | - | - | - | - |
| C | 3 | - | - | - | - | - | - | **.557** | **.593** | - | - |
|  | 8 | - | - | - | - | - | - | **.688** | **.776** | - | - |
|  | 13 | - | - | - | - | - | - | **.711** | **.724** | - | - |
|  | 18 | - | - | - | - | - | - | **.946** | **.945** | - | - |
| O | 5 | - | - | - | - | - | - | - | - | **.706** | **.831** |
|  | 10 | - | - | - | - | - | - | - | - | **.480** | **.498** |
|  | 15 | - | - | - | - | - | - | - | - | **.567** | **.477** |
|  | 20 | - | - | - | - | - | - | - | - | **.878** | **.959** |
|  | CR | .94 | .93 | .90 | .87 | .92 | .94 | .93 | .92 | .91 | .90 |
|  | F1 | - | - |  |  |  |  |  |  |  |  |
|  | F2 | **−.270** | −.146 | - | - |  |  |  |  |  |  |
|  | F3 | **.251** | .059 | **−.145** | .060 | - | - |  |  |  |  |
|  | F4 | −.010 | .009 | −.101 | −.043 | **.164** | .102 | - | - |  |  |
|  | F5 | **.234** | **.205** | −.107 | −.033 | **.193** | .131 | −.022 | −.069 | - | - |

2    *Note.* E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; # = Item
3    administration order; R = Regular version; PW = Positive wording version; CR = Composite reliability; Loadings >
4    .20 and p < 0.01 are shown; Main loadings are in bold; Factor correlations p < 0.01 are in bold.

Descriptive, reliability and validity indices

*Note.* SWLS = The Satisfaction with Life Scale; PA = Positive Affective; NA = Negative

Affective; E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O

= Openness; PW= Positive Wording version; i = number of items in the scales; α =

Cronbach's α; ω = McDonald's ω; Pearson's Correlations > .20 or with p < 0.05 are shown.

1 **Table 3**. Descriptive, reliability and validity indices

| | SWLS | PA | NA | E | N | A | C | O | E_PW | N_PW | A_PW | C_PW | O_PW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PA | .497 | | | | | | | | | | | | |
| NA | −.404 | −.411 | | | | | | | | | | | |
| E | −.231 | .467 | −.192 | | | | | | | | | | |
| N | −.253 | −.404 | .482 | −.216 | | | | | | | | | |
| A | - | - | - | .229 | −.073 | | | | | | | | |
| C | .221 | - | −.163 | - | −.143 | .082 | | | | | | | |
| O | - | .131 | - | .178 | −.074 | .179 | - | | | | | | |
| E_PW | .248 | .472 | −.173 | .943 | - | - | - | .172 | | | | | |
| N_PW | −.229 | −.370 | .448 | - | .876 | - | - | - | - | | | | |
| A_PW | - | - | - | - | - | .924 | - | - | - | - | | | |
| C_PW | .238 | - | −.133 | - | - | - | .925 | - | - | - | - | | |
| O_PW | - | .127 | - | .216 | - | .130 | - | .904 | .197 | - | - | - | |
| i | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| M | 18.13 | 18.41 | 8.39 | 12.24 | 10.82 | 15.92 | 12.72 | 15.42 | 11.86 | 10.74 | 15.62 | 12.65 | 15.43 |
| SD | 3.65 | 2.78 | 3.11 | 3.56 | 3.20 | 2.69 | 3.54 | 2.97 | 3.28 | 2.82 | 2.73 | 3.22 | 2.85 |
| SK | −0.77 | −0.24 | 1.30 | −0.15 | 0.23 | −0.88 | −0.14 | −0.55 | 0.03 | 0.29 | −0.79 | −0.19 | −0.65 |
| K | 0.89 | 0.85 | 1.63 | −0.33 | −0.24 | 1.25 | −0.62 | 0.11 | −0.10 | −0.09 | 1.61 | −0.44 | 0.38 |
| α | .85 | .83 | .78 | .84 | .65 | .82 | .78 | .79 | .82 | .63 | .86 | .80 | .77 |
| ω | .86 | .84 | .83 | .84 | .67 | .85 | .80 | .81 | .86 | .69 | .95 | .86 | .82 |

2 *Note.* SWLS = The Satisfaction with Life Scale; PA = Positive Affective; NA = Negative Affective; E =
3 Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; PW= Positive Wording
4 version; i = number of items in the scales; α = Cronbach's α; ω = McDonald's ω; Pearson's Correlations > .20 or
5 with $p < 0.05$ are shown.