

THE VALIDITY AND RELIABILITY OF THE INDONESIAN IPIP-NEO-120

MEDIANTA TARIGAN

INDONESIA UNIVERSITY OF EDUCATION, BANDUNG, INDONESIA

FADILLAH

BANDUNG INSTITUTE OF TECHNOLOGY, INDONESIA

IPIP-NEO-120 is a scale consisting of 120 items representing the Five Factor Model (FFM) of personality that can be accessed for free on the IPIP website. This study aims to explain the psychometric properties of The Indonesian IPIP-NEO-120 through its validity and reliability tests. The construct validity was tested using confirmatory factor analysis (CFA) and reliability was tested by measuring the composite reliability value and assessing Cronbach's alpha estimation. This study involved 624 participants ranging from 14 to 69 years (67% females, 33% males), who had different levels of education. The results showed that the majority of items were found to be valid in measuring their respective dimensions. The reliability test also had valid results with the construct reliability value exceeding .70 for all dimensions and Cronbach's alpha index ranging from .65 to .90. This study also found that some items were identified as potentially problematic. It is recommended that further analysis may be required for certain items.

Keywords: IPIP-NEO-120; Five Factor Model; Validity; Reliability; Factor analysis.

Correspondence concerning this article should be addressed to Medianta Tarigan, Department of Psychology, Indonesia University of Education, Jl. Dr. Setiabudhi 229, 40154 Bandung, Indonesia. Email: medianta@upi.edu

Over the past few years, Indonesia has experienced a slowdown in psychological test development. Until today, most psychological tests that are widely used by practitioners are considered obsolete. These tests were developed by experts in other countries and translated into Indonesian between the 1940s and 1990s. Moreover, the process of adaptation as well as the quality of the psychometric properties of these tests is unknown (Suwartono et al., 2017). In contrast, American Psychological Association (2017) Ethics Code forbids psychologists from making recommendations based on obsolete tests. This problem of obsolete tests is compounded by the problem of leakage of test items via the Internet, especially for popular instruments used in high stake tests such as preemployment tests or educational tests. It is stated that psychological test materials should be kept private to prevent problems with test administration, scoring, or interpretation, such as when an individual knows the content of the test item (American Educational Research Association et al., 2014). The exposure of test items to the public via the Internet leads to misuse and loss of effective assessment objectives, especially in providing reliable and valid score interpretation results. Thus, the need for newer psychological tests with good psychometric quality is certainly important today. The emergence of newer tests will provide clear benefits for both practitioners and researchers.

In the 2000s, psychological test development in Indonesia began to emerge. However, this effort is still limited by the amount of resources needed, such as research costs and the availability of experts. One of the things that can be done is to adapt newer tests because this is usually considered cheaper and faster (Geisinger, 1994). The adaptation of tests being carried out in Indonesia concerns personality tests based on the basic personality theory of the five major personality factors (PFs). The five major personality factors have been widely researched, accepted, and used in various countries (Goldberg, 1990; John et al., 2008).

This theory has a long history of research dating back to Allport and Cattell's early effort to provide an initial personality structure in the 1930s (John & Srivastava, 1999).

It was observed that there are two distinct historical pathways in research concerning the five major personality factors, that is, the lexical tradition and the questionnaire tradition. Using a lexical approach by asking research participants to rate themselves on hundreds of individual trait words, Goldberg (1981) defines five factors that support the personality model called the "Big Five." In the lexical tradition, a personality model is developed through a systematic process of ordering and naming individual differences in people's behavior and mapping the factors based on the similarity of certain adjectives in the natural English. In contrast to the lexical approach, the questionnaire tradition emphasizes top-down research so that measurements can represent theoretical constructs. This approach assumes that the bottom-up lexical approach is not sufficiently convincing for a truly scientific description of personality (McCrae, 1990). Costa and McCrae (1976) conducted a cluster analysis on 16PFs and found many redundant items, with overlapping meanings (Johnson, 2014). Their longitudinal study uncovered a cluster of personality dimensions called the Five Factor Model (FFM) and developed an instrument called NEO-PI. Although the descriptions and labels of the FFM sometimes differ in content, the formulation by Costa and McCrae seems to be the most widely accepted (Marusic et al., 1996).

Goldberg later initiated a global collaborative effort for researchers in personality test development to contribute to continuously improving the personality item pool on an international network called the International Personality Item Pool (IPIP; Goldberg et al., 2006). In this network, experts around the globe can develop, refine, and adapt several personality inventory item banks on an ongoing basis. All items are freely used for personality scale development, for both scientific research and commercial purposes (Barrick & Ryan, 2003). One of the first personality measures to be created from IPIP was the 300-item IPIP-NEO by Goldberg (1999), which provides a comprehensive personality assessment with the 30 facet scales in the NEO Personality Inventory.

Despite the validity of the interpretation test score, the 300-item IPIP-NEO has one major drawback: it takes a long time which hinders its efficiency (Johnson, 2014). Too many items are suspected to cause participant fatigue, decreasing interest in completing the test, and beginning to respond randomly (Berry et al., 1991). For this reason, several personality measurements were then made shorter, including the IPIP-NEO-120 by Johnson (2014). Although there are fewer item versions of IPIP-NEO such as versions 20, 50, and 100 items, none of these personality inventories can comprehensively measure the five domains (Neuroticism, Extraversion, Openness to experience, Agreeableness, and Conscientiousness) of the Five Factor Model and six facets of each domain (Johnson, 2014). The IPIP-NEO-120 has undergone various language adaptations in several countries, including German (Renner et al., 2014) and Dutch (Blanken et al., 2018) adaptation. This study aims to examine the validity of the Indonesian adaptation of IPIP-NEO-120.

In this study, a back translation process was carried out by translating the instrument into Bahasa Indonesian and then retranslated into English (Beaton et al., 2000). In terms of adapting a test, *Standards* (American Educational Research Association et al., 2014) provide relevant direction that whenever tests are translated evidence of the validity of scores on the adapted versions of the tests should be collected and reported. Standards define validity as the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. This definition is an adaptation of Messick's (1989) validity definition, where validity is a matter of degree and five types of evidence can be collected cumulatively over time to support test score interpretation. One of the types of evidence, structural evidence, seeks to uncover support for underlying structure, most commonly through factor analysis. Factor analysis examines how underlying structure influences the responses on some measured variables. There

are two types of factor analysis: confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). While exploratory factor analysis attempts to discover the nature of constructs or identify the underlying dimensional structure, confirmatory analysis attempts to test whether an a priori dimensional structure is consistent with the structure obtained in a particular set of measures. The International Test Commission (ITC, 2010) which provides guidelines for translating and adapting tests states that researchers are required to address construct equivalence from the test that is being adapted and this can be approached by using CFA. CFA is the most frequently used statistical technique to assess whether a construct in one culture is found in the same form and frequency in another culture (van de Vijner & Poortinga, 2002). Therefore, in this study CFA will be used as a statistical approach to provide evidence of the internal structure validity.

METHOD

Participants

This research uses purposive sampling and the inclusion criteria are Indonesian citizens with Bahasa as their first language. Participants who were willing to engage in the research were given informed consent and completed the test online. This study was ethically approved by the Research Ethics Committee of the Faculty of Psychology, Gadjah Mada University with approval number: 7245/UN1/FPSi.1.3/SD/PT.01.04/2021. Regarding the demographic breakdown from 624 participants, 67% of the sample was female ($n = 418$) and 33% male ($n = 206$) aged between 14-69 years; 15.14% had attended secondary school, 7.65% had a high-school diploma, and 77.2% had a university degree. When participating in this study, all participants had no special knowledge about the nature of IPIP-NEO-120, thus it is unlikely that such knowledge influenced their answering behavior. This information was gathered through the question: "Have you ever taken the same or a similar test to the one you just took?" with a *yes/no* response, that was given after completing IPIP-NEO-120. Completion of the test takes approximately 60 minutes.

Instrument

IPIP-NEO-120 consists of 120 items representing the Five Factor Model (FFM) of personality which can be accessed for free through the IPIP website. Each item is measured on a scale from 1 (*rarely*) to 5 (*almost always*). A back-translation process was carried out with stages as follows. The first stage was to translate the original items where the measuring instrument was translated by two people: the first person was an informed translator who knew the IPIP-NEO-120 personality concept; the second person was the translator who translated without knowing the IPIP-NEO-120 personality concept. The second stage was synthesis. In this stage, translation results carried out by the two previous translators were further processed to be compared and the equivalent word closest to the original meaning was chosen. The third stage was back translation, where items translated into Indonesian were translated back into English. After being translated again, the next process was a review by the linguists who had been previously appointed in this study. In the final stage, finalization was carried out by considering the suitability of the instrument language with the spoken language to find out the differences in the cultural context between the two translators. The following (Table 1) is an example of a back-to-back translation process.

TABLE 1
 Back-to-back translation process

Original statement	Find it difficult to approach others
Result of translation 1	Merasa sulit untuk mendekati orang lain
Result of translation 2	Merasa kesulitan mendekati orang lain
Result of synthesis	Merasa kesulitan melakukan pendekatan kepada orang lain
Back translation	Find it difficult to approach other people

Data Analysis

This study uses CFA method. CFA is used to determine the extent to which all items from The Indonesian IPIP-NEO-120 measure the same factor or are unidimensional. Some of the steps carried out in the CFA method are establishing model specifications, describing the theoretical model, and then describing it in a path diagram. The parameter estimation is carried out using the maximum likelihood method. The next step is to match the model with some parameter indices. Because it is confirmatory, the index that is mostly used is chi-square. If the model does not directly fit, it requires modifications where the measurement error in each item is correlated. The fit index can be increased by reducing the chi-square value in the model. Minimizing the chi-square can be done by freeing measurement errors on several correlated items (Arbuckle, 2005). However, the chi-square is also sensitive to the sample size, where the sample size is over 400 most models are rejected (Byrne, 1998; Stone, 2021). Therefore, other parameter indices are used such as RMSEA, CFI, GFI, AGFI, and TLI which are relatively insensitive to sample size (Brown, 2015; Fan et al., 1999; Marsh et al., 2004). If the model is accepted, it can continue with the significance test. In this study, the model used is 2nd order CFA in each dimension. A minimum cut-off point of .30 is used for item review where a standard loading factor (SLF) > |.30| is accepted as a valid item (Thompson, 2006). This analysis is assisted by Amos Graphics Software.

RESULTS

Descriptive Analysis

In Table 2 the results of the processing of the 624 participants' data are presented.

TABLE 2
 Descriptive analysis

Dimension	Mean	SE	SD	Var	Min	Max	Mean (male)	Mean (female)	Diff mean (z-test)
Neuroticism	69.80	0.51	12.78	163.26	31	105	65.59	71.84	6.25***
Extraversion	77.28	0.32	8.06	64.97	53	106	78.47	76.71	1.75*
Openness to experience	79.57	0.31	7.68	59.03	59	105	80.00	79.36	.65
Agreeableness	83.56	0.24	5.97	35.62	65	99	83.45	83.61	.17
Conscientiousness	86.98	0.42	10.61	112.51	51	117	88.62	86.18	2.44**

Note. SE = standard error; SD = standard deviation.
 * $p < .05$; ** $p < .01$; *** $p < .001$.

Based on Table 2, it can be seen that of the five dimensions, Conscientiousness has the largest average value, 86.98, while Neuroticism has the smallest average value, 69.80. Meanwhile, Neuroticism shows the largest standard deviation, which is 12.78 which indicates that the heterogeneity in the data is quite far from 0, meaning that the data is quite spread out and different from one another. The smallest standard deviation is found in the Agreeableness dimension, which is 5.97.

Confirmatory Factor Analysis

Initial testing on each dimension showed unfit results because some of the fit model criteria were not met, so a modification of the model was carried out by correlating the errors in each item (Tables 3 and 4).

TABLE 3
 Dimensional model fit test results (before correlating errors)

Level of fit	Acceptable value*	Model value				
		Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness
χ^2	$p > .05$	851.06	1368.36	1135.46	1135.46	1135.46
χ^2/df	< 5	3.46	5.56	4.60	4.60	4.60
RMSEA	$< .08$.06	.09	.08	.08	.08
GFI	$> .80$.89	.82	.85	.85	.85
AGFI	$> .80$.87	.78	.81	.81	.81
CFI	$> .80$.89	.75	.64	.64	.64
TLI	$> .80$.88	.72	.60	.60	.60

Note. *acceptable value recommendations from Baumgartner and Homburg (1996); Byrne and Campbell (1999); Carlbach and Wong (2018); Hu and Bentler (1999). *df* = degrees of freedom; RMSEA = root-mean-square error of approximation; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CFI = comparative fit index; TLI = Tucker-Lewis index.

TABLE 4
 Dimensional model fit test results (after correlating errors)

Level of fit	Acceptable value*	Model value				
		Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness
χ^2	$p > .05$	667.34	607.70	530.43	530.43	667.99
χ^2/df	< 5	2.78	2.88	2.38	2.31	2.82
RMSEA	$< .08$.05	.06	.05	.05	.05
GFI	$> .90$.91	.92	.93	.93	.92
AGFI	$> .90$.90	.90	.91	.91	.90
CFI	$> .90$.92	.91	.88	.91	.92
TLI	$> .90$.91	.88	.85	.88	.91

Note. *acceptable value recommendations from Baumgartner and Homburg (1996); Byrne and Campbell (1999); Carlbach and Wong (2018); Hu and Bentler (1999). *df* = degrees of freedom; RMSEA = root-mean-square error of approximation; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 4 shows that in chi-square all dimensions have a p -value $< .05$, which means that the model does not fit. However, the chi-square is quite sensitive to the large sample, so it is advisable to look at other model fit criteria. Even though some values do not exceed the excellent criteria, they still meet the acceptable value. The RMSEA values of each dimension are $< .08$. Among these, the Agreeableness dimension exhibits the closest fit to this threshold. Notably, GFI, AGFI, CFI, and TLI are slightly less than the good fit cut off values ($> .90$) and were all in acceptable areas $> .80$. Based on this result, the confirmatory analysis revealed that each dimension of The Indonesian IPIP-NEO-120 has a good model fit with the construct as in Figures 1 to 5.

The analysis is continued by looking at the validity of each facet against the appropriate dimensions. The results are shown in Table 5.

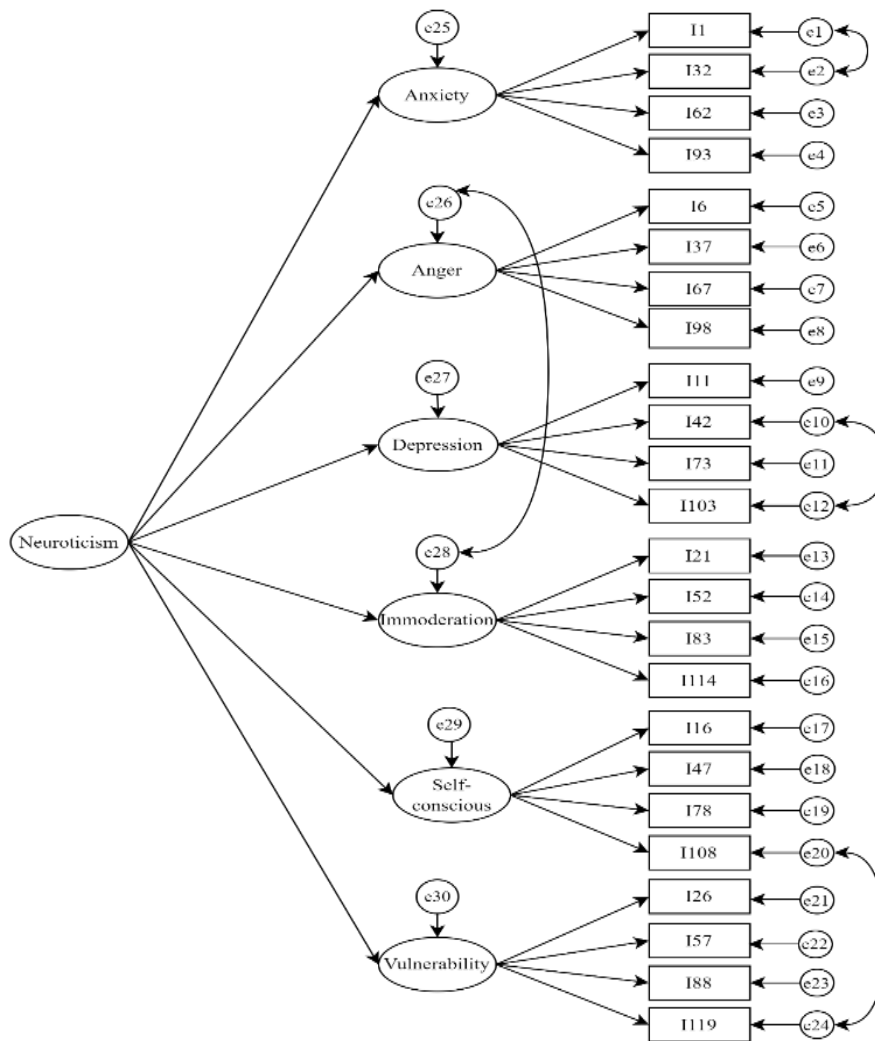


FIGURE 1
The construct of the confirmatory factor analysis of Neuroticism

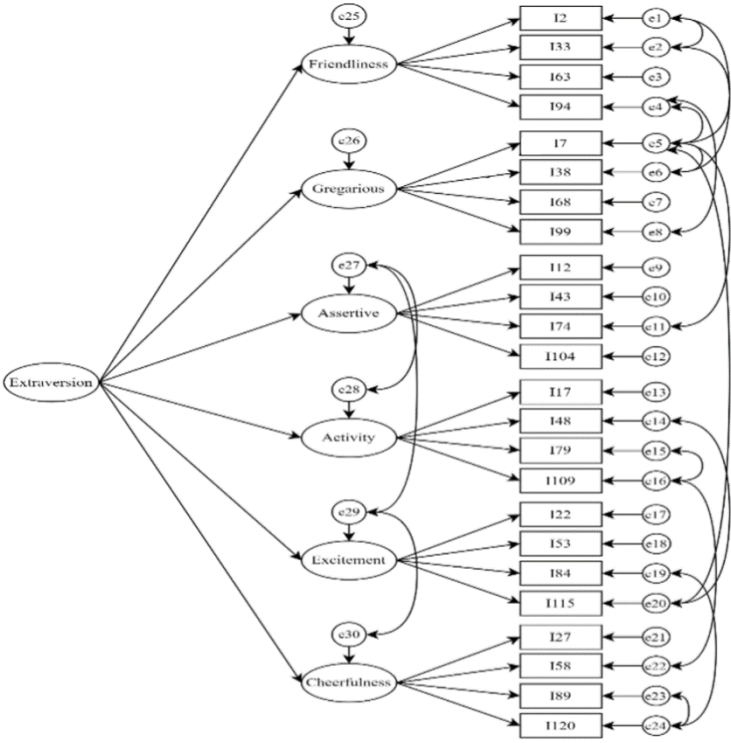


FIGURE 2
The construct of the confirmatory factor analysis of Extraversion

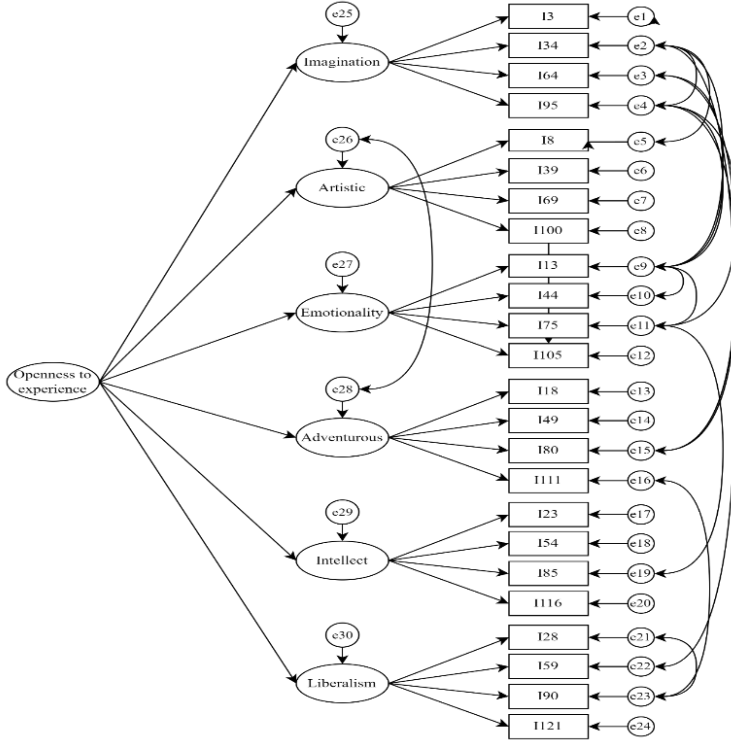


FIGURE 3
The construct of the confirmatory factor analysis of Openness to experience

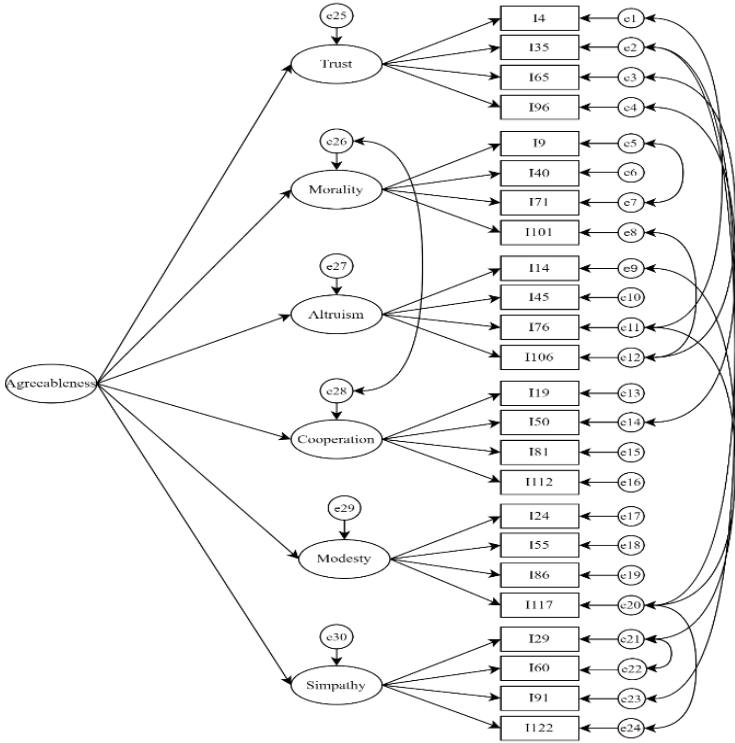


FIGURE 4
The construct of the confirmatory factor analysis of Agreeableness



FIGURE 5
The construct of the confirmatory factor analysis of Conscientiousness

TABLE 5
 Facet standardized loading factor

Facet trait	Standardized loading factor facet					Result
	Domain					
	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	
N_Anxiety	1.02					Valid
N_Anger	.63					Valid
N_Depression	.88					Valid
N_Immoderation	.46					Valid
N_Self-conscious	.85					Valid
N_Vulnerability	.98					Valid
E_Friendliness		1.06				Valid
E_Gregarious		.86				Valid
E_Assertive		.43				Valid
E_Activity		-.01				Not Valid
E_Excitement		.47				Valid
E_Cheerfulness		.58				Valid
O_Imagination			.27			Not Valid
O_Artistic			.83			Valid
O_Emotionality			.52			Valid
O_Adventurous			.66			Valid
O_Intellect			.80			Valid
O_Liberalism			.84			Valid
A_Trust				.40		Valid
A_Morality				.54		Valid
A_Altruism				1.01		Valid
A_Cooperation				-.59		Valid
A_Modesty				-.16		Not Valid
A_Sympathy				1.09		Valid
C_Self-efficacy					.86	Valid
C_Orderliness					.67	Valid
C_Dutifulness					.78	Valid
C_Achievement					.91	Valid
C_Self-discipline					1.04	Valid
C_Cautiousness					-.75	Valid

In this article, we used rotation methods to measure the validity of each indicator. The greater the value of the standardized loading factor, the more it indicates that the facet measures the appropriate dimensions. Table 5 shows that on the dimensions of Neuroticism and Conscientiousness, all facets are valid in measuring the appropriate dimensions (SLF > |.30|). In the Extraversion dimension, the Activity facet is not valid, but the other five facets are valid. On the Openness to experience dimension, the Imagination facet is not valid (SLF < |.30|) but five other facets are valid in measuring dimension. There is one facet that is not valid in measuring the Agreeableness dimension, the Modesty facet (SLF < |.30|), but the other five facets are valid in measuring the Agreeableness dimension.

From Table 6, it can be seen that 103 of 120 items are significant. Conscientiousness is a dimension that has all valid items, while Openness to experience has a higher number of invalid items (eight items).

TABLE 6
 Items' standardized loading factor

Domain Facet trait	Standardized loading factor item					
	Item 1	Item 2	Item 3	Item 4	Σ valid item	Σ not valid item
<i>Neuroticism</i>						
N_Anxiety	.72	.65	.79	.82	4	0
N_Anger	.83	.94	.74	.74	4	0
N_Depression	.81	.57	.87	.17	3	1
N_Immoderation	.02	.31	.60	.73	3	1
N_Self-conscious	.55	.62	.33	.41	4	0
N_Vulnerability	.77	.51	.67	.58	4	0
<i>Extraversion</i>						
E_Friendliness	.65	.62	.70	-.70	4	0
E_Gregarious	.43	.53	.70	.85	4	0
E_Assertive	.32	.72	.53	.63	4	0
E_Activity	.87	.87	.09	-.01	2	2
E_Excitement	.33	.55	-.22	.04	2	2
E_Cheerfulness	.70	-.80	-.73	-.76	4	0
<i>Openness to experience</i>						
O_Imagination	1.38	.18	.01	-.04	1	3
O_Artistic	.73	.52	.50	.67	4	0
O_Emotionality	.23	.51	.40	.72	3	1
O_Adventurous	.38	.60	.60	.28	3	1
O_Intellect	.51	.61	.59	.59	4	0
O_Liberalism	.27	.20	.05	-.36	1	3
<i>Agreeableness</i>						
A_Trust	.75	.50	.50	.71	4	0
A_Morality	.54	-.70	.62	.58	4	0
A_Altruism	.55	.40	.64	.45	4	0
A_Cooperation	.32	-.47	-.65	-.17	3	1
A_Modesty	.42	.67	.67	-.01	3	1
A_Sympathy	.43	.40	.27	.57	3	1
<i>Conscientiousness</i>						
C_Self-efficacy	.68	.66	.66	.54	4	0
C_Orderliness	.66	.53	.81	.70	4	0
C_Dutifulness	.70	.58	.58	.78	4	0
C_Achievement	.66	.39	.67	.63	4	0
C_Self-discipline	.57	.63	.76	.54	4	0
C_Cautiousness	.70	-.80	-.73	-.76	4	0
	Total				103	17
	Percentage				86%	14%

Note. Valid items in bold.

Overall, 14% (17 items) are not valid and these items are recommended to be reviewed or eliminated. Meanwhile, 86% of significant items (103 items) measure the appropriate facet. Furthermore, a reliability test was conducted to test data consistency by computing the items' reliability using construct reliability and Cronbach alpha reliability (Table 7).

The next analysis calculates item reliability to measure internal consistency. The results show that all items have high reliability with a CR index $> .70$. Based on Cronbach's alpha, Neuroticism and Conscientiousness have very high reliability meanwhile Extraversion, Openness to experience, and Agreeableness have high reliability.

TABLE 7
Reliability

Dimension	Construct reliability	Reliability (Cronbach's alpha)
Neuroticism	.94	.90
Extraversion	.88	.76
Openness to experience	.85	.71
Agreeableness	.78	.65
Conscientiousness	.90	.89

Discussion

This psychometric property validity test of The Indonesian IPIP-NEO-120 involved 624 respondents. The analysis begins with testing its internal structure validity using confirmatory factor analysis (CFA). Although this study showed that the overall model fit index is not as optimal as expected it showed adequate model-data fit (RMSEA = .047-.055; GFI = .913-.932; AGFI = .891-.909; CFI = .877-.925; TLI = .847-.914). Thus, the CFA analysis result shows that The Indonesian IPIP-NEO-120 has a fit model. It should be considered that personality trait inventories often perform poorly when their structure is evaluated by CFA due to the inherent complexity of personality (Borkenau & Ostendorf, 1990). Some studies of the FFM internal structure even show a failure to fit the CFA models and do not even approach the recommended index criteria (Hopwood & Donnellan, 2010).

In this study, most of the facet scales meet the criteria of loading factor (SLF $> |.30|$). It was also found that some facet scales show unexpected low loading factors, that is, Modesty, Imagination, and Activity. A previous study of the internal structure of the original IPIP-NEO-120 highlighted that some facet traits did not function sufficiently (e.g., Modesty) which might indicate that this facet should indeed be viewed as an independent trait or maybe contains a greater degree of social desirability than other facets meaning that it may not provide enough variance to support the trait (Kajonius & Johnson, 2019). In the development of the original IPIP-NEO-120, it was also found that the facet scale failed to show the expected highest loadings raising the question of whether the balance numbers of positively and negatively keyed items on IPIP-NEO-120 caused a response bias that might distort the factor structure. However, Johnson (2014) explained that, after carrying out a content-balanced test through the acquiescence index, there was no indication that unequal keying and acquiescence were the problem. Johnson also concluded that although there were several low factor loadings, the factor structure of IPIP-NEO-120 aligns very well with the factor structure of its parent, the NEO-PI-R. The NEO-PI-R is intended to provide a comprehensive description of personality

traits: the individual's characteristic and enduring emotional, interpersonal, experiential, attitudinal, and motivational styles (Costa & McCrae, 2008). In the construction process of IPIP-NEO-120, Johnson (2014) used balancing different strategies in the search for the best selection of items, where it was ensured that all the items selected were closely aligned conceptually with the NEO-PI-R items. This approach has different strategies from Goldberg's (1999) 300-IPIP-NEO scale construction which tried to get a balance between negative and positively scored items to prevent this acquiescence response bias (Vedel et al., 2019).

The result showed that 103 (86%) of the 120 items loaded $> |.30|$ on the expected factor. This is slightly more than in the Danish IPIP-NEO-120 with 100 (83%) items loaded on the expected factor and almost as much as the original IPIP-NEO-120 with 105 (87.5%) items which loaded as expected (Vedel et al., 2019). In this study, two dimensions (Extraversion and Openness to experience) have a higher invalid number of items than the other three dimensions. An item review was then carried out to find out whether the translation process affected the results. After qualitatively examining each problematic item, it was concluded that although the cultural context was considered in the back translation process cultural differences in understanding the item might still be the cause of these items failing to represent the intended facets. It is important to consider that people with Eastern culture respond to certain conditions differently, especially those related to conformity to norms and emotional self-control (Yamaoka, 2014). But it also needs to be noted that in the Danish version, although efforts have been made to change the context of the translation of invalid items, the analysis results still show a weak factor loading (Vedel et al., 2019). The following will describe the results of the analysis of the model for each dimension.

On the Neuroticism dimension, the results of the analysis show that the model is at the good fit level ($\chi^2/df = 2.78 < 5$; RMSEA = .05 \approx .05; GFI = .91 $>$.90; AGFI = .90 \approx .90; CFI = .92 $>$.90; TLI = .91 $>$.90). This is in line with the CFA results on the English IPIP-NEO-120 with RMSEA index = .06, and TLI and CLI $>$.90 (Kajonius & Johnson, 2019). The results also showed that all facets had an index range of .46-1.02 (SLF $>$ |.30|). It can be said that all facets are valid in forming the dimensional structure of Neuroticism. At the item level, all items are valid except Item 4 (n. 103) on the Depression facet and Item 1 (n. 21) on the Immoderation facet. As a comparison, in the Danish IPIP-NEO-120 structure, three of the four items on the Immoderation facet loaded more on other factors but the pattern cannot be known with certainty (Hong et al., 2019; Vedel et al., 2019). Meanwhile, the original IPIP-NEO-120 showed that the weak SLF item is found in the Anxiety and Vulnerability facets, but the Immoderation facet has a strong factor loading.

On the Extraversion dimension, the analysis results show that the fit values are $\chi^2/df = 2.88 < 5$; RMSEA = .06 $>$.05; GFI = .92 $>$.90; AGFI = .90 \approx .90; CFI = .91 $>$.90; TLI = .88 $<$.90. This result is similar to the results of the original IPIP-NEO-120 which shows the value of RMSEA = .06, and TLI and CLI $>$.90 (Kajonius & Johnson, 2019). It can be said that both the adaptation and the original have the same fit model. This study shows that some items are considered invalid: two items (79 and 109) on the Activity level facet and two items (84 and 115) on the Excitement seeking facet. In the analysis of The Danish IPIP-NEO-120, there are also two items from Assertiveness and Activity level facets of Extraversion that have weak factor loading. In the original IPIP-NEO-120, some items of the facets of this dimension also have substantial cross-loadings (Vedel et al., 2019).

The results on the Openness to experience dimension show that the fit values are $\chi^2/df = 2.38 < 5$; RMSEA = .05 \approx .05; GFI = .93 $>$.90; AGFI = .91 $>$.90; CFI = .88 $<$.90; TLI = .85 $<$.90. When compared with the original English NEO-IPIP-120 which has RMSEA = .05, TLI = .91, and CFI = .92, it can be said that both models have a fit model. In the Openness to experience dimension, there is one invalid facet, namely Imagination. Meanwhile, the invalid items are three (34, 64, and 95) on the Imagination facet, one (13) on the Emotionality facet, one (111) on Adventurous, and three (28, 59, and 90) on the Liberalism facet. In the

Danish version, this dimension also shows many items that have a weak loading factor compared to items in other dimensions. Although efforts have been made to change the context of the translation of invalid items, the analysis results still show a weak factor loading on the Libelism facet (Vedel et al., 2019).

Analysis of the Agreeableness dimension showed a value of $\chi^2/df = 2.31 < 5$; RMSEA = .05 \approx .05; GFI = .93 > .90; AGFI = .91 > .90; CFI = .91 > .90; TLI = .88 < .90. This is in line when compared to the result of the structural analysis of the English NEO-IPIP-120 (RMSEA = .05; TLI = .91; CFI = .93). Modesty becomes the only invalid facet in this dimension. While at the item level, there are three invalid items on this dimension: one item on the Cooperation facet (112), one item on the Modesty facet (117), and one item on the Sympathy facet (91). This is different from what was found in the Danish version of the IPIP-NEO-120 test which showed three items with a weak factor loading on the Morality facet.

As in the Agreeableness dimension, the Conscientiousness dimension showed all the required fit indices ($\chi^2/df = 2.82 < 5$; RMSEA = .05 \approx .05; GFI = .92 > .90; AGFI = .90 \approx .90; CFI = .92 > .90; TLI = .91 > .90). These results are in line with the IPIP-NEO-120 original version, which showed fit indices for the Conscientiousness dimension (Kajonius & Johnson, 2019). Not only that, 24 items that represent facets of the Conscientiousness dimension also show good validity (SLF > |.30).

Internal consistency reliability on five dimensions was also analyzed in this study. Based on the results presented, the construct reliability value of the five dimensions is quite high, moving from .78 to .94, and Cronbach's alpha reliability moving from .65 to .90, in line with the reliability value in the original IPIP-NEO-120 where internal consistency reliability moved from .75 to .89 (Johnson, 2014), and also with the internal consistency reliability of the Danish IPIP-NEO-120 that has a reliability index from .75 to .89.

The results of the descriptive analysis in this study showed that the mean scores and standard deviation values on Neuroticism, Extraversion, Openness to experience, and Conscientiousness were very similar to those obtained in both the Danish IPIP-NEO-120 and the original IPIP-NEO-120. However, the mean score of Agreeableness in this study sample was lower ($M = 83.5$) than that of the Danish sample ($M = 97.0$), somewhat similar to the results of the sample used by Johnson (2014; $M = 87.8$). In addition, the standard deviation of the Agreeableness dimension was the lowest among the other dimensions ($SD = 5.9$), resembling the low SD value in the Danish sample ($SD = 8.7$) and the original IPIP-NEO-120 ($SD = 12.5$). It should be noted that the distribution of the sample in this study is mostly students and graduates who do not represent the entire sample population of Indonesia. Further research is needed on more general samples to determine whether there is a difference in results.

Gender differences are thought to influence the results of scores in personality tests. The results of this study indicate that there is a difference in mean scores between gender. The study included a sample of 67% females and 33% males where the gender distribution is similar to the research sample conducted by Johnson on the original IPIP-NEO-120 (females = 60%, males = 40%). This study showed that there was a difference in average scores between females and males on the dimensions of Neuroticism ($p < .001$), Extraversion ($p < .05$), and Conscientiousness ($p < .01$), and no significant difference was found in the dimensions of Agreeableness and Openness to experience. The results shown in this study are quite consistent with previous findings, where female scores higher than male scores on Neuroticism (Costa et al., 2001; Lynn & Martin, 1997; Weisberg et al., 2011). Meanwhile, previous studies showed that females tend to score higher than males on Extraversion (Costa et al., 2001; Feingold, 1994), this study showed conversely. In contrast to some studies' results where females score somewhat higher than males on the Conscientiousness dimension (Costa et al., 2001; Feingold, 1994), this study showed that male scores are higher than female scores. However, these results are not consistent across cultures, and no significant gender differences are typically found

in Conscientiousness (Costa et al., 2001). These findings also clarify the nature of gender differences in personality and highlight the utility of measuring personality at the aspect level (Weisberg et al., 2011).

CONCLUSION

IPIP-NEO-120 as one of the personality inventories has been used for different purposes and has been adapted to many languages. Regarding its use in Indonesia, IPIP-NEO-120 is considered a motivating alternative inventory scale because it gives practitioners and researchers a chance to measure personality using a comprehensive 30-facet inventory with fewer items. The results show that all facets of The Indonesian IPIP-NEO-120 fit the 1-factor model so that the multifactor model theorized according to the FFM principle can be accepted. From this study, it can be seen that the majority of items were found to be valid in measuring their respective dimensions. It is also worth noting that the reliability test produced valid results, with the construct reliability value exceeding .70. It can be concluded that the instrument overall met the criteria of good items, namely (1) having a positive factor load, (2) being valid (significant, p -value < .05), and (3) having correlations among measurement errors that are not more than 3 or, in other words, are unidimensional. Internal consistency reliability of the five dimensions also has high scores. For this reason, The Indonesian IPIP-NEO-120 personality inventory can be considered suitable to be used as an inventory. However, providing evidence for the validity of the internal structure is an ongoing accumulative process and is never completed through the publication of a single study. It is recommended that further analysis may be required for certain items. This study found that some items were identified as potentially problematic. Although previous research on original IPIP-NEO-120 showed that acquiescence response bias possibly does not distort the factor structure, this analysis might provide more information regarding the low loading factor on some items.

REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. American Psychological Association.
- Arbuckle, J. L. (2005). *Amos™ 6.0 user's guide*. Amos Development Corporation.
- Barrick, M. R., & Ryan, M. R. (2003). *Personality and work. Reconsidering the role of personality in organizations*. Jossey Bass.
- Baumgartner, H. R., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13(2), 139-161. [https://doi.org/10.1016/0167-8116\(95\)00038-0](https://doi.org/10.1016/0167-8116(95)00038-0)
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186-3191. <https://doi.org/10.1097/00007632-200012150-00014>
- Berry, D. T. R., Wetter, M. W., Baer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991). Detection of random responding on the MMPI-2: Utility of F , back F , and VRIN scales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 3(3), 418-423. <https://doi.org/10.1037/1040-3590.3.3.418>
- Blanken, T., Dekker, K., & van Someren, E. (2018). *How personality profile similarity can improve comparability between assessment formats: An example of the Mini-IPIP and IPIP-NEO-120 in a Dutch community sample*. PsyArXiv. <http://doi.org/10.31234/osf.io/pjtgx>
- Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, 11(5), 515-524. [https://doi.org/10.1016/0191-8869\(90\)90065-Y](https://doi.org/10.1016/0191-8869(90)90065-Y)

- Brown, T. (2015). *Confirmatory factor analysis for applied research* (II ed.). Guilford Press.
- Byrne, B. M. (1998). *Structural equation modeling with Lisrel, Preliis, and Simplis. Basic concepts, applications, and programming*. Taylor & Francis Group.
- Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30*(5), 555-574. <https://doi.org/10.1177/0022022199030005001>
- Carlback, J., & Wong, A. (2018). A study on factors influencing acceptance of using mobile electronic identification applications in Sweden.
Retrieved from <http://www.diva-portal.org/smash/get/diva2:1214313/FULLTEXT01.pdf>
- Costa, P. T., & McCrae, R. R. (1976). Age differences in personality structure: A cluster analytic approach. *Journal of Gerontology, 31*(5), 564-570. <https://doi.org/10.1093/geronj/31.5.564>
- 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. <https://doi.org/10.4135/9781849200479.n9>
- Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322-331. <https://doi.org/10.1037/0022-3514.81.2.322>
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 56-83. <https://doi.org/10.1080/10705519909540119>
- Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429-456. <https://doi.org/10.1037/0033-2909.116.3.429>
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304-312. <https://doi.org/10.1037/1040-3590.6.4.304>
- Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141-165). Sage.
- Goldberg, L. R. (1990). An alternative "Description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216-1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg University Press.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96. <https://doi.org/10.1016/j.jrp.2005.08.007>
- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O' Cathain, A., Rousseau, M. C., & Vedel, I. (2019). Improving the content validity of the mixed methods appraisal tool: A modified e-Delphi study. *Journal of Clinical Epidemiology, 111*, 49-59. <https://doi.org/10.1016/j.jclinepi.2019.03.008>
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332-346. <https://doi.org/10.1177/1088868310361240>
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). Guilford Press.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102-138). Guilford Press.
- Johnson, J. A. (2014). Measuring thirty facets of the Five Factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78-89. <https://doi.org/10.1016/j.jrp.2014.05.003>
- Kajonius, P. J., & Johnson, J. A. (2019). Assessing the structure of the Five Factor Model of personality (IPIP-NEO-120) in the public domain. *Europe's Journal of Psychology, 15*(2), 260-275. <https://doi.org/10.5964/ejop.v15i2.1671>
- International Test Commission. (2010). International Test Commission guidelines for translating and adapting tests. <http://www.intestcom.org>
- Lynn, R., & Martin, T. (1997). Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *Journal of Social Psychology, 137*(3), 369-373. <https://doi.org/10.1080/00224549709595447>
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

-
- Marusic, I., Bratko, D., & Eterović, H. (1996). A contribution to the cross-cultural replicability of the five-factor personality model. *Review of Psychology*, 3(1-2), 23-35.
- McCrae, R. R. (1990). Traits and trait names: How well is Openness represented in natural languages? *European Journal of Personality*, 4(2), 119-229. <https://doi.org/10.1002/per.2410040205>
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11. <https://doi.org/10.3102/0013189X018002005>
- Renner, W., Rainer, A., Menschik-Bendele, J., & Deakin, P. (2014). Does the Myers-Briggs Type Indicator (R) measure anything beyond the NEO five factor inventory. *Journal of Psychological Type*, 74(1).
- Stone, B. M. (2021). The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.783226>
- Suwartono, C., Amiseso, C. P., & Handoyo, R. T. (2017). Uji Reliabilitas dan Validitas Eksternal [Reliability and external validity test. The Raven's standard progressive matrices]. *Humanitas: Indonesian Psychological Journal*, 14(1). <https://doi.org/10.26555/humanitas.v14i1.5772>
- Thompson, B. (2006). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association. <https://doi.org/10.1037/10694-000>
- van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology*, 33(2), 141-156. <https://doi.org/10.1177/0022022102033002002>
- Vedel, A., Gøtzsche-Astrup, O., & Holm, P. (2019). The Danish IPIP-NEO-120: A free, validated five-factor measure of personality. *Nordic Psychology*, 71(1), 62-77. <https://doi.org/10.1080/19012276.2018.1470553>
- Weisberg, Y. J., De Young, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, 2. <https://doi.org/10.3389/fpsyg.2011.00178>
- Yamaoka, L. (2014). *The effects of adherence to Asian values and extraversion on cardiovascular reactivity: A comparison between Asian and European Americans* [Pitzer senior thesis, paper 56]. Pitzer Student Scholarship at Scholarship @ Claremont. http://scholarship.claremont.edu/pitzer_theses/56
-