Gadjah Mada Journal of Psychology (GamaJoP), Volume 9, Number 2, 2023: (page 232-247)
E-ISSN 2407-7798
https://jurnal.ugm.ac.id/gamajop
DOI: 10.22146/gamajop.73878

Psychometric Properties of Mental Health Scale: An Item Response Theory Approach

Ramadhan Dwi Marvianto[*]
[*]Faculty of Psychology, Universitas Gadjah Mada, Indonesia

Submitted 1 April 2022 **Accepted 4 August 2022** **Published 31 October 2023**

Abstract. Indonesia Family Life Survey (IFLS): the fifth wave used the Centre for Epidemiological Studies Scale (CES-D10) to measure mental health state, which was a depressive symptoms construct. It was primarily used in some studies regarding depressive symptoms. However, there was no specific validation research in the Indonesian context. Thus, this study aimed to (1) investigate the item quality of the depressive symptom measurement from the IFLS fifth wave, (2) measurement precision of this scale, and (3) the measurement invariance based on gender using the IRT approach. This study used data from ILFS-5 in the KP ("*Keadaan Psikologis*" or Mental Health) section, which was CES-D10, consisting of 10 items, and conducted IRT analysis using a 2-PL model. The number of participants in the data was 31,447. CFA analysis resulted in a unidimensional model fit for this scale. Moreover, the finding showed that this scale had good psychometric properties, including item discrimination and item location, except for items 5 and 8. Despite the poor quality of those items, the reliability coefficient, including the items, met a reliable measurement criterion. Also, this scale had much information for assessing medium and severe depressive symptoms. Moreover, the Differential Item Functioning (DIF) analysis indicated that there was no item exhibiting DIF.

Keywords: psychology; quantitative method; psychometrics; mental health scale

Recently, depression has become a serious problem around the world. It was already a significant public health issue, as the increase from 1990 to 2017 was about 49.86% (Liu et al., 2020). Moreover, the World Health Organization (WHO, 2017) published a report stating that around 325 million, or 4.4% of people suffered from depression. The South-east Asia region had the most significant contribution, accounting for about 27% of the total case. Specifically, in Indonesia, the prevalence of depression among Indonesian people in the same year was around 2.50%, or 6 million people (University of Washington, 2021). This number seemed to be perceived as a small amount, yet it might grow remarkably. Therefore, screening and measuring depression were essential to reducing the inclined trend.

In Indonesia, there is a large-scale assessment called the Indonesian Family Life Survey (ILFS). Conducted by RAND (2021), the ILFS is a well-known survey that measured demographic and social-economic, education, and health aspects of Indonesian society. It started in 1993 and ended in 2015 as the fifth wave. This survey, mainly from the fourth wave, included a depression symptom measurement using the Centre for Epidemiological

[*] Corresponding Author: ramadhan.dwi.m@mail.ugm.ac.id

Studies Scale (CES-D10) to gauge participants' mental health condition. CES-D was a short scale designed to measure depressive symptoms in the general population (Roadolff, 1977).

This depressive symptom measurement from ILFS was used mainly in some studies. Starting from IFLS-4, it was applied to search its determinant factor (Kashiwagi et al., 2016), its relationship with recurrent aphthous stomatitis (RAS; Hariyani et al., 2020), spirituality (Mahwati, 2017), smoking behaviour (Liew & Gardner, 2016; Purborini et al., 2021), perceived health, and acute morbidities (Purborini *et al.*, 2021), explore multimorbidity (Hussain et al., 2015) and prevalence (Purborini et al., 2021). Also, the fifth wave was used to estimate a prevalence amount (Astutik et al., 2021), be a screening tool (Pengpid et al., 2019), be an indicator of the quality of life (QoL; Yuniati & Kamso, 2020), explore morbidity (Leung et al., 2021), an exclusion criterion (Peltzer & Pengpid, 2018) and find an association of depressive symptoms with a chronic condition, health, and physical functioning (Astutik *et al.*, 2021), sleep disturbance (Isaura et al., 2020), sociodemographic factors, stressors, lack of social trust, soft drink consumption, lack of religiosity, a chronic condition, and tobacco use (Peltzer & Pengpid, 2018), living alone (Widhowati et al., 2020), and physical activity. Not only the fourth and fifth wave, but the ILFS East version gauged in 2012 was also used as a measurement of mental well-being (Cao & Rammohan, 2016).

However, RAND (2021a) did not mention any technical report about the quality of this measurement. Despite lacking information on the quality, some studies found that CES-D had good psychometric properties in other contexts. Baron et al. (2017) found some valid evidence regarding internal structure and association with other variables using the classical test theory approach. They examined sensitivity and specificity of the South African population. Likewise, Zhang et al. (2012) conducted exploratory factor analysis, reliability, sensitivity and specificity estimation of CES-D-10 Canada for people with HIV. Similarly, in a more specific subgroup population, Björgvinsson et al. (2013) found that this single-factor model fitted with the data obtained, each item had a satisfactory discrimination index and factor loading, and also that this measurement got validity evidence based on association with other variables, which was either a relevant or not relevant construct. However, there was no specific validation of CES-D in the Indonesian context, and the validation was not applied to the newest approach called item response theory (IRT).

IRT was a new approach to construct and evaluate an instrument that some psychologists and educationalists initially developed to produce quality examination and non-exam assessments (Hambleton et al., 1991). Recently, this approach was applied to evaluate some health measurement tools, particularly on depression (Ayis et al., 2018; Chiesi et al., 2017; Giusti et al., 2020; Gustryanti et al., 2017). This approach also provided robust item parameter estimation, which did not depend on the particular sample as much as the classical test theory approach did (Ayala, 2009).

Furthermore, IRT can examine item bias using a method called Differential Item Function (DIF), a method to detect whether items were differently responded to by specific subgroups (Ayala, 2009). The difference between male and female on a depression level became an interesting issue. Some studies found that females had a more significant risk of depression symptoms (Girgus et al., 2017; Nazroo et al., 1998). Therefore, the score obtained by this measurement should be free from gender bias, which the female group tends to respond to highly. This study aimed to (1) investigate the item quality of the depressive symptom measurement from the IFLS fifth wave, (2) measurement precision of this scale, and (3) the DIF based on gender using the IRT approach.

Method

Participants

This project used data sources from IFLS-5. Strauss et al. (2016) reported that this survey involved information on individual, household and community levels from 13 out of 27 provinces in Indonesia. The data sources consisted of 31.447 individuals from 14.714 households. Statistically, more than half of the participants were female (53.25%). In terms of age and scale score, both genders had an adequate mean and standard deviation. See Table 1 for the statistic descriptive.

Table 1.

Statistic Descriptive Based on Gender

	All sample (*N*= 31.447)			Male sample (46.75%)		Female sample (53.25%)	
	Mean	SD	Range	Mean	SD	Mean	SD
Age	37.33	14.93	14 – 101	37.59	14.99	37.09	14.88
Mental health	16.4	4.78	10 – 40	16.31	4.67	16.48	4.88

Instrument

IFLS-5, reported by Strauss et al. (2016), included measurement of depressive symptoms using CES-D. It was constructed by Roadolff (1977), which initially consisted of 20 items. Then, Andresen et al. (1994) analysed to make this scale shorter, and it resulted in 10 items widely used in some research. Also, the scale was placed on Book 3D in "Mental Health" in English or "Kesehatan Psikologis" in Bahasa, which measures symptoms of depression that participants perceived within the past two weeks. Moreover, the scale used the Likert scale for the response format, starting from 1 (Rarely or none) until 4 (Occasionally).

Data Analysis

Descriptive Statistics. It was conducted for each item, including mean, standard deviation (SD), and response proportion. Berk (2006) stated that the mean and SD reflect how easy the participants agreed to the item and how each item's response varied across all participants, respectively. The proportion of response showed the percentage of participants who selected the specific response options. This analysis was conducted using the *base* package in R (R Core Team, 2020).

Dimensionality. This project conducted CFA to investigate the proposed unidimensional model of this scale. To check the fit of the proposed model, it used RMSEA, which is part of non-incremental fit indexes. It is more stable than incremental fit indexes such as CFI, and RMSEA is more appropriate in confirmatory contexts (Rigdon, 1996). Also, factor loading was investigated to examine how well each item represents the latent variable underlying it (Furr & Bacharach, 2013). The CFA was conducted through R using *the lavaan* package (Rosseel, 2012).

IRT Analysis. 2-PL IRT analysis that would be considered for this project were the Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM) as these models also included the estimation of item discrimination i n the analysis besides item difficulty (Ayala, 2009). The use of item discrimination in this project analysis was to examine how each item can differentiate the individual who has high and low ability on the measured construct. This parameter relates to item information with higher value associated with the peaked item information curve (Embretson, 1985). Selecting the model would use Bayesian Information Criteria (BIC) to deal with overfit issues and assess the model selection uncertainty (Wu et al., 2019). The lower value indicated a better fit.

The item characteristic curve (ICC), which used a category characteristic curve or cumulative category characteristic curve (depending on the fittest model between GPCM and GRM), would be generated to visualize item difficulty and discrimination parameters. The analysis above was conducted using *mirt* package in R (Chalmers, 2012).

Reliability Estimation. This project generated an IRT test reliability coefficient to get the reliability estimation for the scale. Kim and Feldt (2010) argued that the use of the IRT test reliability coefficient is more appropriate when IRT is used than CTT approaches such as Alpha. Also, IRT analysis provides reliability estimation for other groups beyond the selected group included in the dataset since CTT is a test- and group-dependent. This reliability estimation needed a latent score, which then this project would use as expected a posteriori (EAP) estimation as it was most popular in use and easy to compute for unidimensional construct (Brown & Croudace, 2015). This estimation also used the *mirt* package in R (Chalmers, 2012).

The Precision of Measurement. Feuerstahler et al. (2020) explained that precision of measurement across the trait ability could be indexed by a function called the test information function. This function was a graph with the y-axis as information value and x-axis as participants' ability scores, which illustrated where ability has the most informative one area and which range of ability region where the scale was reliable. This graph was generated using *the ltm* package (Rizopoulos, 2006).

Different Item Function. This term refers to a method to identify functioning items differently over subgroups of particular populations (Ayala, 2009). Item exhibiting DIF was indicated as an item bias. On the other hand, the Non-DIF item indicated that persons from different subgroups within a specific population had an equal chance of answering the item correctly. The selection of the item exhibiting DIF used BIC value and adjusted significance value using Bonferroni correction to reduce false-positive rate in multiple testing (Haynes, 2013). The analysis also used the *mirt* package in R (Chalmers, 2012).

Results

Item Descriptive Statistic for the Item

Table 2 shows the code and wording of each item, as well as some descriptive statistics. Item means ranged from 1.34 to 1.90, and SD showed between 0.75 and 1.09, indicating that the items were hard to be agreed and had a widespread response, respectively (Berk, 2006). It is supported by the proportion of responses in which participants selected all the category responses. However, most participants chose the first category, which was more than 50%, or .50, in almost all items.

Table 2.

Item Code. Content and Descriptive Statistic For Mental Health Scale (N = 31.447)

Item		Mean	SD	Factor Loading (SE)	Proportion of Response			
Code	Wording				1	2	3	4
MH01	I was bothered by things that usually don’t bother me	1.58	.88	.55 (.006)	.64	.18	.14	.04
MH02	I had trouble concentrating in what I was doing	1.62	.89	.67 (.006)	.62	.19	.15	.04
MH03	I felt depressed	1.45	.82	.69 (.005)	.73	.14	.10	.04

Table 2. (Continued)

Item Code. Content and Descriptive Statistic For Mental Health Scale (N = 31.447)

Item		Mean	SD	Factor Loading (SE)	Proportion of Response			
Code	Wording				1	2	3	4
MH04	I felt everything I did was an effort	1.87	1.09	.53 (.006)	.54	.17	.16	.13
MH05	I felt hopeful about the future	1.90	1.06	-.17 (.006)	.50	.22	.17	.11
MH06	I felt fearful	1.54	.89	.60 (.006)	.68	.15	.12	.05
MH07	My sleep was restless	1.76	1.07	.50 (.006)	.61	.13	.15	.11
MH08	I was happy	1.88	.98	.06 (.006)	.46	.28	.18	.08
MH09	I felt lonely	1.34	.75	.52 (.006)	.81	.09	.08	.03
MH10	I could not get going	1.46	.84	.53 (.006)	.73	.11	.12	.04

Notes. 1 = Rarely or none (< 1 day); 2 = Some days (1-2 days); 3 = Occasionally (3-4 days); 4 = Occasionally (3-4 days).

Dimensionality

This project assumed that the scale was constructed using a unidimensional framework. Then, CFA was conducted to test whether the unidimensional model was supported by the data. It resulted in an RMSEA of .086, which exceeded the cut-off for a close fit model proposed by Hu and Bentler (1998), that is .06. However, RMSEA value between .08 to .10 indicated a mediocre fit (MacCallum et al., 1996). Also, Olino et al. (2008), in their study, claimed their RMSEA finding of .90 as an acceptable value. Therefore, this model can still be considered as a unidimensional model. Nevertheless, considering factor loading, items HM05 and MH08 had a small factor loading coefficient. Item HM05 even had a negative value that was -0,17 and HM08 had 0,06. Both these items were unfavourable items. Thus, it was suggested to have further investigation about the function of these items.

Item Parameter

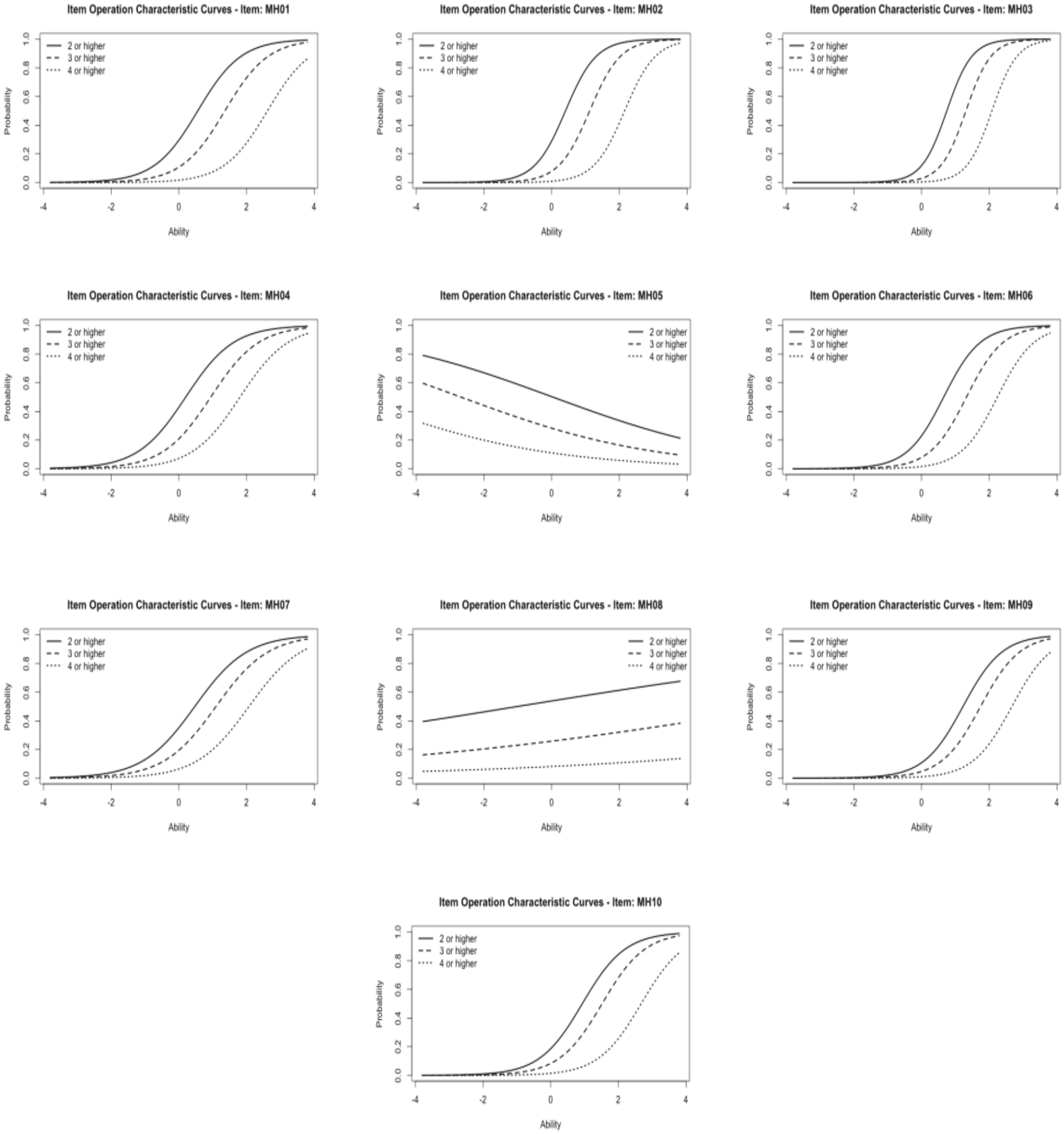
IRT model selection was conducted using GPCM and GRM and evaluated based on the BIC value. It showed that the GPCM (589669.5) model had a higher BIC of 5609.638 than the GRM (584059.9) model. Therefore, this project used the GRM model for the rest of the analysis.

Table 3.
Item Discrimination (A) and Location (B) of Mental Health Scale.

Item	Discrimination (α)	Item location (δ)		
		Threshold 1 (δ_1)	Threshold 2 (δ_2)	Threshold 3 (δ_3)
MH01	1.56 [1.52. 1.60]	.56 [.54, .59]	1.37 [1.33. 1.40]	2.63 [2.56. 2.69]
MH02	2.19 [2.13, 2.25]	.42 [.39, .43]	1.13 [1.11, 1.15]	2.16 [2.12, 2.21]
MH03	2.70 [2.62, 2.79]	.75 [.73, .77]	1.32 [1.29, 1.34]	2.08 [2.04, 2.12]
MH04	1.41 [1.37, 1.45]	.21 [.19, .23]	.95 [.92, .98]	1.81 [1.77, 1.86]
MH05	-.35 [-.37, -.32]	.04 [-.03, .11]	-2.69 [-2.89, -2.48]	-6.02 [-6.45, -5.58]
MH06	1.85 [1.80, 1.90]	.66 [.64, .68]	1.33 [1.30, 1.36]	2.24 [2.20, 2.29]
MH07	1.30 [1.26, 1.33]	.48 [.45, .50]	1.10 [1.07, 1.14]	2.08 [2.03, 2.13]
MH08	.15 [.13, .18]	-1.00 [-1.22, -.78]	6.89 [5.80, 7.97]	15.81 [13.32, 18.30]
MH09	1.74 [1.68, 1.79]	1.23 [1.20, 1.26]	1.75 [1.70, 1.79]	2.67 [2.60, 2.74]
MH10	1.58 [1.54, 1.63]	.94 [.91, .97]	1.52 [1.49, 1.56]	2.68 [2.61, 2.74]

In general, item discrimination ranged from 1.30 to 2.70, which were categorized as a good item in discriminating participants above 1.35 (Baker, 2001). However, MH05 and MH08 showed low and even negative values, indicating that the items were ineffective in discriminating students' ability. It was supported by Figure 1, showing that both items did not have steep curves for each probability in answering the higher category. Also, it had a different pattern in terms of item location as compared to the others. The other items had probability progressively starting from below 1, 1-2, and more than 2 for each threshold in order, while these items had remarkable progression. For instance, MH05, its item location for choosing category two or higher was .04, then the category three or higher was -2.69, which was a significant difference between the two thresholds. Even category 4 was double the previous threshold.

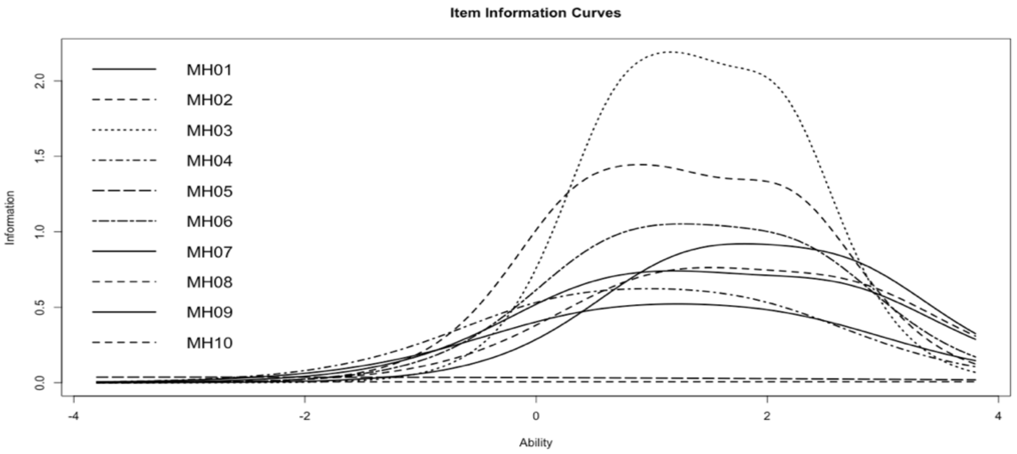
Figure 1.
Cumulative Category Characteristic Curve for Mental Health Scale's Items



In terms of information function, Figure 2 illustrates each item information curve. Almost every item gave much information between ability around 0 to 3, which meant that it could separate the participant who suffered a moderate and severe level of depressive symptom. However, MH05 and MH08 showed the horizontal line indicating that they almost did not give any information for all ability. Therefore, it supported that these items did not work functionally. On the other hand, MH03 was the most informative item as compared to the others.

Figure 2.

Item Information Curve for Mental Health Scale's Items



Reliability Estimation

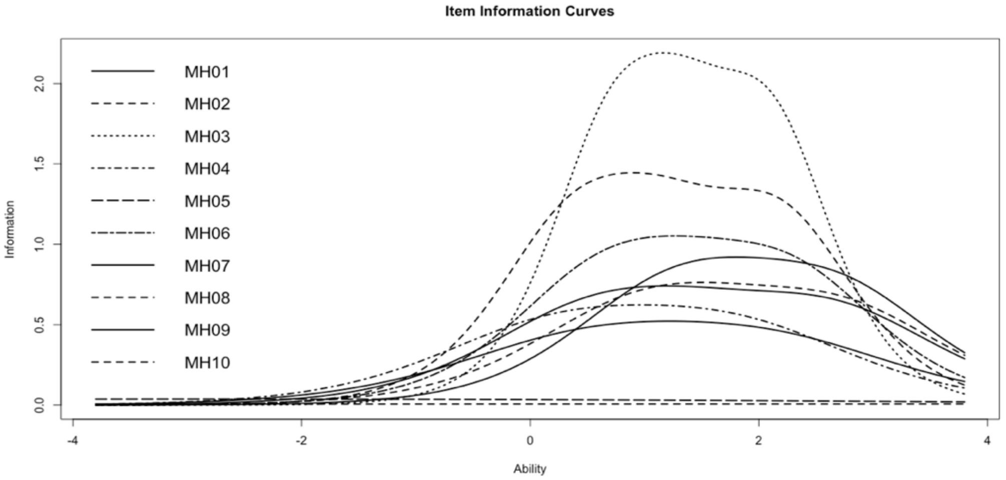
This project evaluated reliability using IRT test reliability, including MH05 and MH08 as the original scale. It resulted in an IRT test reliability coefficient of .756, which met the criteria of good reliability .70 (Nunnally & Bernstein, 1994). Therefore, despite the inclusion of MH05 and MH08, the measurement using the 10-items will still result in reliable results.

Precision of Measurement

Figure 3 showed the test information curve. This scale was suitable for detecting participants who had a moderate and severe level of the depressive symptom as the high information was on ability between around 0 to 3.

Figure 3.

Test Information Curve for Mental Health Scale's Items



Differential Item Function

Based on the DIF analysis, the decrease of BIC value for each item ranged between 9 and 10. The adjusted significance level using Bonferroni correction showed that all items were almost perfectly close to 1, which indicated that all items did not exhibit DIF at all. Then, it can be said that there is no item exhibiting DIF based on gender. Male and female participants had an adequately similar probability of agreeing to each item.

Discussion

Based on our information, studies examining the mental health section's psychometric properties from IFLS-5 are limited. So, this project is providing the psychometric properties of the scale. In general, findings suggest that this scale has good psychometric properties for measuring symptoms of depression in the Indonesia sample.

However, although the data's unidimensional model is supported, this project finds that MH05 and MH08 have low factor loading and item discrimination coefficient. This result is in line with previous study findings that stated that these items had low factor loading (Boey, 1999; Bradley et al., 2010; Lee & Chokkanathan, 2008) when measured in unidimensional construct. Also, some previous studies found that these items were included in a different factor from the other items, namely positive affect (Baron et al., 2017; Boey, 1999; Cheng et al., 2006; Lee & Chokkanathan, 2008), while the remaining items stood for depressed affect. Depressed affect reflected negative effects, for which Diener and Emmons (1984) stated that negative and positive affect were relatively independent. It means those might be separated into different factors.

This finding can be explained by Schroevers et al. (2000) idea that positive affect was broadly tapping with other constructs such as compared to another factor. They also stated that depressed affects from CES-D could represent the essential characteristics of depressive symptom, namely miserable mood, loss of interest, loss of pleasure, weight loss or gain, sleeping problems, problems in motor and cognitive activity, loss of energy, worthlessness, guilty feelings, suicidal thoughts, and concentration problems. Otherwise, positive affects did not reflect the characteristics above. It is also supported by Iwata et al. (1998) study that depressed affect measure could differentiate between depressed patients and non-depressed participants in the Japanese context, while the positive affect could not. Furthermore, findings from item information showed no remarkable information from MH05 and MH08. This can be explained by Cheng et al. (2006) a study arguing that positive affect was marginally contributed to the depression symptoms compared to other factors. They suggested that positive affect items can be removed as these items did not significantly affect the ability of the scale in measuring depression.

In terms of measurement invariance, this project found that there was no item exhibiting DIF. This finding was in line with the study conducted by Stommel et al. (1993), which found that males and females responded differently in the item of "crying" and "talked less" than the remaining items and did not show a different response pattern between males and females. Therefore, the measurement to compare the depression level between males and females could be trusted.

As this study uses data from IFLS-5, which was conducted by a careful and comprehended sampling procedure (Strauss et al., 2016), this finding can be generalized in Indonesian society. However, this project also had some limitations. First, this project merely assumed the unidimensional model of the scale, while some studies found that the model could be more than one, that is, two-, three-factor model and more complex higher-order models. Future studies are suggested to explore the best model for this scale to investigate validity evidence based on the scale's internal structure in the Indonesian context. Second, considering that this project just focused on the quality of the item without giving any attention to the cut-off score and there were no studies concerned with this issue, especially in the Indonesian context, further studies are encouraged to examine the cut-off of this scale. Lastly, this project examined the scale using all subgroup populations without considering ethnicity or age group. As some studies specifically focused on these factors (Baron et al., 2017; Chiesi et al., 2017), future studies may consider including these factors.

Conclusion

The findings suggest that the Mental Health Scale from ILFS-5 has good psychometric properties except for MH05 and MH08. The scale has the best performance in differentiating participants with medium to severe depressive symptoms by using all items. Lastly, there is no item exhibiting DIF, which leads to no bias measurement based on gender. Therefore, this scale is an adequate tool for measuring depressive symptoms for research purposes, especially in the Indonesian context.

Recommendation

This research merely focused on the 2-PL IRT in evaluating the Mental Health Scale. It was limited to the discrimination index, item difficulty, and other psychometric properties included in the 2-PL Polytomous IRT analysis. Also, this research only investigated the psychometric properties from the scale, resulting in exhibiting poor items which then needed to be excluded in scoring or other processes of validation, such as obtaining validity evidence based on association with other variables.

Based on the written limitations, for further research or development, response option formats should be considered. This is done in order to investigate whether the options are ordered or there is an ambiguous response option. Also, this analysis can be evidence to use a certain number of responses, such as 2, 3, or the original responses, 4.

Moreover, the use of 8-item should be investigated to get validity evidence based on association with other variables, and it should be analysed to get a norm for Indonesia's population.

Declaration

Acknowledgement

We are so grateful to the RAND Corporation for giving permission in conducting analysis of mental health scale. Also, we are thankful to Björn Andersson from the Centre of Educational Measurement at Oslo who taught us regarding the principle of item response theory model. Lastly, we appreciate Dwirifqi Kharisma Putra's suggestion on the concept of item response theory used in this paper.

Funding

This research was funded by the authors itself.

Author's Contribution

RDM contributed to write all the sections within this article. author of the scale.

Conflict of Interest

We declare that there is no conflict of interest which can influence this study.

Orcid ID

R. D. Marvianto https://orcid.org/0000-0001-8531-5519

References

Andresen, E. M., Malmgren, J. A., Carter, W. B., & Patrick, D. L. (1994). Screening for depression in well older adults: evaluation of a short form of the CES-D. *American Journal of Preventive Medicine*, *10*(2), 77–84. https://doi.org/https://doi.org/10.1016/S0749-3797(18)30622-6

Astutik, E., Hidajah, A. C., Tama, T. D., Efendi, F., & Li, C. Y. (2021). Prevalence and determinants of depressive symptoms among adults in Indonesia: A cross-sectional population-based national survey. *Nursing Forum*, *56*(1), 37–44. https://doi.org/10.1111/nuf.12508

Ayala, R. J. de. (2009). *The theory and practice of item response theory*. The Guilford Press.

Ayis, S. A., Ayerbe, L., Ashworth, M., & DA Wolfe, C. (2018). Evaluation of the Hospital Anxiety and Depression Scale (HADS) in screening stroke patients for symptoms: Item Response Theory (IRT) analysis. *Journal of Affective Disorders*, *228*(November 2017), 33–40. https://doi.org/10.1016/j.jad.2017.11.037

Baker, F. (2001). The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED458219%5Cnpapers2://p ublication/uuid/53C840DD-C92B-4719-8EC3-AF2076EDCAB3

Baron, E. C., Davies, T., & Lund, C. (2017). Validation of the 10-item Centre for Epidemiological Studies Depression Scale (CES-D-10) in Zulu, Xhosa and Afrikaans populations in South Africa. *BMC Psychiatry*, *17*(1), 1–14. https://doi.org/10.1186/s12888-016-1178-x

Berk, R. A. (2006). *Thirteen Strategies to Measure College Teaching*. Virginia: Stylus Publishing.

Björgvinsson, T., Kertz, S. J., Bigda-Peyton, J. S., McCoy, K. L., & Aderka, I. M. (2013). Psychometric properties of the CES-D-10 in a psychiatric sample. *Assessment*, *20*(4), 429–436. https://doi.org/10.1177/1073191113481998

Boey, K. W. (1999). Cross-validation of a short form of the CES-D in Chinese elderly. *International Journal of Geriatric Psychiatry*, *14*(8), 608–617. https://doi.org/10.1002/(SICI)1099-1166(199908)14:8<608::AID-GPS991>3.0.CO;2-Z

Bradley, K. L., Bagnell, A. L., & Brannen, C. L. (2010). Factorial validity of the center for epidemiological studies depression 10 in adolescents. *Issues in Mental Health Nursing*, *31*(6), 408–412. https://doi.org/10.3109/01612840903484105

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT1. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (a volume in the Multivariate Applications Series)* (pp. 307–333). Routledge.

Cao, J., & Rammohan, A. (2016). Social capital and healthy ageing in Indonesia. *BMC Public Health*, *16*(1), 1–14. https://doi.org/10.1186/s12889-016-3257-9

Chalmers, R. P. (2012). {mirt}: A Multidimensional Item Response Theory Package for the {R} Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheng, S. T., Chan, A. C. M., & Fung, H. H. (2006). Factorial structure of a short version of the Center for Epidemiologic Studies Depression Scale. *International Journal of Geriatric Psychiatry*, *21*(4), 333–336. https://doi.org/10.1002/gps.1467

Chiesi, F., Primi, C., Pigliautile, M., Ercolani, S., della Staffa, M. C., Longo, A., … Mecocci, P. (2017). The local reliability of the 15-item version of the Geriatric Depression Scale: An item response theory (IRT) study. *Journal of Psychosomatic Research*, *96*(December 2016), 84–88. https://doi.org/10.1016/j.jpsychores.2017.03.013

Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology*, *47*(5), 1105–1117. https://doi.org/10.1037/0022-3514.47.5.1105

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics*. Academic Press, Inc.

Feuerstahler, L. M., Waller, N., & MacDonald, A. (2020). *Improving Measurement Precision in Experimental Psychopathology Using Item Response Theory*. *Educational and Psychological Measurement* (Vol. 80). https://doi.org/10.1177/0013164419892049

Furr, M. R., & Bacharach, V. R. (2013). *Psychometric: An Introduction* (2nd ed.). SAGE Publisher.

Girgus, J. S., Yang, K., & Ferri, C. V. (2017). The gender difference in depression: Are elderly women at greater risk for depression than elderly men? *Geriatrics (Switzerland)*, *2*(4). https://doi.org/10.3390/geriatrics2040035

Giusti, E. M., Jonkman, A., Manzoni, G. M., Castelnuovo, G., Terwee, C. B., Roorda, L. D., & Chiarotto, A. (2020). Proposal for Improvement of the Hospital Anxiety and Depression Scale for the Assessment of Emotional Distress in Patients With Chronic Musculoskeletal Pain: A Bifactor and Item Response Theory Analysis. *Journal of Pain*, *21*(3–4), 375–389. https://doi.org/10.1016/j.jpain.2019.08.003

Gustryanti, K., Thongpat, S., & Maneerat, S. (2017). Factors Relating To Depression Among Older People Living in Cimahi, West Java Province, Indonesia. *Belitung Nursing Journal*, *3*(1), 14–22. https://doi.org/10.33546/bnj.50

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory Library*.

Hariyani, N., Bramantoro, T., Nair, R., Singh, A., & Sengupta, K. (2020). Depression symptoms and recurrent aphthous stomatitis—Evidence from a population-based study in Indonesia. *Oral Diseases*, *26*(5), 948–954. https://doi.org/10.1111/odi.13303

Haynes, W. (2013). Bonferroni Correction. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (p. 154). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_1213

Hu, L.-T., & Bentler, P. M. (1998). Fit Indice in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. *Psychological Methdos*, *3*(4), 424–453.

Hussain, M. A., Huxley, R. R., & Al Mamun, A. (2015). Multimorbidity prevalence and pattern in Indonesian adults: An exploratory study using national survey data. *BMJ Open*, *5*(12). https://doi.org/10.1136/bmjopen-2015-009810

Isaura, E. R., Chen, Y. C., Su, H. Y., & Yang, S. H. (2020). The relationship between food security status and sleep disturbance among adults: A cross-sectional study in an indonesian population. *Nutrients*, *12*(11), 1–13. https://doi.org/10.3390/nu12113411

Iwata, N., Umesue, M., Egashira, K., Hiro, H., Mizoue, T., Mishima, N., & Nagata, S. (1998). Can positive affect items be used to assess depressive disorders in the Japanese population? *Psychological Medicine*, *28*(1), 153–158. https://doi.org/10.1017/S0033291797005898

Kashiwagi, S., Tamiya, N., & Sandoval, F. (2016). Factors Associated with Depression amongst Family Caregivers Involved in Care for Community-dwelling Persons of Middle Age and Older: Based on Data from Indonesia Family Life Survey. *Public Policy and Administration Research*, *6*(5), 24–32. Retrieved from www.iiste.org

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, *11*(2), 179–188. https://doi.org/10.1007/s12564-009-9062-8

Lee, A. E. Y., & Chokkanathan, S. (2008). Factor structure of the 10-item CES-D scale among community dwelling older adults in Singapore. *International Journal of Geriatric Psychiatry*, *23*(6), 592–597. https://doi.org/10.1002/gps.1944

Leung, J., Gouda, H., Chung, J. Y. C., & Irmansyah, I. (2021). Comorbidity between depressive symptoms and chronic conditions – findings from the Indonesia Family Life Survey. *Journal of Affective Disorders*, *280*, 236–240. https://doi.org/10.1016/j.jad.2020.11.007

Liew, H. P., & Gardner, S. (2016). The interrelationship between smoking and depression in Indonesia. *Health Policy and Technology*, *5*(1), 26–31. https://doi.org/10.1016/j.hlpt.2015.10.003

Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., & Lyu, J. (2020). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of Psychiatric Research*, *126*(June 2019), 134–140. https://doi.org/10.1016/j.jpsychires.2019.08.002

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Mahwati, Y. (2017). The Relationship between Spirituality and Depression Among the Elderly in Indonesia. *Makara Journal of Health Research*, *21*(1), 13–19. https://doi.org/10.7454/msk.v21i1.6206

Nazroo, J. Y., Edwards, A. C., & Brown, G. W. (1998). Gender differences in the prevalence of depression: Artefact, alternative disorders, biology or roles? *Sociology of Health and Illness*, *20*(3), 312–330. https://doi.org/10.1111/1467-9566.00104

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometrics Theory* (third edit). McGraw-Hill. https://doi.org/10.1007/978-1-4020-9173-5_8

Olino, T. M., Yu, L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkoinis, P. A., & Lewinsohn, P. M. (2008). Measuring depression using item response theory: an examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research*, *21*(1), 76–85. https://doi.org/10.1002/mpr.1348

Peltzer, K., & Pengpid, S. (2018). High prevalence of depressive symptoms in a national sample of adults in Indonesia: Childhood adversity, sociodemographic factors and health risk behaviour. *Asian Journal of Psychiatry*, *33*(December 2017), 52–59. https://doi.org/10.1016/j.ajp.2018.03.017

Pengpid, S., Peltzer, K., & Susilowati, I. H. (2019). Cognitive functioning and associated factors in older adults: results from the Indonesian Family Life Survey-5 (IFLS-5) in 2014-2015. *Current Gerontology and Geriatrics Research*, *2019*, 23–25. https://doi.org/10.1155/2019/4527647

Purborini, N., Lee, M. B., Devi, H. M., & Chang, H. J. (2021). Associated factors of depression among young adults in Indonesia: A population-based longitudinal study. *Journal of the Formosan Medical Association*, (xxxx). https://doi.org/10.1016/j.jfma.2021.01.016

R Core Team. (2020). R: A Language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/

RAND. (2021a). Indonesian Family Life Survey (IFLS) data and documentation | RAND. Retrieved March 19, 2021, from https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/download.html

RAND. (2021b). RAND Indonesian Family Life Survey (IFLS) | RAND. Retrieved March 19, 2021, from https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS.html

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, *3*(4), 369–379. https://doi.org/10.1080/10705519609540052

Rizopoulos, D. (2006). ltm : An R Package for latent variable modeling. *Journal Of Statistical*

Software, *17*(5).

Roadolff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401.

Rosseel, Y. (2012). {lavaan}: An {R} Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Schroevers, M. J., Sanderman, R., Van Sonderen, E., & Ranchor, A. V. (2000). The evaluation of the Center for Epidemiologic Studies Depression (CES-D) scale: Depressed and Positive Affect in cancer patients and healthy reference subjects. *Quality of Life Research*, *9*(9), 1015–1029. https://doi.org/10.1023/A:1016673003237

Stommel, M., Given, B. A., Given, C. W., Kalaian, H. A., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the center for epidemiologic studies depression scale (CES-D). *Psychiatry Research*, *49*(3), 239–250. https://doi.org/10.1016/0165-1781(93)90064-N

Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herwati, Y., Witoelar, F., & Corporation, R. (2016). *The 5th Wave of the Indonesia Family Life Survey (IFLS): Overview and Field Report*. Retrieved from http://www.rand.org/content/dam/rand/pubs/working_papers/WR1100/WR1143z2/RAND_WR1143z2.pdf

University of Washington. (2021). GBD Results Tool | GHDx. Retrieved March 29, 2021, from http://ghdx.healthdata.org/gbd-results-tool

Widhowati, S. S., Chen, C. M., Chang, L. H., Lee, C. K., & Fetzer, S. (2020). Living alone, loneliness, and depressive symptoms among Indonesian older women. *Health Care for Women International*, *41*(9), 984–996. https://doi.org/10.1080/07399332.2020.1797039

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates.* Geneva: World Health Organization. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

Wu, H., Fai Cheung, S., & On Leung, S. (2019). Simple use of BIC to assess model selection uncertainty: an illustration using mediation and moderation models. *Multivariate Behavioral Research*, *55*(1), 1–16. https://doi.org/10.1080/00273171.2019.1574546

Yuniati, F., & Kamso, S. (2020). Assessing the quality of life among productive age in the general population: A cross-sectional study of family life survey in Indonesia. *Asia-Pacific Journal of Public Health*. https://doi.org/10.1177/1010539520956411

Zhang, W., O'Brien, N., Forrest, J. I., Salters, K. A., Patterson, T. L., Montaner, J. S. G., … Lima, V. D. (2012). Validating a shortened depression scale (10 item CES-D) among HIV-Positive people in British Columbia, Canada. *PLoS ONE*, *7*(7), 1–5. https://doi.org/10.1371/journal.pone.0040793