INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 11, Issue 9, (2021) E-ISSN: 2222-6990

An Initial Analysis of Reliability and Validity of a Personality Instrument Using the Rasch Measurement Model

Noorashikeen Mohamed, Wan Shahrazad Wan Sulaiman, Fatimah wati Halim, Mohd Saidfudin Masodi

Programme of Psychology, Centre for Research in Psychology and Human Well-being, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia

Email: shikeen11@yahoo.com

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v11-i9/11251 DOI:10.6007/IJARBSS/v11-i9/11251

Published Date: 21 September 2021

Abstract

This study aims to provide empirical evidence of validity and reliability of a personality instrument using the Rasch Measurement Model. Data was collected from 53 samples from a public university using the 5-Point Likert scale IPIP-NEO 120 (Johnson, 2014) questionnaire. Analysis using Rasch Measurement Model is attained through summary statistics, unidimensionality analysis and persons-items fit measures. Statistically, results from the study have shown that the items in the instrument have a high degree of validity and reliability; therefore, suitable for the measurement of personality characteristics among Malaysians. Premised on the concept item response theory, the Rasch Measurement Model is an ideal mean for instrument validation.

Introduction

Personality is a continuum that can be identified through the approach used to effectively cope with daily life obstacles (Carvalho et al., 2012). The first public-domain resource containing various personality items was the International Personality Item Pool, or IPIP developed by Lewis Goldberg (1999). Beginning at an initial compilation of 1252 items, IPIP has now expanded to 2413 items, all freely accessible on the IPIP website, http:/ipip.ori.org. A portion of these scales were designed to function as proxies for commercial inventory constructs, offering an alternative to these inventories within public domain (Goldberg et al., 2006). A 300-item inventory (Goldberg, 1999) aimed at measuring constructs similar to those quantified by the 30 facet scales in the NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992) was one of the first personality measures to be constructed from the IPIP. Johnson (2000, 2001) developed an adaptation of Goldberg's new inventory to be administered on the web and termed Goldberg's 300-item inventory as IPIP-NEO. Like the NEO PI-R, IPIP-NEO could also yield scores for the five broad domain of the Five-Factor Model

(i.e. Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness to Experience) and the six narrow facets scales within each broad domain (Costa & McCrae, 1992).

Despite vast empirical evidence supporting reliability, validity and usefulness of the 300-item IPIP-NEO, the inventory suffered from one significant shortcoming; it was lengthier than the current 240-item NEO PI-R. The length of the IPIP-NEO could pose difficulty for researchers intending to include an inventory assessing the five major personality factors along with a battery of psychological tests. While there are other IPIP five-factor inventories comprising of 20, 50 and 100 items, none can assess the six narrow facet scales within each of the five broad domains. Hence led to the development of IPIP-NEO 120, a 120-item version of the IPIP-NEO that can accurately and validly represent not just the five domains, but also the 30 facets within the Five-Factor Model (Johnson, 2014).

Item response theory (IRT) has been proposed as a way of enhancing traditional scale construction strategies ever since the publication of Goldberg (1972) new personality scale construction tool (Morizot, Ainsworth, & Reise, 2007). According to experts, IRT was not a replacement for conventional scale construction methods based on classical test theory (CTT). There is still a need to perform fundamental analyses such as internal consistency as well as inter-item and item total correlations. Mainly because IRT modelling may be negatively affected if some items with CTT statistics showed poor psychometric properties (Morizot et al., 2007). Likewise, implementing IRT models did not imply abandoning CTT. More precisely, IRT complemented CTT in providing a comprehensive and thorough analysis of an instrument (Reeve & Fayers, 2005). IRT contributes to conventional approaches through its capacity to see how well items vary across various levels of a trait. For example, the four-item IPIP scales depicting the five major personality domains in addition to Honesty-Humility (Sibley et al. 2011) were verified by Sibley (2012) using IRT as modestly accurate short-form representations of the six major broad-bandwidth personality dimensions beyond "a fairly broad range of each latent trait that centred on average or mean levels of each trait" (p. 26).

There are several IRT-based models, but the simplicity and measurement properties of the Rasch Model makes it a more prominent (Embretson & Reise, 2000). A latest literature review also confirmed that Rasch analysis is an influential psychometric approach to research in psychology (Aryadoust et al., 2019; Edelsbrunner & Dablander, 2019). This model parameterizes items according to its intensity when assessing a latent trait; hence why it is called the one-parameter Rasch model. The Rasch Measurement Model was therefore used in this study to evaluate the validity and reliability of IPIP-NEO 120 (Johnson, 2014) in Malaysian samples.

The Rasch Measurement Model

In social science research, Rasch model analysis was the only available method to assess whether constructs within an instrument measured explicit objectivity within a specific dimension, minimised ambiguous measures, and was a reasonable approximation of accuracy and implicit consistency (Azrilah, 2011). In order to generate a measurement technique with logit function, this model implemented a logarithm for either the odds or probabilistic value, depending on the nature of the study. Like other tools that constructed a model to match the research data, the logit was on a scale with separations of equal interval and measurement began at 0. In this case, the hierarchical relationship between the response of an individual

to an item and the level of construct measured by the scale was defined (Edelen & Reeve, 2007). Every item had its own logit measure, which signified the question's difficulty level. The Rasch Measurement Model theorem was therefore used in this research to assess data integrity and to expand on the Cronbach's alpha for data reliability and validity.

The Rasch model was well-known for transforming ordinal data to interval data, deciding if the latent traits or constructs were within the same dimension while controlling data quality before the study was further executed (Bond & Fox, 2015). The aim of the Rasch Measurement Model was therefore to generate a linear measure, resolve missing data, estimate accuracy and quality of the construct relating to the instrument as well as to detect misfits or outliers. This was achieved in view of three parameters; point measure correlation $0.4 < x < 0.8$, infit and outfit mean square (MNSQ) values $0.5 < y < 1.5$ and z-standard $-2 < z < 2$. It also helped provide parameters to research objectives, either through separability or independence measurement method.

Methodology

This research was conducted via quantitative survey method. It involved the use of self-reported questionnaire with 5-point Likert type scale. Participants comprised of 53 masters and PhD students from a local university. Four samples had to be omitted from the study due to non-response. The remaining samples' age ranged between 22 and 32 years (mean ± standard deviation = 24.94 ± 2.375), where 79.2% (n = 49) were female. Translated version of IPIP-NEO 120 (Johnson, 2014) was administered to all participants within classroom setting. Time for completion of questionnaire was approximately 25 minutes.

For the purposes of this research, the rating scale model was used for data screening analysis via Winsteps software (version 4.4.7) as proposed by other researchers (Bond & Fox, 2015; Azrilah et al., 2013). Summary statistics were derived based on the values from important indicators such as Cronbach alpha, item reliability, individual reliability, person measure and standard error to determine if data obtained were satisfactory for further study. In addition, item and person model parameters were calibrated using the joint maximum likelihood estimation method available on Winsteps. Item difficulties, also referred to as average thresholds, were set at zero setting to define scale metrics. Numerous statistical analyses were used to evaluate model fit, specifically, descriptive statistics for person and item parameters; model fit indices (infit and outfit); reliability indices; and item threshold values.

Results and Discussions

An accurate application of the Rasch model required that certain fundamental criteria be met. First and foremost, the adequacy of the data collected from the survey needed to be evaluated. This was to ensure that the reliability and validity was statistically acceptable and fitting for further analysis. At the same time, it was essential to assess if the latent traits or constructs in the instrument were measuring the specific objectivity within a specific dimension. In line with fulfilling the aforementioned, Rasch Measurement Model analysis was conducted, and results of the survey were analysed in view of three key parameters; the statistical summary, person-item fit measurement and unidimensionality. All of which were important to confirm reliability and validity, as well as to control quality of data.

In terms of reliability values, the indication that ought to be examined were Cronbach alpha (α) value, person reliability value, person measure and valid responses (Azrilah Abdul Aziz, 2010). Kuder-Richardson (KR-20) and coefficient alpha (Cronbach, 1984) test values were used to explain consistency responses observed by the Rasch model interpretation on person and item reliability. In the current analysis, KR-20 was applied to assess reliability within the range of 0.00 to 1.00. Values close to 1.00 suggested that the variables evaluated could be measured. Fraenkel and Wallen (1996) indicated that the reliability of items was adequate when α value was between .70 and .99. Similarly, Kubiszyn and Borich (2000) established that α value within .80 and .90 were acceptable (in Mohamad et al. 2015). However, in social science, the alpha value of .60 was already deemed appropriate (Ghazali 2008), as practiced by several scientists within the field. Table 1 indicated acceptable reliability values for person and item according to Fisher's (2007) rating scale instrument which was based on Rasch literature and his substantial experience of conducting Rasch analysis across various settings.

Table 1
Rating Scale

Person and Item Measurement Reliability	
Poor	<.67
Fair	.67 - .80
Good	.81 - .90
Very Good	.91 - .94
Excellent	>.94

The Rasch model was an appropriate method for achieving the above criteria to meet purposes of this study. The summary statistics for respondents (person) and item (questions) is as shown in Table 2. A total of 53 respondents comprising of masters and PhD students from a local university in Malaysia were included in the analysis. If all variables fell into an accepted range, the research data would predictably fit the model. We tabulated and evaluated the outcomes of their responses.

Table 2
Person-Item Reliability and Separation Index

Research Instrument	Person		Items		Cronbach Alpha (α)
	Reliability	Separation Index	Reliability	Separation Index	
IPIP-NEO 120	.89	2.84	.95	4.36	.90

Statistics were used to measure the test reliability of inter-item consistency. Higher value signified a strong relationship between the test items, whereas lower value suggested a weak relationship between test items. As tabulated in Table 2, the α value obtained was .90, which was significantly higher than the acceptance level of .60 (Garson, 1998; Gliem & Gliem, 2003; Leedy & Ormrod, 2005). This indicates confidence in the instrument's consistency in measuring the students' personality. Rasch also provided person reliability of .89, which provided 'Good' reliability (Fisher, 2007). This means that the assessment was able to define or distinguish personality level of the students. Besides, it allowed for repeatability

(Andrich, 1988), in which the likelihood of the ability pattern, or the position of students on the person-item distribution map, would remain comparable if this group were to be given different sets of personality instruments.

Similarly, as stated in Table 2, the item reliability of IPIP-NEO 120 (Johnson, 2014) was 0.95, which meant 'Excellent' reliability (Fisher, 2007). A higher value implied a strong relationship between the test items, while a lower value indicated that the test items had a weaker relationship. This supports the notion that if the measurement were to be administered to different groups of students, the intensity of each item would remain similar or unchanged. The person separation index of 2.84 was considered good when the value was greater than 2. Likewise, since the value was greater than 2 but not exceeding 5 (Linacre, 2014; Siti Rahayah et al., 2010), the item separation index of 4.36 was also considered excellent.

Table 3
Five-Factor Domains of Personality Reliability Index

No.	Domains	No. of Items	Reliability	Rating Scale
1.	Openness	24	.93	Very Good
2.	Conscientiousness	24	.95	Excellent
3.	Extraversion	24	.92	Very Good
4.	Agreeableness	24	.97	Excellent
5.	Neuroticism	24	.89	Good

In addition to having outstanding reliability as a whole, IPIP-NEO 120 (Johnson, 2014) also displayed a high value of the reliability index for all five domains (i.e. openness to experience, conscientiousness, extraversion, agreeableness and neuroticism) as shown in Table 3. There were excellent reliability indexes in both conscientiousness and agreeableness domains, followed by openness to experience and extraversion with very good reliability indexes. While marginally lower than the other domains, according to the Fisher (2007) rating scale instrument, the neuroticism domain still showed a good reliability index.

In summary, results from the person and item reliability index, along with the high value of reliability in each domain confirmed that the adapted version of the IPIP-NEO 120 (Johnson, 2014) questionnaire was admissible and was suitable to describe personality traits within the local context.

Quality Control with Fit Statistics

Rasch analysis offered fit statistics to verify fundamental measurement assumptions (Wright & Stone, 1979). “Fitting the model” basically implied meeting basic measurement assumptions (e.g. virtually all the easy items should be endorsed correctly by high scorers). Once identified, it was possible to qualitatively analyse “misfits” persons and items to ascertain causes of setbacks. Problems may lie within items with incomprehensible terminology or items that determined a construct different from the one being assessed primarily (i.e. multidimensionality). In short, an understanding of poor fit may lead to refinement and/or discarding of items. Fit statistics, computed for both person and item, assessed the fit of the data to the model. Two misfit indicators assigned in the Rasch model was misfit and infit; infit being susceptible to inconsistent responses in items close to the

person ability level, whereas the misfit was outlier receptive. Mean square fit statistics were defined in such a way that 1.0 was the model-specified uniform randomness value (Wright & Stone, 1979).

Person fit showed the degree to which the person's performance was compatible with the other respondents' manner of using the items. While item fit demonstrated the degree to which a certain item's usage was consistent with how the survey respondents reacted to the other items. Values between .75 and 1.33 logits (log odd units) were considered appropriate (Wilson, 2005) for this form of analysis. With respect to Bond and Fox (2007), the range of 0.60 and 1.40 was equally appropriate. While the interpretation of fit statistical values entailed expertise relevant to a particular measurement context (Bond & Fox, 2015), there was a rule of thumb for such reasonable ranges among item mean-square statistics as displayed in Table 4 (Wright, Linacre, Gustafsson & Martin-Loff, 1994). Nonetheless, the fit statistics should be used to help identify problematic item and person output, not just to conclude which items should be eliminated from a test. Merely excluding the overfitting items could deprive the test from its finest item (Bond & Fox, 2015).

Table 4
Some Reasonable Item Mean Square Ranges for Infit and Outfit

Type of Test	Range
Multiple-choice test (high stakes)	0.8 – 1.2
Multiple-choice test (run of the mill)	0.7 – 1.3
Rating scale (Likert/ survey)	0.6 – 1.4
Clinical observation	0.5 – 1.7
Judged (where agreement is encouraged)	0.4 – 1.2

The infit and outfit MNSQ values of each item and person used in this research was between 0.50 and 1.50. This range was selected because it indicated that the mean-square fit statistics of the parameter level were beneficial to the measurement (Linacre, 2002). Since items within an instrument are priority in constructing a scale besides being the yardstick for measuring a variable, every single misfitting item should be thoroughly revaluated. If the misfit item did not enrich the yardstick concept and was not required for greater precision, it would be a judicious to exclude them. Nonetheless, there should be a logical reason for exclusion of each item (Wright & Stone, 2004).

The infit and outfit MNSQ values of items in the IPIP-NEO 120 instrument are attached in Appendix I. The MNSQ range was between 0.5 to 1.5, and majority of the items were within acceptable range. Out of 120 items, five had infit and outfit MNSQ values > 1.5 and were excluded from further analysis. The process of excluding misfit items was repeated several times until all items had infit and outfit MNSQ values within the accepted range. Overall, a total of seven items (A118, A41, A116, A67, A101, A19, A81) were excluded from the analysis because the infit and outfit MNSQ values were not within the $0.5 < y < 1.50$ range. Details of the misfit items are as shown in Table 5. The infit and outfit MNSQ values of remaining items (after omitting seven misfit items) are shown in Appendix 2. Out of the seven misfit items, four items were from the neuroticism domain, while the remainder were from openness to experience, extraversion and agreeableness domain respectively. Only items within the conscientiousness domain remained unperturbed.

Table 5
Misfit Items

Item	Domains	Sub Domain	Item	Item Measures	Infit MNSQ	Outfit MNSQ
A118 (R)	Openness	Liberalism	Percaya bahawa perbuatan jenayah perlu ditangani segera	-.74	1.82	1.99
A41	Neuroticism	Depression	Tidak menyukai diri sendiri	1.06	1.61	1.55
A116 (R)	Neuroticism	Vulnerable	Kekal tenang dalam keadaan tertekan	-.24	1.38	1.53
A67	Extraversion	Gregarious	Lebih suka menyendiri	.12	1.51	1.52
A101 (R)	Neuroticism	Depression	Selesa dengan diri sendiri	-1.20	1.41	1.52
A19 (R)	Agreeableness	Cooperation	Suka kepada persaingan yang mencabar	-.34	1.48	1.54
A81 (R)	Neuroticism	Immoderation	Mudah menahan godaan	.11	1.41	1.51

Following careful consideration of items, evaluation of samples or persons were conducted. It was also important that the samples were representative of the intended population (Wright & Stone, 2004). Appendix 3 displays the infit and outfit MNSQ values of the respondents or person in this research. Also, the selected MNSQ range was 0.5 to 1.5, and most samples were within acceptable range. Those whom were outside the range were omitted from the analysis because the item reliability could be affected by unexpected or inconsistent responses from misfit individuals, causing a measurement distortion. This process of omission was performed a few times until all respondents had infit and outfit MNSQ values within the acceptable range. Overall, a total of 15 respondents were omitted from the analysis because the infit and outfit MNSQ values were not within the $0.5 < y < 1.50$ range. In addition to fit statistics, the principal component analysis of residuals was used to determine whether a substantial factor existed in the residuals after the estimation of the primary measurement dimension (Smith, 2002; Aryadoust et al., 2019).

Unidimensionality

The unidimensionality analysis was necessary to ensure that the items in the instrument measured specific objectivity within the same domain. For data representing a single dimension of difficulty and ability, Rasch dimensionality analysis tested non-random variation observed (Sick, 2010; Hudiya et. al., 2017). Rasch model analysis then extended the key component analysis to the residuals, to assess whether the variance measured what was intended; and to examine the instrument's capability to estimate in a uniformed dimension with an appropriate level of distraction (Azrilah et al., 2017).

The unidimensionality test was carried out in this study to further examine whether the objects in the structures were consistently measured within the dimensions. The variance

explained by measures had to be above 40%, with unexplained variance less than 15% in the first contrast, based on strength of at least 3 items (Linacre, 2004). The unidimensionality criterion was also reported by Conrad, Conrad, Dennis and Funk (2011) to be 40%, which is aligned with Linacre's recommendation, with above 30% is considered a modest measurement dimension.

Table 6
Principal Component Analysis of Standardized Residual Correlations for Items

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units				
		Eigenvalue	Observed	Expected
Total raw variance in observations	=	193.4950	100.0%	100.0%
Raw variance explained by measures	=	80.4950	41.6%	41.4%
Raw variance explained by persons	=	5.9926	3.1%	3.1%
Raw Variance explained by items	=	74.5024	38.5%	38.3%
Raw unexplained variance (total)	=	113.0000	58.4%	100.0%
58.6%				
Unexplained variance in 1st contrast	=	17.9278	9.3%	15.9%
Unexplained variance in 2nd contrast	=	9.1744	4.7%	8.1%
Unexplained variance in 3rd contrast	=	7.5923	3.9%	6.7%
Unexplained variance in 4th contrast	=	7.1144	3.7%	6.3%
Unexplained variance in 5th contrast	=	5.9450	3.1%	5.3%

Once the misfit items and persons were detected and removed, the principal component analysis was carried out. As seen in Table 6, the raw variance explained by the measures was 41.6%, with an unexplained variance of 17.93 and 9.3% in the first contrast. This demonstrates that the raw variance explained by the measure was over 40%, suggesting that the items in the questionnaire were unidimensional. Despite the slightly high eigenvalue of 17.93 in the first contrast, the unexplained variance was still less than 15% based on strength of at least 3 items. Therefore, it was still acceptable. It was interesting to discover that the Malaysian samples perceived the items in IPIP-NEO 120, five-factor personality domains as a unidimensional personality assessment. Further investigation involving a larger sample size will be beneficial in validating this preliminary observation.

Item Polarity Analysis

The evaluation of item polarity is an indication used to convey that the items used shifted in the direction expected by the construct being observed. Instruments showing a positive index for all items demonstrated a synchronous movement in a parallel direction to measure the formed constructs. For items with negative index, the researcher must re-examine whether the data needs to be revised or dropped. This is because this marker suggests the existence of an item or person who responded contradictorily to the variable (Linacre, 2003). According to Bond and Fox (2001), item polarity or point-measure correlation (PTMEA Corr.) is also an early identification of construct validity.

As tabulated in Table 7, a total of 20 items (n = 120) had negative PTMEA Corr. values which meant that the items were not moving in the same direction as anticipated. Hence, it is necessary to refer to the item texts and rating scale to figure out what has caused this

unintended result. The two viable options are to fix the issue by rescoring the rating scale or discarding the item. As with most of the other items, when the correlations of the item measure are all positive, evaluating items with infit means squares > 1.5 and diagnosing the causes for their occurrence was a helpful rule of thumb (Wright & Stone, 2004). If extreme infit misfit was detected in only a few items and there was no reasonable answer, the easiest option was to omit these items. Nevertheless, the unforeseen misfit is worthy of reconsideration and diagnosing, especially since the original purpose for the item being included in the test was because they were hypothesized to fit. Therefore, all the negative PTMEA Corr. values were re-examined, taking everything into account as suggested by Wright and Stone (2004) before any deletion took place.

Table 7
Partial Item Statistics: Correlation Order

No.	Item No/Domains/ Sub-Domain	Item Measure	Infit MNSQ	Outfit MNSQ	PTMEA Corr.
1	A88_O6 (R)_Liberal	.38	1.13	1.14	-.27
2	A35_C1_Efficacy	-.51	.95	.98	-.13
3	A43_O3_Emotionality	-1.20	.65	.67	-.13
4	A108_O4 (R)_Adventure	.11	.95	.97	-.12
5	A10_C2_Order	-1.26	1.59	1.70	-.11
6	A20_C4_Achievement	-1.95	.88	.88	-.09
7	A24_A5 (R)_Modesty	.72	1.24	1.26	-.09
8	A101_N3 (R)_Depress	-1.51	1.72	1.83	-.09
9	A81_N5 (R)_Immoderate	-.13	.88	.88	-.08
10	A116_N6 (R)_Vulnerable	-.43	1.47	1.48	-.08
11	A84_A5 (R)_Modesty	-.09	1.28	1.27	-.07
12	A25_C5_Discipline	-.35	.70	.72	-.06
13	A95_C1_Efficacy	-.43	.88	.94	-.06
14	A117_E6_Cheerful	-1.26	.95	.97	-.04
15	A72_E3_Assertive	-.43	.81	.82	-.04
16	A111_N5 (R)_Immoderate	-.47	.81	.83	-.04
17	A96_N2 (R)_Anger	.41	1.01	1.01	-.02
18	A50_C4_Achievement	-1.03	.81	.83	-.01
19	A42_E3_Assertive	-.09	1.03	1.03	-.01
20	A58_O6_Liberal	-.33	1.24	1.21	-.01

Conclusions

This study aimed to assess validity and reliability of IPIP-NEO 120 (Johnson, 2014) as a research instrument. The application of Rasch Measurement Model in the identifying validity and reliability of the research instrument was advantageous because the model could describe the constructs of valid items, besides providing a consistent description of the measured constructs in accordance with the theoretical assumptions. Interestingly, this model could be efficiently used on items that could be reliably measured and used with valid response patterns. Results generally suggested commendable adequacy of psychometric properties for the scales within IPIP-NEO 120 questionnaire. The Rasch model was able to check for different reliability indices that may vary across levels of the latent trait being measured by the items. This was done by assessing the local error, in addition to reliability

analyses conventionally used. Among the limitations of the study was its relatively small sample size (n=53) and its lack of demographic diversity. Future studies should consider replicating the validity and reliability of the five-factor personality model using a bigger and more diverse group of samples, so that a Malaysian five-factor personality norm could be established. Nevertheless, the study design used to evaluate suitability of items in research instruments to fit the model was supported by the results obtained. Therefore, it was essential to improve and strengthen quality of items in the instrument in order to measure the intended construct effectively.

References

Andrich, D. (1988). Rasch models for measurement. Newbury Park: Sage

Aryadoust, V., Tan, H. A. H. & Ng, L. Y. (2019). A Scientometric Review of Rasch Measurement: The Rise and Progress of a Specialty. Front. Psychol. 10:2197.

Abdul Aziz, A. (2011). Rasch Model Fundamentals: Scale Construct and Measurement Structure: Integrated Advance Planning Sdn. Bhd.

Abdul Aziz, A., Masodi, M. S., & Zaharim, A. (2017). Asas Model Pengukuran Rasch: Pembentukan Skala & Struktur Pengukuran. Bangi: Penerbit Universiti Kebangsaan Malaysia.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Oxford: Psychology Press.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd Ed. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd Ed. Oxford: Psychology Press.

Carvalho, L. F., Primi, R. & Meyer, G. J. (2012). Application of the Rasch model in measuring personality disorders. Trends Psychiatry Psychother., 34 (2):101-9.

Conrad, K. M., Conrad, K. J., Dennis, M. L., & Funk, R. (2012). Validation of the self-help improvement scale to the Rasch measurement model GAIN methods report 1.0. http://gaincc.org/_data/files/Psychometrics_and_Publications/Working_Papers/Conrad_et_al_2011_SPS_Report.pdf.

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

Cronbach, L. J. (1984). Essentials of Psychological Testing (4th Edition). New York: Harper & Row

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modelling to questionnaire development, evaluation and refinement. Quality Life Research, 16, 5-18.

Edelsbrunner, P. A., & Dablander, F. (2019). The psychometric modelling of scientific reasoning: A Review and Recommendations for Future Avenues. Edu. Psychol. Rev. 31, 1–34.

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

Fisher, J. W. P. (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions 21(1): 1095.

Fraenkel, J. R., & Wallen, N. E. (1996). How to design and evaluate research in education (3rd Ed.). New York: McGraw-Hill.

Darusalam, G. (2008). Kesahan dan Kebolehpercayaan Dalam Kajian Kuantitatif and Kualitatif. Jurnal Institut Perguruan Islam. April.

Gliem J. A., & Gliem R. R. (2003). Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education, Columbus, 82-88.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. Multivariate Behavioral Research Monograph, 7, No. 72-2.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. in Mervielde, Deary, De Fruyt, & Ostendorf (Eds.), Personality Psychology in Europe, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: 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, 84-96.

Johnson, J. A. (2000). Web-based personality assessment. Paper presented at the 71st Annual Meeting of the Eastern Psychological Association, Baltimore, MD.

Johnson, J. A. (2001, May). Screening massively large data sets for non-responsiveness in web-based personality inventories. Invited talk to the joint Bielefeld-Groningen Personality Research Group, University of Groningen, The Netherlands. Retrieved from http://www.personal.psu.edu/~j5j/papers/ConferencePapers/2001BGPRG.pdf

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.

Leedy, P. D., & Ormrod, J. E. (2005). Practical research planning and design (8th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Linacre, J. (2002). What do Infit and Outfit, Mean Square and Standardized mean? Rasch Measurement Transactions.

Linacre, J. M. (2003). Size vs. significance: Standardized chi-square fit statistic. Rasch Measurement Transactions, 17(1), 918.

Linacre, J. M. (2004). From Microscale to Winsteps: 20 years of Rasch software development. Rasch Measurement Transactions. 17:958.

Linacre, J. M. (2014). WINSTEPS Rasch Measurement (Version 3.81.0). Beaverton, OR: Winsteps.com.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. Procedia-Social and Behavioral Sciences, 204, 164-171.

Morizot, J., Ainsworth, A. T. & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (407-423). New York: Guilford Press.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers and R. D. Hays (Eds.), Assessing

quality of life in clinical trials: Methods of practice (2nd Ed., pp. 55-73). Oxford: Oxford University Press.

Sibley, C. G., Luyten, N., Purnomo, M., Moberly, A., Wootton, L. W., Hammond, M. D., Sengupta, N., Perry, R., West-Newman, Wilson, M. S., McLellan, L., Hoverd, W. J. & Robertson, A. (2011). The Mini-IPIP6: Validation and extension of a short measure of the Big-Six factors of personality in New Zealand. New Zealand Journal of Psychology, 40, 142-159.

Sibley, C. G. (2012). The Mini-IPIP6: Item Response Theory analysis of a short measure of the big-six factors of personality in New Zealand. New Zealand Journal of Psychology, 41, 21-31.

Sick, J. (2010). Assumptions and requirements of Rasch measurement. JALT Testing & Evaluation SIG Newsletter, 14 (2): 23-29.

Smith, E. Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J. Appl. Measurement. 3, 205–231.

Wilson, M. (2005). Constructing Measures: An Item Response Modelling Approach. Mahwah, NJ: Lawrence Erlbaum Associates

Wright, B. D., Linacre, J. M., Gustafson, J. E. & Martin-Loff, P. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (2004). Making Measures. Chicago: The Pheneron Press.