Primenjena psihologija
Vol. 16, No. 1, pp. 3-28, 2023

Research Article

A Rasch analysis of the International Personality Item Pool Big Five Markers Questionnaire: Is longer better?

Hanif Akhtar [1,2] ✉ and Bambang Sumintono[3]

[1] *Faculty of Psychology, Universitas Muhammadiyah Malang*
[2] *Doctoral School of Psychology, ELTE Eötvös Loránd University*
[3] *Faculty of Education, Universitas Islam Internasional Indonesia*

ABSTRACT

The 50-item International Personality Item Pool version of the Big Five Markers (IPIP-BFM) is an open-source and widely used measure of the big five personality traits. A short version of this measure (IPIP-BFM-25) has been developed using the classical test theory approach. No study was performed to examine the psychometric properties of a longer and shorter version of IPIP-BFM Indonesia using modern test theory. This study aimed to evaluate the psychometric properties of the Indonesian version of IPIP-BFM as well as IPIP-BFM-25 using Rasch analysis. The analysis was conducted in order to test their dimensionality, rating scale functioning, item properties, person responses, targeting, reliability, and item bias on 1003 Indonesian samples. The findings showed that both IPIP-BFM and IPIP-BFM-25 Indonesia have some adequate psychometric properties, especially regarding category function, item properties, reliability, and item bias. However, the emotional stability and intellect scales did not meet the assumption of unidimensionality, and all items on the scales were too easy to endorse by participants. In general, longer measures outperformed shorter measures in terms of person separation and reliability. Further testing and refinement must be conducted.

Keywords: Five-Factor model, IPIP Big Five Markers, Rasch analysis, DIF

UDC: 159.923.072
DOI: 10.19090/pp.v16i1.2401
Received: 11.04.2022.
Revised: 16.06.2022.
Accepted: 18.07.2022.

✉ Corresponding author email: hanifakhtar@umm.ac.id

Introduction

The Big Five personality trait is the most widely recognized personality model in psychology to date. This model explains that there are five main factors in an individual's personality; namely extraversion, agreeableness, conscientiousness, emotional stability, and intellect (Goldberg, 1992). Since many researchers use the Big Five model as a predictor of several outcomes in their studies, many instruments have been developed, such as the Big Five Inventory (BFI; John & Srivastava, 1999), the NEO PI-R (Costa & McCrae, 1995), the Ten-Item Personality Inventory (TIPI; Gosling et al., 2003), and the Trait Descriptive Adjective (TDA; Goldberg, 1992).

One of the most popular instruments to measure The Big Five personality traits is the 50-item International Personality Item Pool representation of the Goldberg (1992) markers for the Big-Five factor structure, hereinafter referred to as IPIP-BFM. This instrument is available on the International Personality Item Pool (IPIP) website and is free to use by anyone (Goldberg et al., 2006). The main advantage of IPIP-BFM is that this scale is cost-free and widely used by researchers, making the research findings comparable to existing studies. This measure has been adapted and validated in numerous different countries, such as Croatia (Mlačić & Goldberg, 2007), Poland (Strus et al., 2017), Scotland (Gow, Whiteman, Pattie, & Deary, 2005), China (Zheng et al., 2008), New Zealand (Guenole & Chernyshenko, 2005), Portugal (Oliveira, 2017), and Indonesia (Akhtar & Azwar, 2018).

Akhtar and Azwar (2018) have adapted IPIP-BFM in Indonesian samples using forward-backward-translation methods. Moreover, they developed a short version of the measure, IPIP-BFM-25, using the classical test theory (CCT) approach. The items for IPIP-BFM-25 were selected from the parent measure to maximize loading on the primary factor and minimize the cross-loading factor from the exploratory factor analysis (EFA). The study indicated that the Indonesian version of both IPIP-BFM and IPIP-BFM-25 has adequate Cronbach's alpha (ranging from .70 to .86), satisfactory factorial validity, and has a high correlation with BFI. However, due to its intrinsic limits, the CTT cannot protect the IPIP-BFM from psychometric criticism. The CTT, for example, cannot determine the response category functioning of IPIP-BFM (see Kean et al., 2017; Petrillo et al., 2015). In addition, selecting items based on the loading factor only may narrow item content, restricting the breadth of the item content on the full scale (Smith et al., 2000). Kline (2000) also noted that the main drawback of using EFA is the tendency to select items that are essentially paraphrases of each other in order to form a factor based on correlational analysis. Although redundant in the meaning, those items will have a high correlation and thus will have high loading on a factor.

Given the extensive usage of IPIP-BFM, it is critical to evaluate its structural validity using modern test theory, which avoids many of the CTT's flaws (see Bond & Fox, 2015). Rasch rating scale model can be beneficial for several reasons. First, Rasch models can look at how people of various abilities respond to a set of IPIP-BFM items (Raykov & Marcoulides, 2015). Rasch analysis can reveal the relative endorsability of items using item-person maps, which display both items and persons on the same logit scale according to item difficulty estimates (Bond & Fox, 2015). Second, Rasch analysis can convert the ordinal scale data obtained using the IPIP-BFM into linear, interval scale data using the raw score-to-logit transformation (see Andrich & Marais, 2019). Third, the Rasch model enables the examination of the functioning of the ordered response categories. In this respect, it is assumed that the category thresholds will be arranged in value in the same order as the response categories (Adams et al., 2012). Fourth, Rasch model measurement analysis provides reliability figures for items in the measurement instrument and persons. Rasch model analysis uses the separation to measure not only the person reliability but also item reliability (Fisher, 1992). In order to achieve this aim, the current study employed Rasch analysis to evaluate the psychometric properties of the IPIP-BFM.

Previous studies aimed to validate the IPIP-BFM using the Rasch model were conducted by Apple and Neff (2012) in a Japanese sample. The results of their study indicated the possible existence of additional factors within the Intellect and Agreeableness factors, as well as additional item fit problems within each hypothesized construct. Moreover, emotional stability items had a moderate floor effect, indicating that some items for this construct may have been too difficult for participants to endorse. On the other hand, the Agreeableness item-person map suggests a strong ceiling effect, indicating that the items for this particular construct were too easily endorsable by the participants. These findings could be a reference point to examine the appropriateness of the American-developed five-factor model personality trait instrument for measuring an Indonesian population.

Currently, there is no study to analyze the IPIP-BFM using the Rasch model in Indonesia. Moreover, although a short version of IPIP-BFM has been performed well in EFA and Cronbach's alpha analysis, there is no evidence that the chosen items are not redundant. Considering the benefits of Rasch analysis, this study aims to fill the gap in analyzing the psychometric properties of the Indonesian IPIP-BFM using Rasch analysis. Thus, this study aims to test IPIP-BFM dimensionality, rating scale functioning, item properties, person responses, targeting, reliability, and differential item functioning (DIF) across gender using the Rasch model. Second, this study aims to compare the targeting and reliability of the 50-item and 25-item versions of IPIP-BFM Indonesia.

Methods

Participants

A total of 1019 participants participated in this study. An online survey was used for the data collection of this study. Participants were recruited during 2021 using various strategies, including advertisements on social media (Facebook, Instagram, and WhatsApp groups) and encouraging our colleagues to share the advertisements. After data screening to filter participants with careless responses (*e.g.*, responding '*Strongly Agree*' to all items despite reverse wording), the final dataset contains the data from 1003 participants. The total sample consisted of 409 (40.8%) males and 594 (59.2%) females aged 15 to 50 (M = 19.9, SD = 4.8). The education of the sample consisted of 334 (33.3%) junior high school, 526 (52.4%) senior high school, 97

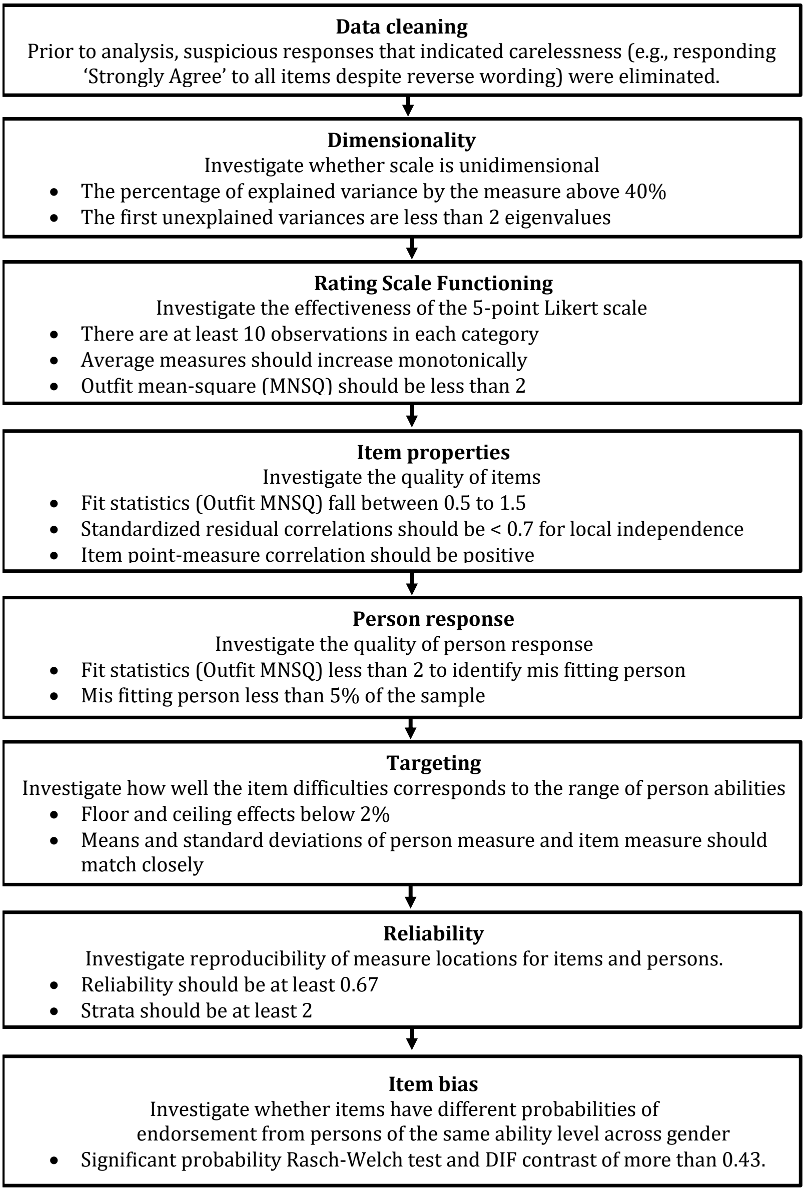
(9.6%) bachelor, 37 (3.6%) master, and 9 (0.8%) doctoral. On the first page of the questionnaire, it was stated that it was strictly anonymous and voluntary to address ethical concerns. Thus, by completing the questionnaire, the respondents have given their consent. All procedures have been approved by the ethical committee of the Faculty of Psychology, Universitas Muhammadiyah Malang.

Instrument

The Indonesian version of the IPIP-BFM was administered to all participants. The IPIP-BFM Indonesia contains 50 items that are used to measure five personality traits: extraversion, emotional stability, agreeableness, conscientiousness, and intellect. Each of the five personality traits is evaluated using a 10-item response, with each item rated using a 5-point Likert rating scale, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). IPIP-BFM-25 was a short version of IPIP-BFM that contains 25 items. The items have been adapted and validated into Indonesian by Akhtar and Azwar (2018) using the forward-backward-translation method.

Analysis procedures

The Rasch model analysis was performed using the Rasch Rating Scale Model (RSM), an extension of the Rasch model for polytomous items (Andrich, 1978; Andrich & Marais, 2019) in Winsteps 3.73 computer software (Linacre, 2012). The analysis was conducted for each dimension separately. Following Lim and colleagues' (2009) recommendations, this study conducted seven key Rasch evaluations for validity evidence: dimensionality, rating scale functioning, item properties, person responses, targeting, reliability, and item bias. The data analysis processes and a summary of the objectives of each process are shown in Figure 1. The scale level properties (dimensionality, targeting, and reliability) of IPIP-BFM-25 were compared. The analysis for IPIP-BFM-25 was conducted using the same dataset but only involving items from IPIP-BFM-25.



Data cleaning

Prior to analysis, suspicious responses that indicated carelessness (e.g., responding 'Strongly Agree' to all items despite reverse wording) were eliminated.

Dimensionality

Investigate whether scale is unidimensional

- The percentage of explained variance by the measure above 40%
- The first unexplained variances are less than 2 eigenvalues

Rating Scale Functioning

Investigate the effectiveness of the 5-point Likert scale

- There are at least 10 observations in each category
- Average measures should increase monotonically
- Outfit mean-square (MNSQ) should be less than 2

Item properties

Investigate the quality of items

- Fit statistics (Outfit MNSQ) fall between 0.5 to 1.5
- Standardized residual correlations should be < 0.7 for local independence
- Item point-measure correlation should be positive

Person response

Investigate the quality of person response

- Fit statistics (Outfit MNSQ) less than 2 to identify mis fitting person
- Mis fitting person less than 5% of the sample

Targeting

Investigate how well the item difficulties corresponds to the range of person abilities

- Floor and ceiling effects below 2%
- Means and standard deviations of person measure and item measure should match closely

Reliability

Investigate reproducibility of measure locations for items and persons.

- Reliability should be at least 0.67
- Strata should be at least 2

Item bias

Investigate whether items have different probabilities of endorsement from persons of the same ability level across gender

- Significant probability Rasch-Welch test and DIF contrast of more than 0.43.

Figure 1. Data analyses flow

Results

Dimensionality

Dimensionality analysis investigated whether all scales are unidimensional. The unidimensionality was investigated using the Principal Components Analysis of Rasch measures and residuals. The scale is fundamentally unidimensional if the percentage of explained variance by the measure is greater than 40%, and the first unexplained variances are less than 2 eigenvalues (Linacre, 2012).

For the IPIP-BFM, the analysis showed that the extraversion, agreeableness, and conscientiousness scales met the assumptions of unidimensionality. However, the emotional stability and intellect scales had eigenvalues of unexplained variance larger than 2. The raw variance explained by measures for extraversion, conscientiousness, agreeableness, emotional stability, and intellect was 48.7%, 41.6%, 45.3%, 49.3%, and 41.0%, respectively. The unexplained variances in the first contrast for extraversion, conscientiousness, agreeableness, emotional stability, and intellect were 1.9, 1.7, 1.9, 2.1, and 2.1, respectively.

For the IPIP-BFM-25, the analysis showed better unidimensionality with the extraversion, agreeableness, conscientiousness, and intellect scales met the assumptions of unidimensionality, but not for the emotional stability scales. The raw variance explained by measures for extraversion, conscientiousness, agreeableness, emotional stability, and intellect was 53.6%, 47.0%, 55.4%, 54.4%, and 49.0%, respectively. The unexplained variances in the first contrast for extraversion, conscientiousness, agreeableness, emotional stability, and intellect were 1.6, 1.6, 1.6, 2.1, and 1.8, respectively.

Table 1

Rating Scale Model category statistics for the total sample (N=1,003)

IPIP-BFM dimension	Category	Frequency	Percentage	Average measure	Outfit MNSQ	Andrich threshold
Extraversion	1 (SD)	571	6%	-1.87	1.13	NONE
	2 (D)	1916	19%	-.094	0.97	-2.60
	3 (N)	3824	38%	-0.02	0.87	-1.15
	4 (A)	3044	30%	0.97	1.01	0.68
	5 (SA)	675	7%	2.13	1.09	3.07
Agreeableness	1 (SD)	96	1%	-0.57	1.85	NONE
	2 (D)	512	5%	-0.16	1.20	-2.37
	3 (N)	2536	25%	0.41	0.84	-1.47
	4 (A)	5295	53%	1.61	0.89	0.29
	5 (SA)	1591	16%	3.13	1.02	3.55
Conscientiousness	1 (SD)	215	2%	-1.08	1.54	NONE
	2 (D)	1198	12%	-0.45	1.11	-2.64
	3 (N)	3348	33%	0.32	0.84	-1.06
	4 (A)	3974	40%	1.36	0.89	0.67
	5 (SA)	1295	13%	2.45	1.02	3.02
Emotional stability	1 (SD)	861	9%	-2.45	1.05	NONE
	2 (D)	3118	31%	-1.14	0.97	-3.09
	3 (N)	3359	33%	-0.12	0.84	-0.70
	4 (A)	2349	23%	0.77	1.04	0.66
	5 (SA)	343	3%	1.51	1.27	3.12
Intellect	1 (SD)	133	1%	-0.73	1.26	NONE
	2 (D)	1149	12%	-0.27	1.11	-2.83
	3 (N)	3894	39%	0.31	0.87	-1.20
	4 (A)	3742	37%	1.36	0.91	0.86
	5 (SA)	1112	11%	2.51	1.04	3.16

Note. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree

Rating scale functioning

Rating scale functioning investigated the effectiveness of the 5-point Likert scale (Van Zile-Tamsen, 2017) for each dimension. Several essential criteria suggested by Linacre (2002b, 2004) to diagnose the effectiveness of the rating scale used are a) there are at least ten observations in each category; b) average measures should increase monotonically; c) outfit mean-square (MNSQ) should be less than 2.

Results illustrated that there were no disordered thresholds. At least 10 individuals chose each category in the five scales of IPIP-BFM Indonesia. The average measure by category moved up monotonically with the rating scale. All categories had outfit mean squares of less than 2.0 in each step. Hence, all the Likert scale categories were well functioning and fully understood by respondents. The rating scale model category is presented in Table 1.

Item properties

Item properties analysis examines the quality of items. Item fit is considered good if the fit statistics for outfit MNSQ fall between 0.5 and 1.5 (Linacre, 2002a). Standardized residual correlations represent item local dependence, with correlations greater than 0.7 indicating that two items share more than half of their random variance and only one has to be retained (Linacre, 2012). Meanwhile, the item point-measure correlation index should be positive, indicating no polarity. Negative correlations indicate reverse-coded item miskeying, whereas near-zero correlations indicate items that are very easy or difficult to endorse or that measure a different construct. A correlation of less than 0.4 can be used to identify items for wording investigation (Wolfe & Smith, 2007).

Out of the 50 items of IPIP-BFM, 49 had outfit mean squares that fell between 0.5 and 1.5, and one had outfit mean squares above 1.5. Furthermore, all of the items had point-measure correlations greater than 0.4. Item I8 was the most difficult item to endorse, and item A3 was the easiest item to endorse among all items. The complete item fit information for the five scales of IPIP-BFM is shown in Table 3. The largest standardized residual correlations ranged from -.30 to .22 for extraversion, from -.32 to .11 for agreeableness, from -.33 to .19 for conscientiousness, from -.30 to .48 for emotional stability, and from -.36 to .24 for intellect, which indicated that the items could be viewed as locally independent.

Table 2

Item properties for the five scales of IPIP-BFM

Item	Measure	SE	Outfit MNSQ	PTME
Extraversion				
E1 Am the life of the party	-0.53	0.05	0.80	0.67
E2 Don't talk a lot	-0.11	0.04	0.95	0.70
E3 Feel comfortable around people	-0.52	0.05	0.96	0.64
E4 Keep in the background	0.63	0.04	1.27	0.62
E5 Start conversations	-0.66	0.05	0.91	0.65
E6 Have little to say	-0.05	0.04	0.96	0.73
E7 Talk to a lot of different people at parties	-0.34	0.05	0.75	0.76
E8 Don't like to draw attention to myself	0.69	0.04	0.98	0.66
E9 Don't mind being the centre of attention	0.45	0.04	1.29	0.59
E10 Am quiet around strangers	0.44	0.04	1.15	0.68
Agreeableness				
A1 Feel little concern for others	0.49	0.05	1.37	0.62
A2 Am interested in people	-0.64	0.06	0.74	0.68
A3 Insult people	-0.92	0.06	1.19	0.54
A4 Sympathise with others' feelings	-0.60	0.06	0.78	0.64
A5 Am not interested in other people's problems	0.56	0.05	1.10	0.62
A6 Have a soft heart	0.64	0.05	1.35	0.59
A7 Am not really interested in others	0.32	0.05	0.98	0.63
A8 Take time out for others	0.29	0.05	0.98	0.55
A9 Feel others' emotions	-0.18	0.05	0.78	0.63
A10 Make people feel at ease	0.04	0.05	0.81	0.66
Conscientiousness				
C1 Am always prepared	-0.50	0.05	0.87	0.67
C2 Leave my belongings around	-0.03	0.05	1.38	0.62

C3 Pay attention to details	0.00	0.05	0.91	0.67
C4 Make a mess of things	-0.37	0.05	0.98	0.66
C5 Get chores done right away	0.80	0.04	0.97	0.65
C6 Often forget to put things back in their proper place	0.52	0.04	1.36	0.63
C7 Like order	-0.17	0.05	0.91	0.69
C8 Shirk my duties	-0.43	0.05	1.16	0.60
C9 Follow a schedule	-0.04	0.05	0.72	0.71
C10 Am exacting in my work	0.21	0.05	0.82	0.65
Emotional stability				
ES1 Get stressed out easily	-0.11	0.04	1.16	0.68
ES2 Am relaxed most of the time	-0.02	0.04	1.00	0.66
ES3 Worry about things	0.47	0.05	1.13	0.65
ES4 Seldom feel blue	0.05	0.04	0.90	0.71
ES5 Am easily disturbed	0.08	0.04	1.01	0.69
ES6 Get upset easily	-0.15	0.04	0.87	0.73
ES7 Change my mood a lot	0.47	0.05	0.89	0.71
ES8 Have frequent mood swings	0.03	0.04	0.89	0.75
ES9 Get irritated easily	-0.27	0.04	0.90	0.72
ES10 Often feel blue	-0.54	0.04	1.34	0.63
Intellect				
I1 Have a rich vocabulary	0.25	0.05	0.87	0.64
I2 Have difficulty understanding abstract ideas	0.73	0.05	1.02	0.64
I3 Have a vivid imagination	-0.56	0.05	1.09	0.64
I4 Am not interested in abstract ideas	0.20	0.05	1.03	0.61
I5 Have excellent ideas	-0.31	0.05	0.78	0.65
I6 Do not have a good imagination	-0.70	0.05	0.85	0.63
I7 Am quick to understand things	-0.08	0.05	0.96	0.60
I8 Use difficult words	*0.95*	*0.05*	*1.62*	*0.51*
I9 Spend time reflecting on things	-0.30	0.05	1.03	0.57
I10 Am full of ideas	-0.18	0.05	0.78	0.63

Note: italic item (I8) misfit for outfit MNSQ criteria. PTME = Point-measure correlation.

Table 3

Summary Statistics for Five Scales of IPIP-BFM

Construct	Item measure		Person measure		Person MNSQ Outfit		Item MNSQ Outfit		Item reliability	Person reliability	Item separation	Person Separation	Cronbach's alpha
	Mean	SD	Mean	SD	Mean	SD	Mean	SD					
IPIP-BFM													
Extraversion	0	0.49	0.16	1.33	1.00	0.78	1.00	0.17	0.99	0.85	10.55	2.34	0.87
Agreeableness	0	0.53	1.47	1.46	1.01	0.87	1.01	0.23	0.99	0.8	9.68	2.02	0.83
Conscientiousness	0	0.40	0.91	1.34	1.01	0.87	1.01	0.21	0.99	0.83	8.21	2.19	0.86
Emotional stability	0	0.29	-0.40	1.46	1.01	0.85	1.01	0.15	0.97	0.86	6.21	2.48	0.89
Intellect	0	0.51	0.89	1.26	1.00	0.8	1.00	0.23	0.99	0.8	10.36	2.00	0.82
IPIP-BFM-25													
Extraversion	0	0.47	0.43	1.61	1.00	0.92	1.00	0.19	0.99	0.76	9.15	1.76	0.77
Agreeableness	0	0.47	1.79	1.80	1.00	1.09	1.00	0.24	0.98	0.69	7.69	1.48	0.72
Conscientiousness	0	0.60	1.19	1.85	1.00	0.95	1.00	0.14	0.99	0.77	10.80	1.81	0.79
Emotional stability	0	0.34	-0.56	0.34	1.00	0.96	1.00	0.12	0.98	0.78	6.87	1.90	0.80
Intellect	0	0.62	1.20	1.61	1.00	0.93	1.00	0.12	0.99	0.71	11.65	1.56	0.72

Person response

Person response analysis examines the quality of a person's response. Person response was indicated by person fit statistics. Smith and Wolfe (2007) proposed using an outfit MNSQ value of 2 to identify misfits, which is expected to be less than 5% of the sample. Person misfits may indicate low-quality self-report data, for example, due to careless responses or misunderstanding of items (Kottorp, et al., 2003).

Out of the 1,003 participants, 911 (90.82%) reported an acceptable fit on the extraversion scale, 887 (88.43%) on the agreeableness scale, 898 (89.53%) on the conscientiousness scale, 899 (89.63%) on the emotional stability scale, and 896 (89.33%) on the intellect scale. At all scales, there were approximately 10% of misfitting persons. Although the percentage of misfitting persons was noticeably higher than expected, the average mean person's outfit MNSQ was at the ideal of approximately 1 (Table 3). The nature of the anonymous online survey might explain the occurrence of careless responses.

Targeting

Item-person targeting examined how well the distribution of item difficulties corresponds to the range of persons' abilities (Linacre, 2012). Floor and ceiling effects below 1% are very good, and effects between 1% and 2% are good (Fisher, 2007). In addition, the means and standard deviations of the person and item measures should match closely (Bond & Fox, 2015).

For IPIP-BFM, two participants (0.19%) scored maximum scores on the extraversion scale, ten participants (0.99%) on the agreeableness scale, six participants (0.19%) on the conscientiousness scale, six participants (0.59%) on the emotional stability scale, and seven participants (0.69%) on the intellect scale. Less than 1% of participants scored the maximum score of the five scales, and none of the participants scored minimum scores on the five scales of IPIP-BFM, indicating very good ceiling and floor effects. The means and SDs of the person and item measure are presented in Table 3.

For IPIP-BFM-25, six participants (0.59%) scored maximum scores on the extraversion scale, seventeen participants (1.69%) on the agreeableness scale, nine participants (0.89%) on the conscientiousness scale, fifteen participants (1.49%) on the emotional stability scale, and fourteen participants (1,39%) on the intellect scale. Less than 2% of participants scored the maximum score of the five scales, and none of the participants scored minimum scores on the five scales of IPIP-BFM, indicating good ceiling and floor effects. The means and SDs of the person and item measure are presented in Table 3.

Item-person maps based on item difficulty measures were generated for each dimension to further examine the item-participant match. Figures 2 and 3 show the distribution of the person and item measures. The construct is laid out vertically, with the most difficult items at the bottom and the highest-abled person at the top. As shown in the figure, the difficulty of the items was lower than the levels of conscientiousness, agreeableness, and intellect of the participants, while the difficulty of the items was equal to the middle level of extraversion and emotional stability of the participants. Special attention should be directed to the agreeableness, conscientiousness, and intellect scales, as a majority of participants were above the range of the items, indicating that the items were too easy to endorse.

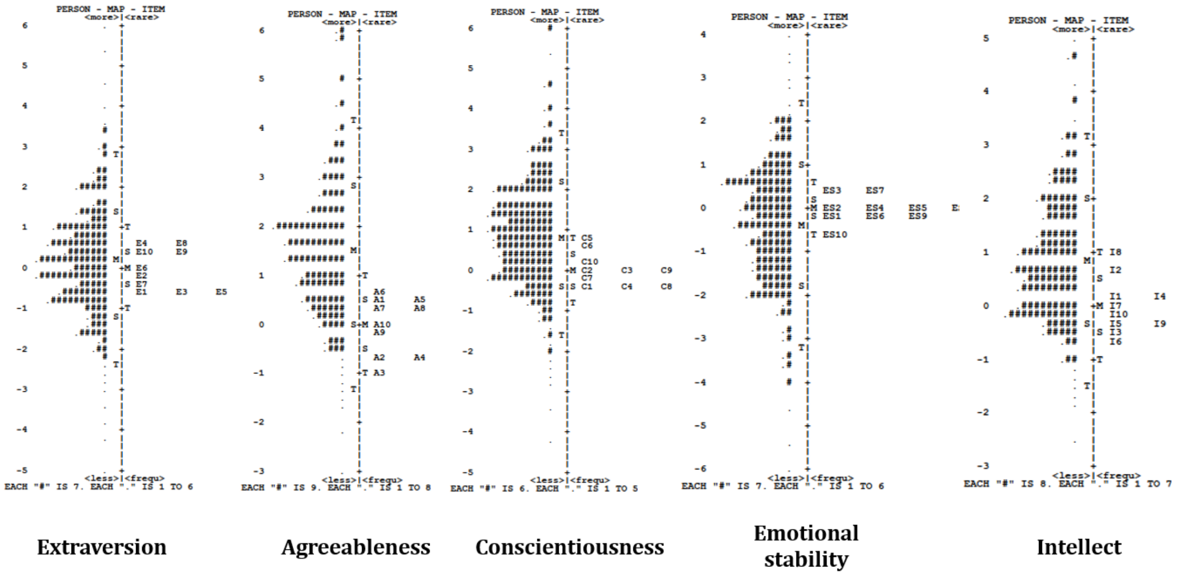


Figure 2. Wright map for all scale of IPIP-BFM

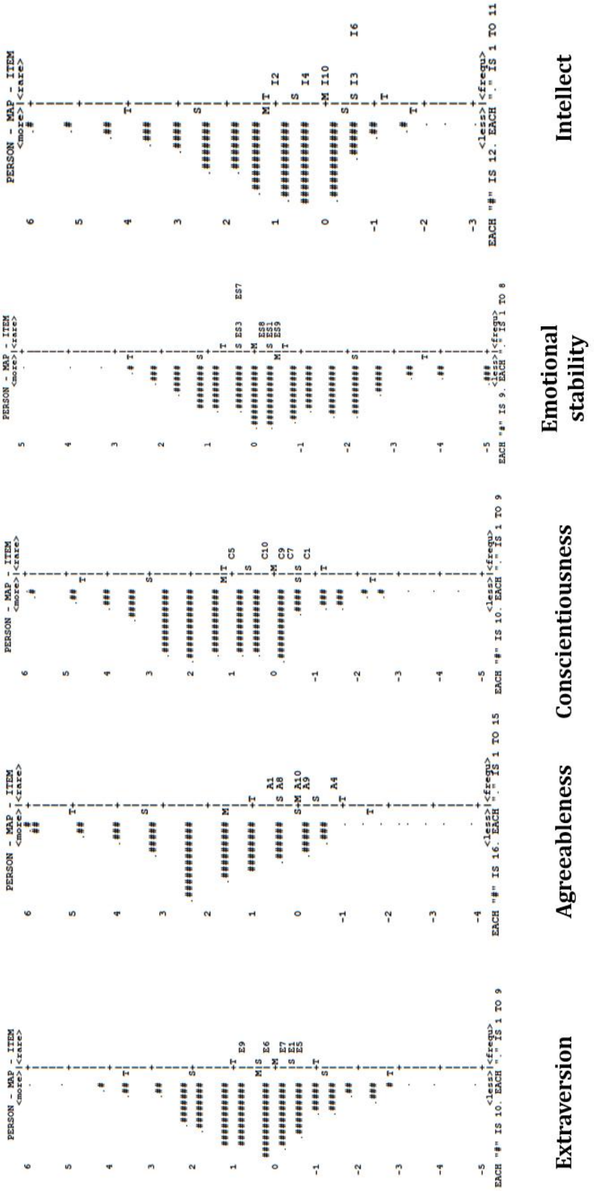


Figure 3. Wright map for all scale of IPIP-BFM-25

Reliability

Reliability analysis examined the reproducibility of measure locations for items and persons. We examined the reliability of the IPIP-BFM concerning the Rasch person separation reliability (Wright & Masters, 1982) and Cronbach's alpha. Fisher (2007) suggested that for fair quality, reliability and strata should be at least 0.67 and 2, and for good quality, they should be at least 0.81 and 3. The strata index is calculated as (4 S + 1)/3, where S refers to the person or item separation index (Wright & Masters, 1982).

All five scales in IPIP-BFM had satisfactory items, persons, and alpha reliability indices above 0.8 (Table 2). The item separation for all scales above six, and person separation for all scales above two, results in the item and person strata above three. It indicates that all scales have good quality (Fisher, 2007). The sample is large enough to confirm the item's hierarchical difficulty in the instrument, and all scales are sensitive enough to differentiate various levels of a person's abilities.

For the short scale, the downgrade of the measured quality was indicated by the lower person reliability and person separation. The scales are not sensitive enough to differentiate various levels of a person's abilities. A separation index implies that the scales can consistently identify only two levels of person's abilities. However, all five scales have a fair quality of reliability.

Item bias

Lastly, an item bias was tested using differential item functioning (DIF) by gender and education to test the generalized validity. According to ETS guidelines, a slight to moderate DIF was regarded to be present if the difficulty parameters had a significant probability and a DIF contrast of more than 0.43 (Zwick et al., 1999).

Overall, the differences in the item difficulty of the males and females were small. The highest contrast was -0.42 logits for item E6. However, none of the items with DIF contrast more than 0.43, indicating that females and males attached similar meanings to the items of IPIP-BFM. The differences in the item difficulty of people with secondary school and higher education were slightly higher. The highest contrast was -1.01 logits for item A3. Eight items had a DIF contrast of more than 0.43 (Table 4), indicating that people with secondary school and higher education attached different meanings to some items of IPIP-BFM.

Table 4

Item bias for the five scales of IPIP-BFM

Item	Gender		Education	
	DIF Contrast[a]	Rasch-Welch t[a]	DIF Contrast[b]	Rasch-Welch t[b]
E4 Keep in the background	0.00	0.00	0.77	6.07*
A3 Insult people	-0.07	-0.63	-1.01	-5.84*
A5 Am not interested in other people's problems	0.02	0.23	0.68	5.06*
A6 Have a soft heart	0.08	0.84	0.61	4.58*
I1 Have a rich vocabulary	-0.11	-1.20	-0.57	-4.16*
I7 Am quick to understand things	0.00	0.00	-0.59	-4.26*
I8 Use difficult words	0.15	1.62	0.91	7.04*
I9 Spend time reflecting on things	-0.18	-1.87	-0.46	-3.27*

Notes: DIF contrast[a] = the difference in the difficulty of the item between males and females. A negative DIF contrast[a] indicates that the item is more difficult for females. DIF contrast[b] = the difference in the difficulty of the item between people with secondary school and higher education. A negative DIF contrast[b] indicates that the item is more difficult for people in secondary school.

* *p* < .05

Discussion

In general, our findings show that the IPIP-BFM Indonesia has some adequate psychometric properties, especially in terms of category function, item properties, and item bias. The use of five-point Likert scale categories was well functioning and fully understood by participants. No item was identified as DIF across gender. This indicates that females and males attached similar meanings to the items of IPIP-BFM. Therefore, comparing the big five personalities between genders can be conducted fairly using the items in this IPIP-BFM. However, some items were detected to have DIF by education levels. If personality traits are compared across education levels for some reason, the conclusion should be made cautious since several items might be biased.

In terms of scale-level psychometric properties, the longer version of the test outperforms the shorter one. As predicted, IPIP-BFM has higher reliability, separation, and better item targeting than IPIP-BFM-25. However, for research purposes, IPIP-BFM-25 provides adequate reliability as all scales have reliability above 0.70. Therefore, when the resource is limited, and a shorter measure should be used, then IPIP-BFM-25 provides an adequate measure of the big five personality traits. For example, researchers who wish to place personality, not as the primary variable of interest are suggested to use IPIP-BFM-25.

In terms of unidimensionality, the emotional stability and intellect scales did not meet the assumption of unidimensionality. This finding is similar to Apple and Neff (2012) findings. The shorter scale has better unidimensionality in this case. This is not surprising since personality is a complex construct with many facets (Cooper et al, 2010). Selecting items with high loading and low cross-loading factor would ensure unidimensionality, but it could narrow the content and range of endorsability. However, these reports are better than the Japanese version of IPIP-BFM, especially in terms of reliability and item properties (Apple & Neff, 2012). The explanation can be addressed in the adaptation process since Apple and Neff (2012) changed the original version of five-point Likert-type categories into four response categories by omitting the middle category response.

In general, all items in extraversion, agreeableness, conscientiousness, and intellect scales were too easy to endorse by participants. The inability of the scale items to cover participants with maximum scores of the constructs in our findings is in line with Apple and Neff (2012) findings. This is not surprising because the instrument was developed based on a CTT approach, which has two major problems: having sets of redundant items and having skewed response categories for most items (Petrillo, et al., 2015). This problem seems more severe on a shorter scale.

Since the items for the short scale were selected based on the loading factor, some items are redundant for the content and the item difficulty. The present research suggests that adding more difficult items to the scales could enhance their measurement precision. Specifically for short scale, items selected should have a broader range of endorsability, which may improve future psychological research involving Indonesian participants.

The mean for a person measure on the agreeableness scale was the highest among the five scales in IPIP-BFM and IPIP-BFM-25, indicating that the item on the agreeableness scale is too easy to be endorsed by Indonesian samples. This result could be explained by previous literature in cultural psychology, where it is popularly termed the "interdependent self" (Markus & Kitayama, 1998). Eastern societies (such as Indonesia) are considered more group-oriented than Western societies, which are more individualistic (Hofstede & McCrae, 2004; Kitayama, Markus, & Kurokawa, 2000). It may be like Indonesians to consider others’ feelings in their daily life, which affects the participants' responses, as such in the current instrument.

One limitation of the study is that participants were recruited using an online survey, which means they may not be representative of the Indonesian population. The internet has not yet reached several regions of Indonesia. It is possible that Indonesians from rural areas would interpret the IPIP-BFM content differently than expected. This must be empirically examined. A second possible limitation is that our study solely focused on the internal psychometric characteristics of the IPIP-BFM. While these findings are promising, the external validity measure could address to what degree this measure corresponds to any theoretically related constructs. A third limitation is that the investigation of IPIP-BFM-25 used the same dataset as the IPIP-BFM. Smith and colleagues (2000) suggested administering the short form on an independent sample to avoid overestimating the correlation. However, the procedure used in this study is considered enough to describe the reliability and separation loss of using the short form.

Conclusion

IPIP-BFM Indonesia has some adequate psychometric properties, especially in terms of category function, person and item reliability, and item properties. In general, longer measures outperformed shorter measures regarding person separation and reliability. However, further testing and refinement must be developed. Based on the Rasch model analysis in this study, the following suggestions can be made to improve the test's precision. First, more difficult items should be added to represent a wider variety of endorsement abilities. Second, the item detected as a misfit (I8 Use difficult words) should either be revised or deleted. Third, if a short version is needed, then the item selection for the short scale should not solely be chosen based on the loading factor on EFA but also considering the broad of the content and item difficulty.

Authors note

HA performed study design, data collection, statistical analysis, data interpretation, manuscript preparation, literature search, and funds collection. BS performed statistical analysis, data interpretation, manuscript preparation, and literature search.

Conflict of Interests

The authors declare that there is no conflict of interest.

Data availability statement

The datasets presented in this study can be found in online repositories at https://osf.io/uca3p/.

Funding

H.A. received funding from the Faculty of Psychology, University of Muhammadiyah Malang, Indonesia: Grant no. E.2.b/749.a/F.Psi-UMM/VIII/2021

References

Adams, R. J., Wu, M. W., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. https://doi.org/10.1177/0013164411432166

Akhtar, H., & Azwar, S. (2018). Development and validation of a short scale for measuring big five personality traits: The IPIP-BFM-25 Indonesia. *Journal of Innovation in Psychology, Education and Didactics, 22*(2), 167–174.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*(3), 665–680. https://doi.org/10.1177/001316447803800308

Andrich, D. & Marais, I. (2019*). A Course in Rasch Measurement Theory, measuring in the educational, social and health sciences*. Singapore: Springer.

Apple, M. T., & Neff, P. (2012). Using Rasch Measurement to Validate the Big Five Factor Marker Questionnaire for a Japanese University Population. *Journal of Applied Measurement*, *13*(3), 1–17.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Third edition). New York ; London: Routledge, Taylor and Francis Group.

Costa Jr, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*(1), 21–50. https://doi.org/10.1207/s15327752jpa6401_2

Fisher, W. P., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions*, 6, 238

Fisher, W.P. Jr. (2007). Rating scale instrument quality criteria. *Rasch measurement transactions*, 21(1), 1095. (Downloaded from http://www.rasch.org/rmt/rmt211m.htm)

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

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

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's 'IPIP' Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*(2), 317–329. https://doi.org/10.1016/j.paid.2005.01.011

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Guenole, N., & Chernyshenko, O. S. (2005). The Suitability of Goldberg's Big Five IPIP Personality Markers in New Zealand: A Dimensionality, Bias, and Criterion Validity Evaluation. *New Zealand Journal of Psychology, 34*(2), 86–96.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*(1999), 102–138.

Kean, J., Bisson, E. F., Brodke, D. S., Biber, J., Gross, P. H. (2017). An introduction to item response theory and Rasch analysis: *Application using the Eating Assessment Tool (EAT-10). Brain Impairment, 19*(1), 91–102. https://doi.org/10.1017/BrImp.2017.31

Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, Emotion, and Well-being: Good Feelings in Japan and the United States. *Cognition & Emotion*, *14*(1), 93–124. https://doi.org/10.1080/026999300379003

Kline, P. (2000). The future of personality measurement. In J. Mohan (Ed.), *Personality across cultures: Recent developments and debates* (pp. 336-351). New Delhi, India: Oxford University Press.

Kottorp, A., Bernspång, B., & Fisher, A.G. (2003). Validity of a performance assessment of activities of daily living for people with developmental disabilities. *Journal of Intellectual Disability Research*, *47*(8), 597–605. https://doi.org/10.1046/j.1365-2788.2003.00475.x

Lim, S.M., Rodger, S., & Brown, T. (2009). Using Rasch analysis to establish the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation, 16*(5), 251–260. https://doi.org/10.12968/ijtr.2009.16.5.42102

Linacre, J. M. (2002a). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2002b). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2012). *A User's Guide to WINSTEP & MINISTEP: Rasch-Model Computer Programs*. Chicago: Winsteps.com.

Markus, H. R., & Kitayama, S. (1998). The cultural psychology of personality. *Journal of Cross-Cultural Psychology, 29*(1), 63–87. https://doi.org/10.1177/0022022198291004

Mlačić, B., & Goldberg, L. R. (2007). An Analysis of a Cross-Cultural Personality Inventory: The IPIP Big-Five Factor Markers in Croatia. *Journal of Personality Assessment*, *88*(2), 168–177. https://doi.org/10.1080/00223890701267993

Oliveira, J. P. (2017). Psychometric Properties of the Portuguese Version of the Mini-IPIP five-Factor Model Personality Scale. *Current Psychology, 38,* 432–439. https://doi.org/10.1007/s12144-017-9625-5

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Raykov, T., Marcoulides, G. A. (2015). On the relationship between classical test theory and item response theory. *Educational and Psychological Measurement, 76*(2), 325–338. https://doi.org/10.1177/0013164415576958

Smith, E. V., Jr., & Wolfe, E. W. (2007). Understanding Rasch measurement: Instrument development tools and activities for measure validation using Rasch models: Part 2–Validation activities. *Journal of Applied Measurement, 8*(2), 204–234.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12(*1), 102–111, https://doi.org/10.1037//1040-3590.12.1.102.

Strus, W., Cieciuch, J., & Rowiński, T. (2017). The Polish adaptation of the IPIP-BFM-50 questionnaire for measuring five personality traits in the lexical approach. *Roczniki Psychologiczne/Annals of Psychology*, *17*(2), 347–366.

Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education,* 58(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0

Wolfe, E.W., & Smith, E.V. (2007). Instrument development tools and activities for measure validation using Rasch models: Validation activities. *Journal of Applied Measurement, 8*(2), 204–234.

Wright, B., Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Zheng, L., Goldberg, L. R., Zheng, Y., Zhao, Y., Tang, Y., & Liu, L. (2008). Reliability and concurrent validation of the IPIP Big-Five factor markers in China: Consistencies in factor structure between Internet-obtained heterosexual and homosexual samples. *Personality and Individual Differences*, *45*(7), 649–654. https://doi.org/10.1016/j.paid.2008.07.009

Zwick, R., Thayer, D.T., Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36(*1), 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x