JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v14i2.46520
http://journal.uinjkt.ac.id/index.php/jp3i

PHQ-9: Validation and Investigation Response Format using Nominal Response Model

Ramadhan Dwi Marvianto[1] **& Sri Kusrohmaniah**[1]

Faculty of Psychology, Universitas Gadjah Mada, Indonesia

koes_psi@ugm.ac.id

Abstract

The Patient Health Questionnaire 9 (PHQ-9) is a widely used tool for depression screening, but its internal structure varies across different contexts and potentially leads to misinterpretations of the depression construct it measures. This study aims to investigate the internal structure and response format effectiveness of the Indonesian version of the PHQ-9. Data were collected from 1,310 participants who completed the PHQ-9 questionnaire, sourced from the Faculty of Psychology of the Universitas Gadjah Mada (UGM)'s database. Data analysis included confirmatory factor analysis (CFA), item factor analysis (IFA), and item response theory (IRT) using a nominal response model (NRM). Results indicated that a two-factor model demonstrated a better fit than a single-factor model, which was categorised as a marginal fit. Furthermore, nearly all items functioned effectively in their response format, except for items 5 (poor appetite) and 9 (suicidal thoughts), which showed suboptimal functioning in the highest categories. These findings support the practical use of the PHQ-9 and underscore the value of collapsing categories when extreme responses are rarely endorsed to improve measurement precision.

Keywords: Patient Health Questionnaire - 9, depression, Item Response Theory, Nominal Response Model

Abstrak

Patient Health Questionnaire 9 (PHQ-9) adalah instrumen yang umum digunakan untuk skrining gejala depresi, namun struktur internal dari instrumen ini bervariasi di berbagai konteks dan berpotensi menyebabkan interpretasi yang salah terhadap konstruk tersebut. Studi ini bertujuan untuk menyelidiki faktor struktur dan efektivitas format respons versi Indonesia dari PHQ-9. Data dikumpulkan dari 1.310 peserta yang mengisi kuesioner PHQ-9, yang bersumber dari basis data Fakultas Psikologi Universitas Gadjah Mada (UGM). Analisis data meliputi confirmatory factor analysis (CFA), item factor analysis (IFA), dan item response theory (IRT) menggunakan nominal response model (NRM). Hasil menunjukkan bahwa two-correlated factors model menunjukkan indeks ketepatan model yang lebih baik daripada single-factor model, yang mana tergolong dalam model fit yang marginal. Selain itu, hampir semua item berfungsi secara efektif dalam format responsnya, kecuali item 5 (nafsu makan buruk) dan 9 (pikiran bunuh diri) yang menunjukkan bahwa nilai tertinggi tidak berfungsi secara optimal. Temuan ini mendukung penggunaan praktis PHQ-9 dan menekankan pentingnya menggabungkan kategori saat respons ekstrem jarang dipilih untuk meningkatkan ketepatan hasil pengukuran.

Kata kunci: *Patient Health Questionnaire - 9, depresi, Item Response Theory, Nominal Response Model*

Introduction

According to the World Health Organization (2022), mentally healthy individuals manage daily stress, recognise their potential, work productively, and contribute to their community. The prevalence of individuals experiencing depressive symptoms reaches 27.86% in adult individuals (Purborini et al., 2021) and 22.6% (Leung et al., 2021) to 23.65% (Tama et al., 2021) in the general population based on data from Indonesian Family Life Survey and 6.1% of the total population under 15 years of age (Balitbangkes Ministry of Health RI, 2019).

Mental health problems, especially depression, not only have an incapacity effect, but also other harmful effects such as being a risk factor of suicide attempt (Chen et al., 2022; Chiang et al., 2022; Hawton et al., 2013; Paljärvi et al., 2023; Williams et al., 2022), anxiety, substance abuse, and personality disorders (Hawton et al., 2013; Reutfors et al., 2021). Sustained depression can interfere with individual cognitive performance which causes them to experience decreased productivity at work (Nafilyan et al., 2021) so that they experience difficulties in their lives such as in their economic and work lives. Given the detrimental effects, efforts are needed to overcome this depression issue.

Nevertheless, the current condition of health facilities in Indonesia is considered not to have sufficient capacity to deal with this. It was proven by the imbalance in the ratio of mental health workers to the number of people they have to support, which is 1:223,587 (WHO standard, 1:30,000) for psychiatrists (Kemenkes RI, 2022), 1:81,468 (WHO standard, 1:30,000) for clinical psychologists (Ikatan Psikolog Klinis Indonesia (IPK), 2022), and 2:462,875 (standard 25:10,000) for psychiatric nurses (Kemenkes RI, 2022). In terms of health facilities, currently only 47% Regional General Hospitals (Rumah Sakit Umum Daerah or abbreviated as RSUD) have mental services, and only 45% public health centres (Pusat Kesehatan Masyarakat or abbreviated as Puskesmas) have mental health services with trained health workers (Kemenkes RI, 2022). These data indicate that the handling of mental health in Indonesia, especially in the realm of curation (healing), is not yet optimal.

To minimise the burden, mental health prevention can be done through early detection of mental health problems. One example is screening for depressive disorders in pregnant women so that they can prevent women from conceiving and caring for children in a depressed state (Alhusseini et al., 2023; Heslin et al., 2022; Waqas et al., 2022). In addition, screening can also improve the quality of referrals (Blake, 2022) so that the most effective and efficient interventions can be obtained by individuals and further reduce the burden on mental health facilities. Indeed, the screening process needs to follow a standard measuring instrument. In the clinical context, standardised measuring instruments must meet several qualities according to Hidayat and Primasari (2011), which are validity, reliability and feasibility. The concept of validity is developed by American Educational Research Association (AERA) et al. (2014) as evidence that supports the interpretation of measurement results for a particular measurement objective. So, the use of a measurement instrument must be supported by scientific evidence, both theoretical and empirical.

According to American Educational Research Association (AERA) et al. (2014), an instrument should demonstrate five evidence of validity to ensure strong interpretation. First, it should provide evidence that its items represent the intended construct (validity based on item content). Second, it must align with the theoretical framework through a suitable factor structure (validity based on internal structure). Third, participants' mental processes during the test should match the expected outcomes (validity based on response process). Fourth, it should not favour any group within the population (validity based on test consequences). Lastly, it should correlate with other theoretically related instruments (validity based on association with other variables).

In Indonesia, one of the instruments for screening symptoms of depression is the Patient Health Questionnaire 9 (PHQ-9), which was developed Kroenke et al. (2001). This questionnaire is one part of an extensive questionnaire called the Patient Health Questionnaire (PHQ). The complete questionnaire from the PHQ itself measures several symptoms of mental health disorders based on the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders-IV), which include major depressive disorder, panic disorder, other anxiety disorders, and bulimia nervosa, as well as sub-threshold disorders such as other depressive disorder, probable alcohol abuse/dependence, somatoform and binge eating disorder. Kroenke et al. (2001) further explained that the PHQ-9 is a measurement module that focuses on diagnosing major depression with nine symptom indications, which include interest, mood, sleep difficulty, lack of energy, poor appetite, pessimism, trouble concentrating, moving/speaking, and suicidal thoughts. In their publication, Kroenke and his colleagues also conducted sensitivity and specificity tests to provide a related picture of a cut-off score that indicates whether an individual can be said to be depressed based on their response to the indications presented.

The PHQ-9 had been widely used in several countries, with satisfactory sensitivity and specificity values, such as in Kenya (Tele et al., 2023), Scotland, (Beswick et al., 2022)America (Chung et al., 2023; Mufson et al., 2022), Uganda (Kaggwa et al., 2022), Vietnam ((Le Hoang Ngoc et al., 2021), Lithuania (Pranckeviciene et al., 2022) and Peru (Smith et al., 2022)). These data indicate that the use of PHQ-9 for depression symptom screening is supported with empirical studies in those countries, that it can classify whether individuals are having depression or not based on their score. However, despite being one of the most widely used instruments for assessing depression, a systematic search using the keywords "PHQ-9" and "depression" in Scopus revealed that there are still psychometric properties issues, particularly related to the factor structure of the scale and the performance of individual items.

Research evaluating the factor structure of PHQ-9 reported different numbers of its factors. The single-factor model has been observed in countries like the Philippines, South Africa, and Vietnam, though some samples showed covariance between residues (Murray et al., 2022). Arrieta et al. (2017), using confirmatory factor analysis to evaluate the validity of the PHQ in a rural community context in Mexico, found that the instrument exhibited a one-factor structure (CFI = 0.91, NNFI = 0.88, factor loadings > 0.36). However, they also reported that a two-factor solution demonstrated similarly good fit indices. Since the two-factor solution did not substantially improve model fit, the study concluded that a single-factor structure best represents the PHQ-9. Fonseca-Pedrero et al. (2023) found that a one-factor model demonstrated strong fit among Spanish adolescents ($\chi^2(27)=203.80$, CFI = .993, TLI = .991, RMSEA = .054, SRMR = .043) and reported satisfactory reliability ($\omega = .87$). Similarly, Gómez-Gómez et al. (2023) confirmed a single-factor structure for the PHQ-9 in a large Spanish primary-care sample, with item loadings ranging from .55 to .77 and high internal consistency.

Beyond CFA, studies employing Rasch analysis also support the unidimensionality of the PHQ-9 across general and student populations. Using data from Danish primary-care patients, K. S. Christensen & Sparle-Christensen (2023) found that the PHQ-9 fit the Rasch model after minor category rescoring ($\chi^2(24)=18.16$, p = .795) with acceptable reliability (PSI = .80) and only 3.95 % significant t-tests, confirming unidimensionality. Similarly, Wilton & Horton (2020) reported that, after adjusting for item dependencies (items 1–2 and 3–4), the PHQ-9 achieved satisfactory Rasch fit, supporting a single underlying depression dimension.

In contrast, research in Latin America has indicated potential multidimensionality. A recent validation of the PHQ-9 in a non-clinical Colombian sample supported a bifactor model as the best-fitting structure for the instrument. The confirmatory factor analysis showed that this model, which includes a general depression factor and two specific first-order factors—cognitive/affective (items 1, 2, 6–9) and somatic (items 3–5)—provided an excellent fit to the data ($\chi^2(17) = 85.03$, CFI = .984, TLI = .965, RMSEA = .061, SRMR = .021) (Berrío et al., 2024). Similar to this finding, Cassiani-Miranda et al. (2016) identified a two-factor model comprising a nonsomatic/affective factor (items 1, 2, 6, and 9) and a somatic factor (items 3–5, 7, and 8), which together explained 42.8% of the total variance (KMO = 0.889).

Similar to the findings from Latin America, the Indonesian sample demonstrated a better fit for the two-factor model, distinguishing between somatic and cognitive/affective dimensions (Hall et al., 2021). This variation highlights the need for further research on the PHQ-9's factor structure, as it may impact its practical interpretation. Therefore, this problem needs to be studied and tested using scientific methods such as confirmatory factor analysis (CFA) to confirm which one is the correct model on the PHQ-9 measurement.

However, the research that investigates the factor structure of PHQ-9 in the Indonesian context is still rare. The systematic search from this study did not yet permit us to find a validity testing study, through factor analyses, in the Indonesian context. This could be due to the lack of empirical evidence to support the usefulness of PHQ-9 in clinical practice, and therefore, this measurement instrument must be re-examined empirically and theoretically. Moreover, evaluations using contemporary approaches such as IRT (item response theory) are less of a focus in the investigation of this measuring instrument. Previous studies have examined the psychometric properties of the PHQ-9 by focusing primarily on its factor structure and by using factor loadings as indicators of item performance (e.g., Arrieta et al., 2017). Such findings provide insights into how strongly each item relates to the underlying depression construct. Yet, factor loadings can offer only limited information about the functioning of individual response categories. Therefore, the PHQ-9 assessment focuses on factor structures and investigations using the IRT approach to look at psychometric properties at the test and item level, which need to be carried out. In addition, this PHQ-9 has never been evaluated in terms of answer choices. This is indicated by the absence of research that focuses on the effectiveness of the answer choices in this measuring instrument. Investigations into the functioning of response options are gaining attention, as they provide insight into whether the available categories meaningfully contribute to the raw score as intended. This perspective is critical because ineffective or redundant categories may reduce the precision of measurement and compromise the interpretability of results.

Several studies have used the Nominal Response Model (NRM) that was developed by Bock (1972, 1997), to evaluate the effectiveness and order of scores on the answer choices. This NRM model is able to see the ability of the selection of answers in differentiating the theta (θ) individuals in choosing a particular category, so that later the effectiveness of the choice of responses can be seen as done by Moulton et al. (2019) and Matlock Cole et al. (2018). In addition, NRM is also able to detect out-of-order scores of answer choices (disorder responses) from a measuring instrument, as done in research by Fujimoto et al (2018). In addition, NRM is also able to see the functioning of the mean (sometimes, in doubt, neutral) as was done in the research by DeMars & Dary Erwin (2004) and Murray et al. (2016). Despite these advantages, the application of NRM remains limited in the broader field of psychometrics and is rarely used in scale validation, including the PHQ-9.

These findings lead to a factor dilemma in the PHQ-9 structure, especially the difference of findings in Indonesia compared to other countries, which can lead to misinterpretation of the measurement results of this instrument. In addition, PHQ-9 investigations using contemporary analysis have not been carried out in Indonesia, especially at the level of response options. Therefore, this study aims to investigate factor structure and effectiveness of response options from PHQ-9 in the Indonesian sample. Investigation of factor structures will be answered by testing single-factor and two-correlated factors models, to answer the different findings in the two models and the effectiveness of the choice of responses will be seen from the analysis of the nominal response model.

The theoretical implication of this research is to strengthen the empirical foundation of the PHQ-9 factor structure in Indonesia while advancing the underexplored use of NRM in psychometric studies. By demonstrating the utility of NRM in evaluating response option functioning, this study provides an empirical illustration for psychologists and researchers of how this model can enrich instrument validation. Practically, the findings offer evidence on the suitability of NRM within Item Response Theory (IRT) frameworks and contribute to methodological learning materials that can be incorporated into psychometric courses, thereby encouraging wider application of NRM in future research and practice.

Nominal Response Model

The nominal response model was first coined by Bock (1972), which stated that the probability of an individual choosing a particular response (k) can be determined by the following equation,

$$P(X = k|\theta) = \frac{e^{(z_k)}}{\sum_{k=1}^{m} e^{(z_k)}} \quad 1$$

where k refers to a particular answer choice, θ (theta) is the individual latent score level and z_k is a linear function of theta. Furthermore, z_k it can be explained in the equation below.

$$z_k = \; a_k\theta + c_k \quad 2$$

The parameter a_k refers to the slope parameter and c_k is an intercept parameter. That is, the probability of an individual with a certain theta to respond to the k category is obtained from the exponential function of the multiplication between theta and the slope plus the intercept of the item category.

Then, Thissen et al. (2010) created a parameter that updates the parameter that Bock had previously triggered. This parameter is known as Thissen's parameterization, which is described in the following equation,

$$T(u = k|\theta, a_i^*, a_k, c_k) = T(k) = \frac{e^{(z_k)}}{\sum_{k=1}^{m} e^{(z_k)}} \quad 3$$

where,

$$z_k = a_i^* a_{k+1}^s \theta + \; c_{k+1} \quad 4$$

which $a_i^*$is the overall slope parameter, a_{k+1}^s is the scoring function of a response (k), and c_{k+1} the same intercept parameters as Bock's parameterization. However, there are some restrictions in the identification of this model, namely as follows.

$$a_1^s = 0; \; a_m^s = m - 1; \; c_1 = 0 \quad 5$$

The implication is that items that have four categories will have a scoring function and intercept value of 0 for the first category ($a_1^s = 0$), have two values that vary in the second and third categories, while they will have a constant value of 3 ($m = 4$; $a_4^s = 4 - 1$), have an intercept value of 0 in the first category ($c_1 = 0$), and other intercepts have their own variations in value. Thus, this study will use Thissen's parameterisation in assessing PHQ-9.

Methods

Participants

This study utilized one dataset from research Marvianto and Widhiarso (2018) and two additional datasets from the Mental Health Service Unit at Universitas Gadjah Mada, encompassing 1,310 individuals who completed the PHQ-9. The sample was predominantly female (66%; 867 subjects), with 81 subjects (6%) not disclosing their gender. The average age of participants was 22 years (SD=9), primarily within the 13-20 age range, though 122 subjects (9%) did not provide age information. Regarding depression severity, responses were categorized as minimal (≤4; 12%), mild (5-9; 26%), moderate (10-14; 21%), moderately severe (15-19; 19%), and severe (20-27; 21%; see Table 1).

Table 1. Summary of demographic characteristics of research subjects (n = 1,310)

Demographic	Amount	Percentage
Gender		
Man	361	28%
Woman	867	66%
Don't Want to Disclose	1	0%
missing	81	6%
Age		
13-20 years	590	45%
21-30 years	433	33%
31-40 years	104	8%
41-50 years	32	2%
> 50 years	29	2%
missing	122	9%
Category		
At a minimum	157	12%
Mild	343	26%
Moderate	281	21%
Moderately severe	249	19%
Severe	280	21%

Sources: Personal Data (2025).

Intruments

This study used secondary data and employed the Patient Health Questionnaire-9 (PHQ-9) as the instrument administered to participants. The PHQ-9 is a screening measure of depressive symptoms developed by Kroenke et al. (2001). It comprises nine items that ask participants about indications of major depression over the past two weeks. The indicators are based on the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) and include loss of interest, depressed mood, sleep difficulties, low energy, poor appetite, pessimism, trouble concentrating, psychomotor changes (moving/speaking), and suicidal thoughts. Participants rated their experiences during the past two weeks on a 4-point Likert scale (0 = not at all or rarely; 1 = several days; 2 = more than half the time; 3 = nearly every day).

Data Collection

This study did not collect primary data; instead, it employed secondary data obtained from a previous study and from a unit within Universitas Gadjah Mada. The study received ethics approval from the Ethics Committee of the Faculty of Psychology, Universitas Gadjah Mada (No. 6512/UN1/FPSi.1.3/SD/PT.01.04/2023). Furthermore, use of the secondary data was authorized by the original researchers and the unit mentioned above.

Data Analysis

This study uses the Confirmatory Factor Analysis (CFA) method to investigate factor structures that underlie the PHQ-9 measurement. Prior to that, this research also provides descriptive statistics of each item as well as Pearson’s correlation matrix. Then, this research also tries to do Item Factor Analysis (IFA) to compare measurement models when they are treated as interval, categorical, or nominal data. The two methods are also one of the IRT assumptions, namely the unidimensionality assumption.

Furthermore, this study conducted a Nominal Response Model (NRM) analysis by first testing the assumption of local independence. The parameters obtained from the NRM estimation use Thiessen's parameterisation so that the overall slope, scoring function, and intercept are obtained, which will be processed into Category Boundaries Discriminations (CBDs) and several other output features of IRT.

This study uses the RStudio software along with the supporting packages. CFA analysis is typically performed using the Lavaan package (Rosseel, 2012). Then, descriptive statistical analysis and Pearson’s correlation matrix were performed using the package base (R Core Team, 2025), and NRM analysis using the package mirt (Chalmers, 2012). Meanwhile, IFA analysis was performed using the software plus (Muthén & Muthén, 2011) and the robust maximum likelihood (MLR) estimator.

Results and Discussion

Descriptive Statistic

Appendix 1 presents descriptive statistics for each PHQ-9 item. Generally, response choice 1 (several days) was most frequently selected. While responses were varied for most items, items 8 (worry) and 9 (suicidal thoughts) showed lower proportions. Univariate normality was acceptable, as skewness values did not exceed ±2 (Kim, 2013). Nevertheless, as Finney and DiStefano (2013) emphasise, even when observed variables approximate normality, data from psychological instruments are often ordinal and prone to distributional deviations. To minimise potential bias in parameter estimates, standard errors, and chi-square statistics, this study employed the robust maximum likelihood estimator (MLR), which provides more reliable results under both normal and non-normal conditions.

Item Correlation Matrix

Furthermore, item correlation matrix analysis was performed using the Pearson formula, which shows that the correlation between items formed is a positive correlation ranging from .402 to .782 (see Appendix 2). That is, all items can be said to be related to one another. Furthermore, the direction of the relationship, which is entirely positive on all items, has the potential to create a common factor that underlies the nine points. In addition, the correlation of each item with the total score also shows a fairly high value and has a positive direction.

Confirmatory Factor Analysis

This study implemented two measurement models: (1) a single-factor model where the nine PHQ items represent one latent construct, depression, and (2) a two-correlated factors model, with somatic factors for items 3 to 5 and cognitive/affective factors for items 1, 2, and 6 to 9. Confirmatory factor analysis (CFA) was conducted, treating items as interval data (models A and B), categorical data (models C and D), and nominal data (models E and F) to assess the model fit. Meanwhile, the statistical summary and goodness of fit index can be seen in Table 2 and a summary of the model can be seen in Figure 1.

Table 2. Summary of statistical tests and goodness of fit index

Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	AIC
A	242.351	36	< .001	.952	.936	.087	.036	28,791.433
B	166.039	26	< .001	.969	.957	.071	.029	28,698.411
C	-	-	-	-	-	-	-	26,947.258
D	-	-	-	-	-	-	-	26,855.133
E	-	-	-	-	-	-	-	26,898.926
F	-	-	-	-	-	-	-	26,809.551

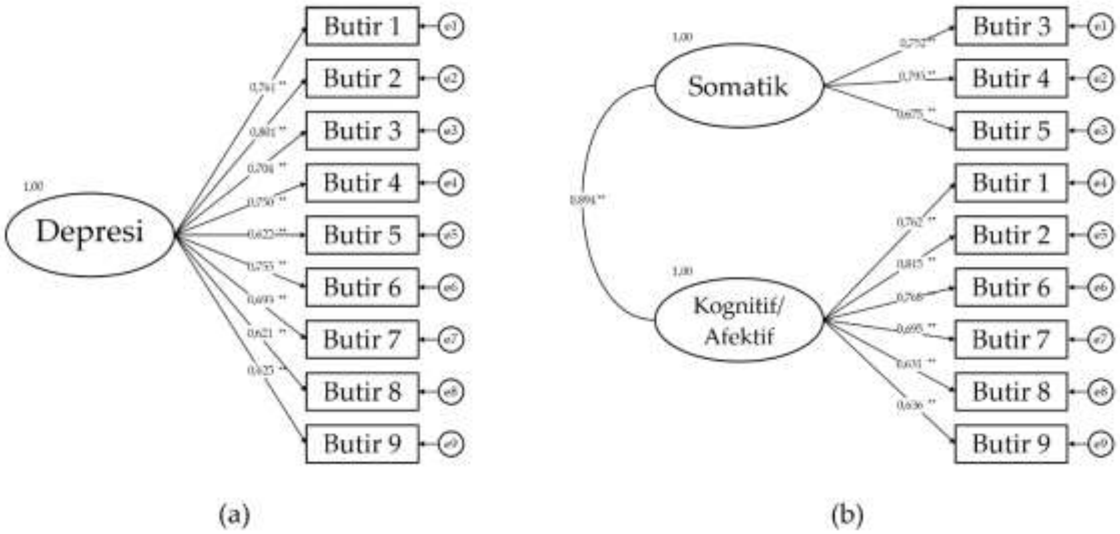
Notes. χ^2 = *Chi-squared*; df = *degree of freedom*; p = *p-value of chi-square test*; CFI = *Comparative Fit Index*; TLI = *Tucker-Lewis Index*; RMSEA = *Root Mean Square of Error Approximation*; SRMR = *Standardized Root Mean Square of Residual*; BIC = *Akaike Information Criteria*.

Sources: Personal Data (2025).

Analysis using a robust maximum likelihood (MLR) estimator through CFA shows that the single-factor model (model A) has a satisfactory index value, namely CFI above the critical value 0.95 with a value of .952 (Hu & Bentler, 1999) and SRMR below the critical value .05 (Hu & Bentler, 1999). On the other hand, this model has a TLI which is below the critical limit value but still in the range above the critical value .90 (Bentler & Bonett, 1980), which is considered an acceptable fit. Meanwhile, the RMSEA value is also below the critical value .05 or .08 (Hu & Bentler, 1999), but a value below .10 (Browne & Cudeck, 1992) can still be said to be a mediocre fit. Thus, model A is a marginal fit model and is evidence of the unidimensionality assumption to perform IRT analysis.

On the other hand, the two-correlated factors model (model B) shows CFI and TLI values above the critical value of .95 and SRMR below the critical value of .05, which indicates that this model fits with the data. Then, the RMSEA in this model is at a tolerable value, which is below 0.08 (Hu & Bentler, 1999), so that it can be said that this model is fit with the data. Compared to model A, model B has a smaller AIC value (28,698.411) than model A (28,791.433), which means that model B is a better fit than model A. However, the correlation parameter between factors shows a relatively large value (see Figure 1), which is .894, indicating that the two factors are similar or based on a higher factor.

Then, the results of the analysis by treating the data as categorical and nominal show that the AIC value in the categorical model ($AIC_C = 26{,}947.258$; $AIC_D = 26{,}855.133$) has a greater value than the nominal ($AIC_E = 26{,}898.926$; $AIC_F = 26{,}809.551$). This means that the analysis using the nominal data model is a better fit than the categorical and interval models. Thus, this supports the use of ordinal data-based IRT analysis, particularly analysis using nominal data using the NRM model.



Sources: Personal data (2025).

Figure 1. PHQ-9 measurement model with (a) single-factor model and (b) two-correlated factors model.

Local Independence Assumption

Local Independence assumption is tested using the correlation method between residues of each item called Q3 (Yen, 1984). The results of the analysis show that the correlation between the residuals is negative and close to 0 (see Appendix 3). That is, the residuals of each item only have a very small or even negligible correlation. In addition, Christensen et al. (2017) state that a Q3 value that is below .20 indicates the absence of local dependence. Thus, all items in the PHQ-9 measurement are locally independent of one another.

Model Comparison

Before using NRM, this study tried to explore the suitability of the data with several other models, namely the Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM). Testing the accuracy of the model using the C2 method was developed by Cai and Monroe (2014). This method produces C2 values along with the degree of freedom values and their significance, CFI, TLI, and RMSEA.

Table 3 shows that all three models demonstrated satisfactory accuracy (CFI and TLI > .95; Hu & Bentler, 1999). However, these conventional cutoffs were initially developed for continuous data and may not be entirely appropriate for categorical indicators. Recent studies highlight that for ordinal data, model evaluation should rely more on the C2 statistic, which provides RMSEA2 specifically suited for limited-information estimators (Maydeu-Olivares & Joe, 2014). Moreover, Xia & Yang (2019) caution that applying traditional cutoffs to DWLS or ULS analyses often leads to overoptimistic conclusions. In this study, the NRM model yielded the lowest C2 value and an

RMSEA2 of .06, well below the tolerance limit of .08, thereby supporting NRM as the best-fitting model.

Table 3. Summary of goodness-of-fit indices and statistics on the GPCM, GRM, and NRM models using the C2 statistic

Model	C2	df	p-value	CFI	TLI	RMSEA2
GPCM	290.762	27	< .001	.978	.970	.086
GRM	291.411	27	< .001	.978	.970	.086
NRM	75.310	9	< .001	.994	.978	.075

Notes. C2 = *C2 type-statistic* (Cai & Monroe, 2014); df = *degree of freedom*; p = *p-value of chi-square test*; CFI = *Comparative Fit Index*; TLI = *Tucker-Lewis Index*; RMSEA = *Root Mean Square Error Approximation*.

Sources: Personal Data (2025).

Nominal Response Model

NRM analysis produces several parameters, namely discriminatory power or overall item slope (a), scoring function for each response choice (ak_o... ak_3), and intercept (d_o... d_3). Overall slope value and scoring function are then processed into a scoring weight for each item. Furthermore, the difference from scoring weights 1- 0, 2-1, and 3-2 is categorised as category discrimination boundaries or CBD (a_1... a_3).

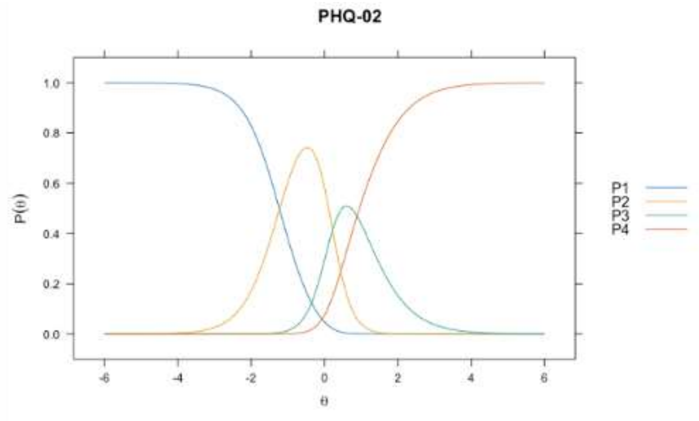
Then, the results of the analysis show that five items from PHQ-9 (Items 3, 5, 7, 8, and 9) have a discriminating power (a) which is classified as moderate according to Baker and Kim (2017) because it is between .650 and 1.340 (See Table 4). In addition, items 4 and 6 show that the power of discrimination on this item is relatively high, namely between 1.340 and 1.690. Meanwhile, the remaining two items, namely items 1 and 2, include items that have very high discriminatory power (> 1.70).

Furthermore, these a values are used to obtain a scoring weight for each answer choice. For example, in point 2, the scoring weight of choice 0 (never) will be obtained from the following calculation results: a where , ak_0 this value is constant 0, so that the first-choice value for all items will be constant at 0. Meanwhile, the scoring weight of the choice of the two values will change according to the scoring function for each option.

For example, for item 2, response choice 2, which has a value $ak_1$of .944, will be multiplied by $a$that item of 2.227 to obtain a scoring weight of 2.102. In addition, for response option 3 (more than half of the time referred to), having a value $a$of 2.227 will be multiplied by a value ak_2 of 2.360, which results in a scoring weight of 5.256. Finally, response choice 4 (almost every day), scoring value function (ak_3) is three multiplied by the $a$item value of 2.227, so that it produces a scoring weight worth 6.681.

After that, the scoring value weight on each item is processed to see Category Discrimination Boundaries (CBDs). As an example, point 2, the first CBD value, namely the discriminatory power value between response options 1 and 2 (a_1) of 2.102, is obtained from the results of the reduced scoring weight of choice 2 (2.038) minus scoring weight choice 1 (0). A similar process was also carried out for the second (a_2) and third (a_3) CBD, each of which was obtained from reduced scoring weight 3 (5.256) with 2 (2.102) and 4 (6.681) with 3 (5.256) resulting $a_2$in 3.153 and $a_3$1.425.

Then, in terms of CBDs, it was found that almost all of the CBD had a value greater than 1. This means that the choice of response k with k-1 can distinguish specific theta opportunities in choosing that response. For example, in the PHQ-05 item, the probability of an individual with a specific theta in a category will be illustrated as shown in Figure 2.



Sources: Personal data (2025).

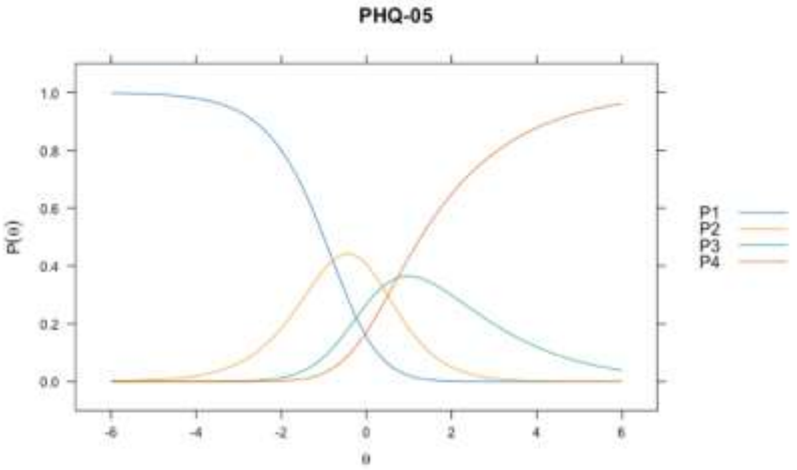
Figure 2. Item characteristic curve pada PHQ-02

To see the meaning of CBD between categories 4 and 3 ($a_3 = 1.425$) in point 2, a simulation can be carried out based on Figure 2. For example, an individual with a theta of 0.1 will have the opportunity to answer category four by 10% and category three by 32 % whereas an individual with a theta of .2 will have the chance to choose category four by 14% and category three by 40%. These two examples show that response choices 4 and 3 have quite different chances of being answered by a specific theta, namely, .1 and .2. However, when theta is between the crossing of the category 3 (green) and category 4 (orange) lines, or category the threshold between categories 3 and 4 (δ_3) is .886 or equivalent to .9, so the chances of the individual answering categories 3 and 4 are 47% and 48%, respectively.

Then, as a comparison, a_3 $(.607)$ point 5 can be simulated based on Figure 3. Individuals with a theta of 0.1 will have a 29% and 20% chance of choosing categories 3 and 4, respectively. If the individual's theta rises to .2, then the odds are 30% and 22% for categories 3 and 4, respectively. This shows that the difference in probability between categories 3 and 4 for a particular theta is not different, as shown in Figure 2. Figure 3 on the green and orange lines in the theta area 0 to .2. Meanwhile, if you use the example of a theta of 2, you will find that the probability of choosing categories 3 and 4 is 30% and 65%, the same as the ICC shown in Figure 3 where the orange line at theta 2 is far above the green line at the same theta.

Furthermore, the option response function (ORF) for each item can be seen in Appendix 4, which provides a general illustration regarding the functioning of CBDs in each item based on the curve in each category.

In terms of test information, Appendix 5 shows that all items in general can optimally provide information to individuals who have a theta between -1 and +1 or individuals with moderate abilities. Nonetheless, some items are more inclined to provide information to individuals who have high abilities above +1, such as items PHQ-08 and PHQ-9, where the peak points are above +1.



Sources: Personal data (2025).

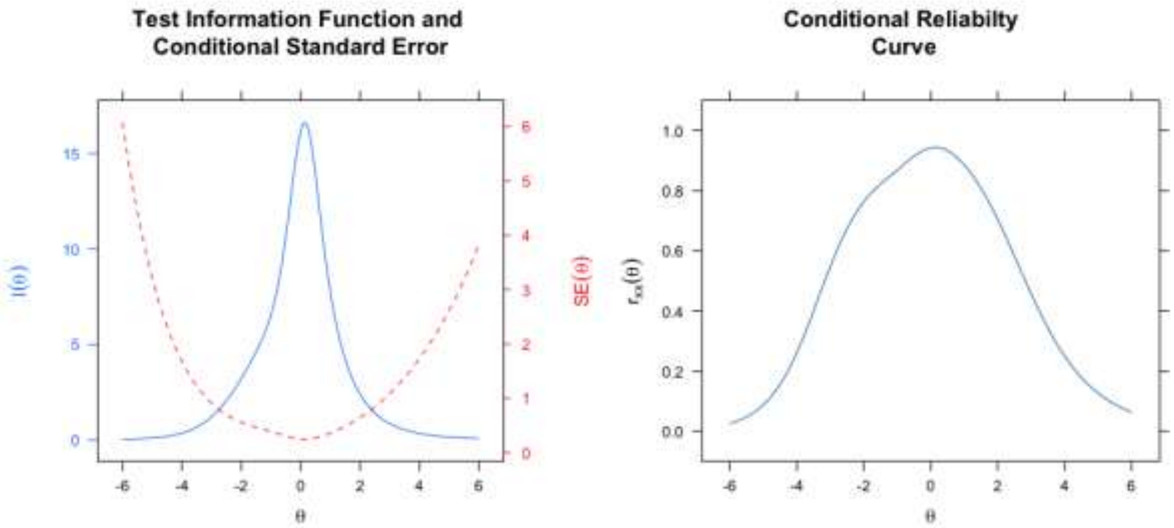
Figure 3. Item characteristic curve on the PHQ-05

Item Fit

Based on Appendix 6, statistical values $S-\chi^2$ using the package mirror uncorrected and with False Discovery Rate correction (FDR; Benjamini & Hochberg, 1995) indicate that there are slight differences. The fit items resulting from the uncorrected package show that almost all items fit with the data (p > .05) except for item PHQ-08. This does not indicate that item 8 is not a valid measure of the construct. However, this means that NRM cannot model the responses from the existing data, so another model is recommended for this item.

On the other hand, using the FDR correction, it was found that all items fit the model. That is, the NRM model can explain the responses in the field data obtained.

Precision of Measurement



Sources: Personal data (2025).

Figure 4. Test information function and conditional standard error (left) and conditional reliability curve (right)

The information value in each item can then be processed to obtain a test information function curve (see Figure 4 on the left), which gives meaning regarding the accuracy of measurements on the continuum capability or theta. From the curve, it can be seen that the peak point of the information is above 15, or more precisely 16.638 at a theta of .14, which means that this measurement can provide optimal information on theta at that figure. The theta, then, is calculated to be a conditional reliability curve (see **Figure 4** on the right), which shows a measurement reliability above .70 at theta -1.96 to 1.69 with I(θ) above 3.334. This means that this measurement has a satisfactory level of reliability (over .70; Nunnally & Bernstein, 1994) when measuring individuals with abilities between -1.96 and +1.69.

Table 4. Summary of analyses results using mirt

Item	a	Scoring function				Intercept				Scoring Weight				Category Discrimination Boundary		
		ak_0	ak_1	ak_2	ak_3	d_0	d_1	d_2	d_3	ak_0	ak_1	ak_2	ak_3	a_1	a_2	a_3
1	1.858	0	.829	2.379	3	0	2.938	2.290	1.336	0	1.540	4.420	5.574	1.540	2.880	1.154
2	2.227	0	.944	2.360	3	0	2.521	1.691	.345	0	2.102	5.256	6.681	2.102	3.153	1.425
3	1.316	0	.889	2.124	3	0	1.766	1.585	1.498	0	1.170	2.795	3.948	1.170	1.625	1.153
4	1.714	0	.878	2.335	3	0	2.666	2.694	2.176	0	1.505	4.002	5.142	1.505	2.497	1.140
5	0.972	0	1.247	2.375	3	0	.958	0.543	.101	0	1.212	2.308	2.916	1.212	1.096	.607
6	1.580	0	.902	1.911	3	0	1.427	1.326	.927	0	1.425	3.019	4.740	1.425	1.594	1.721
7	1.181	0	.914	2.272	3	0	.827	0.169	-.380	0	1.079	2.683	3.543	1.079	1.604	.860
8	1.074	0	.887	2.099	3	0	.083	-0.950	-2.072	0	.953	2.254	3.222	.953	1.302	.968
9	1.070	0	1.470	2.420	3	0	-1.060	-2.187	-3.459	0	1.573	2.589	3.210	1.573	1.016	.621

Catatan. a = *slopes*; $ak_o \ldots ak_3$ = *scoring coefficient*; $d_o \ldots d_3$ = *intercept*; response format 1 = tidak pernah; 2 = beberapa hari; 3 = lebih dari separuh waktu yang dimaksud, 4 = hampir setiap hari; $a_1 \ldots a_3$ = *category boundary discriminations*.

Sources: Personal Data (2025).

Then, the estimation of the reliability of this measurement can also be done using the marginal reliability formula. The results of the analysis show that the marginal reliability value of this measurement is 0.884, which can be said to be satisfactory because it is above the standard reliability value for a measurement, which is .70.

This study seeks to investigate the factor model structure of the PHQ-9 measurement by creating single-factor and two-correlated factor models in several data type treatments. As a result, this study shows that the two-correlated factors model has a better goodness of fit index compared to the single-factor model. This means that a model with two factors, namely somatic and cognitive/affective, is considered more appropriate than a model that only measures depression in the context of the Indonesian sample.

The finding that the unidimensional model is classified as a marginal fit is also found in several countries. Meanwhile, the research by Murray et al (2022) showed that PHQ-9 is a good fit in a single-factor model in other countries such as Ghana, Romania and Jamaica. In the Ghana sample, the initial model had CFI, TLI, RMSEA and SRMR values that did not meet the critical values, so the covariance between the residuals of item 3 with items 4 and 6 was carried out. Similar findings also occurred in the Romanian and Jamaican samples, which required a residual covariance as in the Ghana sample; however, a single-factor fit model was not produced in these two samples. Not only that, but the research also conducted by Rahman et al. (2022) supports that the single-factor model was initially not fit with the data, so modifications were made. Like previous studies, modifications were made by providing covariance between residues in items 3-6, 6-9, 6-4, 6-9, and 3-9 so that the single-factor model has a more satisfactory index than the two-factor model. This method of covariance between residues is also carried out to modify the single-factor model to meet the critical value of this measurement (Yu et al., 2012).

Besides the need for modification because the single-factor model is not fit, the two-correlated factor model is also found to have a better index. That is, with a model in which somatic factors are represented by items 3 to 5 and cognitive/affective factors are represented by items 1, 2, and 6 to 9, a more fit model is obtained compared to depression, which is represented by the nine items. The two-correlated factor model, which is more fit than the single-factor model, is also supported by many previous studies (Bi et al., 2021; Chilcot et al., 2013; Elhai et al., 2012; Familiar et al., 2015; Krause et al., 2008, 2010; Vu et al., 2022; Wang et al., 2023). Regarding the implications of the two-factor model, the .893 correlation value between factors obtained in this study is relatively strong (strong correlation; .70-.89; Schober & Schwarte, 2018). These results were also found in studies that found the two-factor model to be more fit because the correlation between factors was found to have a relatively strong value, namely > .70 (Bi et al., 2021; Cassiani-Miranda & Scoppetta, 2018; Elhai et al., 2012; Familiar et al., 2015; Krause et al., 2008) and moderate, namely 0.30-0.70 (Chilcot et al., 2013; Vu et al., 2022). This high correlation allows for the potential for higher factors underlying these two factors.

Before examining the implications of high correlation between factors, several studies have shown that the single-factor model is also fit to the data without modification (López-Guerra et al., 2022; Wang et al., 2023; Yu et al., 2012)and marginal fit, or there are one or two indices that do not meet the criteria (Elhai et al., 2012; López-Guerra et al., 2022). Then, due to the finding that the correlation between factors is high, López-Guerra et al. (2022) investigated a model that tries to explain the existence of a significant factor that underlies the PHQ-9 measurement after it was found that the two-factor model was a better fit than the single-factor model. In his research, he tested nested bi-factor models with single-factor models and two correlated factors. This bi-factor model then shows more fit results than the two nested models. In addition, in this study, it was found that factor loading for the depression factor is more dominant than factor loading for the somatic and cognitive factors. In addition, in the discussion, this study also explains that the use of the multidimensional model is less needed, so that unidimensional models can still be used.

Furthermore, the IRT analysis in this study has been supported by the assumption of unidimensionality obtained from the marginal CFA results fit without covariance between residuals, and local test results dependence shows that each item is in a local condition of independence from other items, which strengthens the model of using IRT analysis. The results of the IRT analysis show that the NRM model has the best statistics and goodness of fit index compared to the GPCM and GRM models. This finding is also supported by CFA findings, which reveal that IFA analysis treating items as ordinal data has the smallest AIC value compared to analysis treating items as interval and ordinal data. This has implications for the finding that the use of nominal data types is more suitable for viewing the order of answer choices and not forcing the data to have a particular order of options, which order of options has become a characteristic of the GPCM and GRM models (de Ayala, 2022).

The results of the NRM analysis show that, in general (global fit), the NRM model can explain the existing data with the output parameters. First, the overall parameters, the slope or discrimination parameter, indicate that all items have values that are classified as good, namely moderate (items 3, 5, 7, 8, and 9), high (item 3) and very high (items 1 and 2). This means that all of these items, in general, can distinguish individuals with low and high theta in responding to all of these items.

In addition, the NRM model that is formed can produce slope and intercept parameters that are able to explain individual opportunities to choose response options based on the individual's theta. This opportunity then underlies the Option Response Function (ORF) of each item. ORF can explain the value of CBDs obtained by each item. The results showed that most of the items had CBDs that exceeded 1 for each k response choice with k-1 response choices, which meant that the choice of these options could differentiate individuals with higher and lower theta. This is similar to the findings of studies using NRM in evaluating the effectiveness of response choices where CBD values exceeding 1 indicate the ability of response choices to discriminate individuals (Preston et al., 2011; Reise et al., 2021a, 2021b) theta. In other words, the response options in this PHQ-9 as a whole function effectively.

However, several CBDs were found that were quite far below 1, namely a_3 in point 5 (poor appetite) and a_3 in point 9 (suicidal thoughts), where both values were 0.60. This low CBD makes the chances of individuals with a theta not much different in choosing option 4 or 3, so that the two choices are not effective for differentiating individual theta. Therefore, it is recommended to combine response choices 4 with response choices 3. This method was also carried out in previous studies, such as research conducted by Reise et al. (2021a), who simplified the scoring measuring instrument, which initially had three answer choices and was changed to two answer choices (dichotomous responses).

Based on the findings and practices carried out in previous research, this study provides recommendations in terms of presentation and scoring, namely the presentation of items 1 (interest), 2 (mood), 3 (sleep difficulty), 4 (lack of energy), 6 (pessimism), 7 (trouble concentrating), and 8 (moving/shooting) can use the same four answer options as practice in general with the same scoring procedure (0 = never or rarely; 1 = several days; 2 = more than half of the time prescribed; 3 = nearly every day). Then, for items 5 (poor appetite) and 9 (suicidal thought), presentation can be made using the same four answer choices as practice in general, with the scoring changed, namely by changing the score 3 (almost every day) to 2, also because the two items are not effectively able to distinguish individual abilities. Alternatively, items 5 and 9 can be presented with three possible answers that read “never or rarely”, “several times”, and “almost every day” with scores ranging from 0 to 1, respectively. A summary of these recommendations can be seen in **Table 5**.

This study also found that PHQ-9 measurements can provide optimal information on individual theta, which ranges from -1.96 to 1.69. This means that the measurement on theta has a low (Nunnally & Bernstein, 1994) error value, and the measurement results are at a satisfactory reliability value, which is above .70. From the conditional reliability, one reliability value is obtained for the PHQ-9 measurement, namely marginal reliability. The findings of this study indicate that the marginal reliability value is classified as satisfactory (> .70).

However, only simple structure factor investigations were carried out, namely the single-factor model and two correlated factors, so this research was limited to the two models. However, other modelling is still possible to do in investigating the most fit factor structures for measuring PHQ-9, such as the two-factor model with items that are different in terms of factor representation. In addition, more complex models such as second-order factor or bi-factor can be explored more deeply for PHQ-9 and its contemporary analysis. In addition, this study only uses the GPCM and NRM models as a comparison for the NRM model, so that it is still possible to have a better model to explain the data from this measurement, and future research is expected to include other models that are characteristically suitable to explain PHQ-9 measurements. Furthermore, although the results of this study indicate that most of the items in PHQ-9 have effective response options, there are findings that items 5 (poor appetite) and 9 (suicidal thought) have inadequate response effectiveness for answer choices 4 and 3, so that they theoretically can be combined into one response, i.e. almost every day. In addition, this study also provides recommendations for scoring and presentation in the discussion section, but these recommendations also need to be reviewed to ensure the functioning of the answer choices in accordance with the recommendations provided. Then, this study only discusses the factor structures of the PHQ-9 to look for the factor structures so that it only gets validity evidence based on internal structures so that future research is advised to examine other validity evidence such as exploring the relationship between PHQ-9 and other construct measurements such as quality life, personality, or maladaptive personality and other constructs that are theoretically proven to be related.

Table 5. Recommendations for serving and scoring the PHQ-9 based on the results of the NRM analysis

Item	Response format and scoring
1, 2, 3, 4, 6, 7, and 8	- Score 0: tidak pernah atau jarang - Score 1: beberapa kali - Score 2: lebih dari separuh waktu yang ditentukan - Score 3: hampir setiap hari.
5 and 9	- Score 0: tidak pernah atau jarang - Score 1: beberapa kali - Score 2: lebih dari separuh waktu yang ditentukan - Score 2: hampir setiap hari

Sources: Personal Data (2025).

Conclusion

Based on various findings from the results of the analysis conducted, this study concluded that the single-factor model on the PHQ-9 measurement is classified as a marginal fit. However, in terms of data, the two-factors model with somatic factors contains items 3 (sleep difficulty), 4 (lack of energy) and 5 (poor appetite) as well as cognitive/affective factors which include items 1 (interest), item 2 (mood).), item 6 (pessimism), item 7 (trouble concentrating), item 8 (moving/striking) and item 9 (suicidal thought) are classified as better fit than the single-factor model. Meanwhile, the NRM results show that the NRM model fits the data and results in the conclusion that almost all items have answer choices that function effectively in differentiating individual theta. However, 2 CBD were found in categories 4 and 3 (a_3) in items 5 and 9, which had low scores, so it was said that both items did not have good discrimination. In addition, the NRM results also show optimal measurement precision in the range of theta -1.96 or raw-score 2 to theta 1.69 or raw-score 24 with marginal reliability of measurement at .884. Therefore, the use of PHQ-9 and its scoring could be supported by this evidence.

Acknowledgment

The researcher would like to thank MSU and RH for the constructive inputs in the process of working on this thesis. In addition, the researchers would like to thank the units and centres working in the field of mental health at the Faculty of Psychology UGM for their assistance in granting permission to researchers to study their datasets. In addition, the researchers also thank WW, who has helped and stimulated researchers to always study psychometrics, and also MDKP, who has agreed to become a research mentor to study analytical models in psychometrics more deeply.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors Contribution

RDM: Conceptualisation; Methodology; Formal analysis; Validation; Visualisation. SK: Writing – review & editing; Supervision. All authors read and approved the final manuscript.

References

Alhusseini, N., Farhan, H., Yaseen, L., Abid, S., Imad, S. S., & Ramadan, M. (2023). Premarital mental health screening among the Saudi population. *Journal of Taibah University Medical Sciences*, *18*(1), 154–161. https://doi.org/10.1016/j.jtumed.2022.06.013

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baker, F. B., & Kim, S.-H. (2017). The Basics of Item Response Theory Using R. Dalam *Measurement: Interdisciplinary Research and Perspectives* (Vol. 16, Nomor 3). Springer. https://doi.org/10.1080/15366367.2018.1462078

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Dalam *Source: Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 57, Nomor 1).

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Beswick, E., Quigley, S., Macdonald, P., Patrick, S., Colville, S., Chandran, S., & Connick, P. (2022). The Patient Health Questionnaire (PHQ-9) as a tool to screen for depression in people with multiple sclerosis: a cross-sectional validation study. *BMC Psychology*, *10*(1). https://doi.org/10.1186/s40359-022-00949-8

Bi, Y., Wang, L., Cao, C., Fang, R., Li, G., Liu, P., Luo, S., Yang, H., & Hall, B. J. (2021). The factor structure of major depressive symptoms in a sample of Chinese earthquake survivors. *BMC Psychiatry*, *21*(1). https://doi.org/10.1186/s12888-020-02993-3

Blake, C. (2022). Depression Screening Implementation: Quality Improvement Project in a Primary Care Clinic for First Responders. *Workplace Health and Safety*, *70*(12), 543–550. https://doi.org/10.1177/21650799221119147

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Bock, R. D. (1997). The Nominal Categories Model. Dalam W. J. van der Linden & R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory* (hlm. 33–49). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_2

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods &amp; Research*, *21*(2), 230–258. https://EconPapers.repec.org/RePEc:sae:somere:v:21:y:1992:i:2:p:230-258

Cai, L., & Monroe, S. L. (2014). A New Statistic for Evaluating Item Response Theory Models for Ordinal Data. CRESST Report 839. *National Center for Research on Evaluation, Standards, and Student Testing*.

Cassiani-Miranda, C. A., & Scoppetta, O. (2018). Factorial structure of the Patient Health Questionnaire-9 as a depression screening instrument for university students in Cartagena, Colombia. *Psychiatry Research*, *269*, 425–429. https://doi.org/10.1016/j.psychres.2018.08.071

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

Chen, X., Mo, Q., Yu, B., Bai, X., Jia, C., Zhou, L., & Ma, Z. (2022). Hierarchical and nested associations of suicide with marriage, social support, quality of life, and depression among the elderly in rural China: Machine learning of psychological autopsy data. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.1000026

Chiang, Y.-H., Ma, Y.-C., Lin, Y.-C., Jiang, J.-L., Wu, M.-H., & Chiang, K.-C. (2022). The Relationship between Depressive Symptoms, Rumination, and Suicide Ideation in Patients with Depression. *International Journal of Environmental Research and Public Health*, *19*(21). https://doi.org/10.3390/ijerph192114492

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., Sykes, N., Hansford, P., & Hotopf, M. (2013). The factor structure of the PHQ-9 in palliative care. *Journal of Psychosomatic Research*, *75*(1), 60–64. https://doi.org/10.1016/j.jpsychores.2012.12.012

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Chung, T. H., Hanley, K., Le, Y. C., Merchant, A., Nascimento, F., De Figueiredo, J. M., Wilcox, H. C., Coryell, W. H., Soares, J. C., & Selvaraj, S. (2023). A validation study of PHQ-9 suicide item with the Columbia Suicide Severity Rating Scale in outpatients with mood disorders at National Network of Depression Centers. *Journal of Affective Disorders*, *320*, 590–594. https://doi.org/10.1016/j.jad.2022.09.131

de Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory* (Second edition). The Guilford Press.

DeMars, C. E., & Dary Erwin, T. (2004). SCORING NEUTRAL OR UNSURE ON AN IDENTITY DEVELOPMENT INSTRUMENT FOR HIGHER EDUCATION. Dalam *Research in Higher Education* (Vol. 45, Nomor 1).

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., Chan, P. K., Slembarski, R., Liberzon, I., Galea, S., & Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, *199*(3), 169–173. https://doi.org/10.1016/j.psychres.2012.05.018

Familiar, I., Ortiz-Panozo, E., Hall, B., Vieitez, I., Romieu, I., Lopez-Ridaura, R., & Lajous, M. (2015). Factor structure of the Spanish version of the patient health questionnaire-9 in Mexican women. *International Journal of Methods in Psychiatric Research*, *24*(1), 74–82. https://doi.org/10.1002/mpr.1461

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. Dalam *Structural equation modeling: A second course, 2nd ed.* (hlm. 439–492). IAP Information Age Publishing.

Fujimoto, K. A., Gordon, R. A., Peng, F., & Hofer, K. G. (2018). Examining the Category Functioning of the ECERS-R Across Eight Data Sets. *AERA Open*, *4*(1). https://doi.org/10.1177/2332858418758299

Hall, B. J., Patel, A., Lao, L., Liem, A., Mayawati, E. H., & Tjipto, S. (2021). Structural validation of The Patient Health Questionnaire-9 (PHQ-9) among Filipina and Indonesian female migrant domestic workers in Macao: STRUCTURAL VALIDATION OF PHQ-9. *Psychiatry Research*, *295*. https://doi.org/10.1016/j.psychres.2020.113575

Hawton, K., Casañas i Comabella, C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: A systematic review. *Journal of Affective Disorders*, *147*(1), 17–28. https://doi.org/https://doi.org/10.1016/j.jad.2013.01.004

Heslin, M., Jin, H., Trevillion, K., Ling, X., Nath, S., Barrett, B., Demilew, J., Ryan, E. G., O'Connor, S., Sands, P., Milgrom, J., Bick, D., Stanley, N., Hunter, M. S., Howard, L. M., & Byford, S. (2022). Cost-effectiveness of screening tools for identifying depression in early pregnancy: a decision tree model. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-08115-x

Hidayat, R., & Primasari, I. (2011). Metodologi Penelitian Psikodiagnostika. *Buletin Psikologi*, *19*(2), 81–92.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ikatan Psikolog Klinis Indonesia (IPK). (2022). *Pusat Data Strategis dan Statistik Ikatan Psikolog Klinis (IPK) Indonesia*. https://data.ipkindonesia.or.id/

Kaggwa, M. M., Najjuka, S. M., Ashaba, S., & Mamun, M. A. (2022). Psychometrics of the Patient Health Questionnaire (PHQ-9) in Uganda: A Systematic Review. Dalam *Frontiers in Psychiatry* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fpsyt.2022.781095

Kemenkes RI. (2022). *Sistem Kesehatan Jiwa di Indonesia*.

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52. https://doi.org/10.5395/rde.2013.38.1.52

Krause, J. S., Bombardier, C., & Carter, R. E. (2008). Assessment of Depressive Symptoms During Inpatient Rehabilitation for Spinal Cord Injury: Is There an Underlying Somatic Factor When Using the PHQ? *Rehabilitation Psychology*, *53*(4), 513–520. https://doi.org/10.1037/a0013354

Krause, J. S., Reed, K. S., & McArdle, J. J. (2010). Factor structure and predictive validity of somatic and nonsomatic symptoms from the patient health questionnaire-9: A longitudinal study after spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, *91*(8), 1218–1224. https://doi.org/10.1016/j.apmr.2010.04.015

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Le Hoang Ngoc, T., Le, M. A. T., Nguyen, H. T., Vo, H. V., Le, N. Q., Tang, L. N. P., Tran, T. T., & Le, T. Van. (2021). Patient Health Questionnaire (PHQ-9): A depression screening tool for people with epilepsy in Vietnam. *Epilepsy and Behavior*, *125*. https://doi.org/10.1016/j.yebeh.2021.108446

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*(Pt A), 236–240. https://doi.org/10.1016/j.jad.2020.11.007

López-Guerra, V. M., López-Núñez, C., Vaca-Gallegos, S. L., & Torres-Carrión, P. V. (2022). Psychometric Properties and Factor Structure of the Patient Health Questionnaire-9 as a Screening Tool for Depression Among Ecuadorian College Students. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.813894

Marvianto, R. D., & Widhiarso, W. (2018). Adaptasi dan Evaluasi Properti Psikometris Skala Academic Motivation Scale (AMS) versi Bahasa Indonesia. *Gadjah mada Journal of Psychology*, *4*(1), 87–95.

Matlock Cole, K. L., Turner, R. C., & Gitchel, W. D. (2018). A Study of Reverse-Worded Matched Item Pairs Using the Generalized Partial Credit and Nominal Response Models. *Educational and Psychological Measurement*, *78*(1), 103–127. https://doi.org/10.1177/0013164416670211

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Moulton, S. E., Young, E. L., & Sudweeks, R. R. (2019). Examining the Psychometric Properties of the SRSS-IE With the Nominal Response Model Within a Middle School Sample. *Assessment for Effective Intervention*, *44*(4), 227–240. https://doi.org/10.1177/1534508418777866

Mufson, L., Morrison, C., Shea, E., Kluisza, L., Robbins, R., Chen, Y., & Mellins, C. A. (2022). Screening for depression with the PHQ-9 in young adults affected by HIV. *Journal of Affective Disorders*, *297*, 276–282. https://doi.org/10.1016/j.jad.2021.10.037

Murray, A. L., Booth, T., & Molenaar, D. (2016). When Middle Really Means “top” or “bottom”: An Analysis of the 16PF5 Using Bock’s Nominal Response Model. *Journal of Personality Assessment*, *98*(3), 319–331. https://doi.org/10.1080/00223891.2015.1095197

Murray, A. L., Hemady, C. L., Do, H., Dunne, M., Foley, S., Osafo, J., Sikander, S., Madrid, B., Baban, A., Taut, D., Ward, C. L., Fernando, A., Thang, V. Van, Eisner, M., Hughes, C., Fearon, P., Valdebenito, S., Tomlinson, M., Pathmeswaran, A., & Walker, S. (2022). Measuring Antenatal Depressive Symptoms Across the World: A Validation and Cross-Country Invariance Analysis of the Patient Health Questionnaire-9 (PHQ-9) in Eight Diverse Low-Resource Settings. *Psychological Assessment*. https://doi.org/10.1037/pas0001154

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User’s Guide* (Sixth Edition). Muthén & Muthén.

Nafilyan, V., Pabon, M. A., & de Coulon, A. (2021). *The Causal Impact of Depression on Cognitive Functioning: Evidence from Europe*. www.iza.org

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

Paljärvi, T., Tiihonen, J., Lähteenvuo, M., Tanskanen, A., Fazel, S., & Taipale, H. (2023). Psychotic depression and deaths due to suicide. *Journal of affective disorders*, *321*, 28–32. https://doi.org/10.1016/j.jad.2022.10.035

Pranckeviciene, A., Saudargiene, A., Gecaite-Stonciene, J., Liaugaudaite, V., Griskova-Bulanova, I., Simkute, D., Naginiene, R., Dainauskas, L. L., Ceidaite, G., & Burkauskas, J. (2022). Validation of the patient health questionnaire- 9 and the generalized anxiety disorder-7 in Lithuanian student sample. *PLoS ONE*, *17*(1 January). https://doi.org/10.1371/journal.pone.0263027

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, *71*(3), 523–550. https://doi.org/10.1177/0013164410382250

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*, *120*(7), 1434–1443. https://doi.org/https://doi.org/10.1016/j.jfma.2021.01.016

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Rahman, M. A., Dhira, T. A., Sarker, A. R., & Mehareen, J. (2022). Validity and reliability of the Patient Health Questionnaire scale (PHQ-9) among university students of Bangladesh. *PLoS ONE*, *17*(6 June). https://doi.org/10.1371/journal.pone.0269634

Reise, S. P., Hubbard, A. S., Wong, E. F., Schalet, B. D., Haviland, M. G., & Kimerling, R. (2021a). Response Category Functioning on the Health Care Engagement Measure Using the Nominal Response Model. *Assessment*. https://doi.org/10.1177/10731911211052682

Reise, S. P., Hubbard, A. S., Wong, E. F., Schalet, B. D., Haviland, M. G., & Kimerling, R. (2021b). Response Category Functioning on the Health Care Engagement Measure Using the Nominal Response Model. *Assessment*. https://doi.org/10.1177/10731911211052682

Reutfors, J., Andersson, T. M.-L., Tanskanen, A., DiBernardo, A., Li, G., Brandt, L., & Brenner, P. (2021). Risk Factors for Suicide and Suicide Attempts Among Patients With Treatment-Resistant Depression: Nested Case-Control Study. *Archives of Suicide Research*, *25*(3), 424–438. https://doi.org/10.1080/13811118.2019.1691692

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

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Smith, M. L., Sanchez, S. E., Rondon, M., Gradus, J. L., & Gelaye, B. (2022). Validation of the patient health Questionnaire-9 (PHQ-9) for detecting depression among pregnant women in Lima, Peru. *Current Psychology*, *41*(6), 3797–3805. https://doi.org/10.1007/s12144-020-00882-2

Tama, T. D., Astutik, E., & Reuwpassa, J. O. (2021). Predictors of depressive symptoms based on the human capital model approach: Findings from the Indonesia family life survey. *Yale Journal of Biology and Medicine*, *94*(3), 395–406. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117204066&partnerID=40&md5=6a75db4c86e5e4e468acc5ebe8903118

Tele, A. K., Carvajal-Velez, L., Nyongesa, V., Ahs, J. W., Mwaniga, S., Kathono, J., Yator, O., Njuguna, S., Kanyanya, I., Amin, N., Kohrt, B., Wambua, G. N., & Kumar, M. (2023). Validation of the English and Swahili Adaptation of the Patient Health Questionnaire–9 for Use Among Adolescents in Kenya. *Journal of Adolescent Health*, *72*(1), S61–S70. https://doi.org/10.1016/j.jadohealth.2022.10.003

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. Dalam *Handbook of polytomous item response theory models.* (hlm. 43–75). Routledge/Taylor & Francis Group.

Vu, L. G., Le, L. K., Dam, A. V. T., Nguyen, S. H., Vu, T. T. M., Trinh, T. T. H., Do, A. L., Do, N. M., Le, T. H., Latkin, C., Ho, R. C. M., & Ho, C. S. H. (2022). Factor Structures of Patient Health Questionnaire-9 Instruments in Exploring Depressive Symptoms of Suburban Population. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.838747

Wang, Y., Liang, L., Sun, Z., Liu, R., Wei, Y., Qi, S., Ke, Q., & Wang, F. (2023). Factor structure of the patient health questionnaire-9 and measurement invariance across gender and age among Chinese university students. *Medicine (United States)*, *102*(1), E32590. https://doi.org/10.1097/MD.0000000000032590

Waqas, A., Koukab, A., Meraj, H., Dua, T., Chowdhary, N., Fatima, B., & Rahman, A. (2022). Screening programs for common maternal mental health disorders among perinatal women: report of the systematic review of evidence. *BMC Psychiatry*, *22*(1). https://doi.org/10.1186/s12888-022-03694-9

Williams, S. Z., Lewis, C. F., Muennig, P., Martino, D., & Pahl, K. (2022). Self-reported anxiety and depression problems and suicide ideation among black and latinx adults and the moderating role of social support. *Journal of Community Health*, *47*(6), 914–923. https://doi.org/10.1007/s10900-022-01127-y

World Health Organization (WHO). (2022, Juni 17). *Mental health: strengthening our response*. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Dalam *APPLIED PSYCHOLOGICAL MEASUREMENT* (Vol. 8, Nomor 2).

Yu, X., Tam, W. W. S., Wong, P. T. K., Lam, T. H., & Stewart, S. M. (2012). The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Comprehensive Psychiatry*, *53*(1), 95–102. https://doi.org/10.1016/j.comppsych.2010.11.002

Appendix

Appendix 1. Item descriptive statistics

Item	Mean	SD	Skewness	Kurtosis	Response proportion			
					0	1	2	3
Item 1	1.592	.931	.213	-.976	.092	.446	.242	.221
Item 2	1.469	.984	.196	-.996	.163	.403	.237	.197
Item 3	1.748	1.069	-.165	-1.298	.138	.315	.208	.339
Item 4	1.750	.99	-.098	-1.157	.099	.348	.256	.296
Item 5	1.437	1.073	.126	-1.235	.234	.314	.234	.218
Item 6	1.671	1.123	-.137	-1.381	.191	.276	.204	.329
Item 7	1.380	1.088	.216	-1.240	.256	.324	.203	.217
Item 8	1.033	1.036	.632	-.807	.392	.313	.166	.129
Item 9	0.669	.949	1.196	.217	.599	.202	.128	.070

Notes. SD = *Standard deviation*; 0 = Tidak pernah; 1 = Beberapa hari; 2 = Lebih dari separuh waktu yang dimaksud; 3 = Hampir setiap hari.

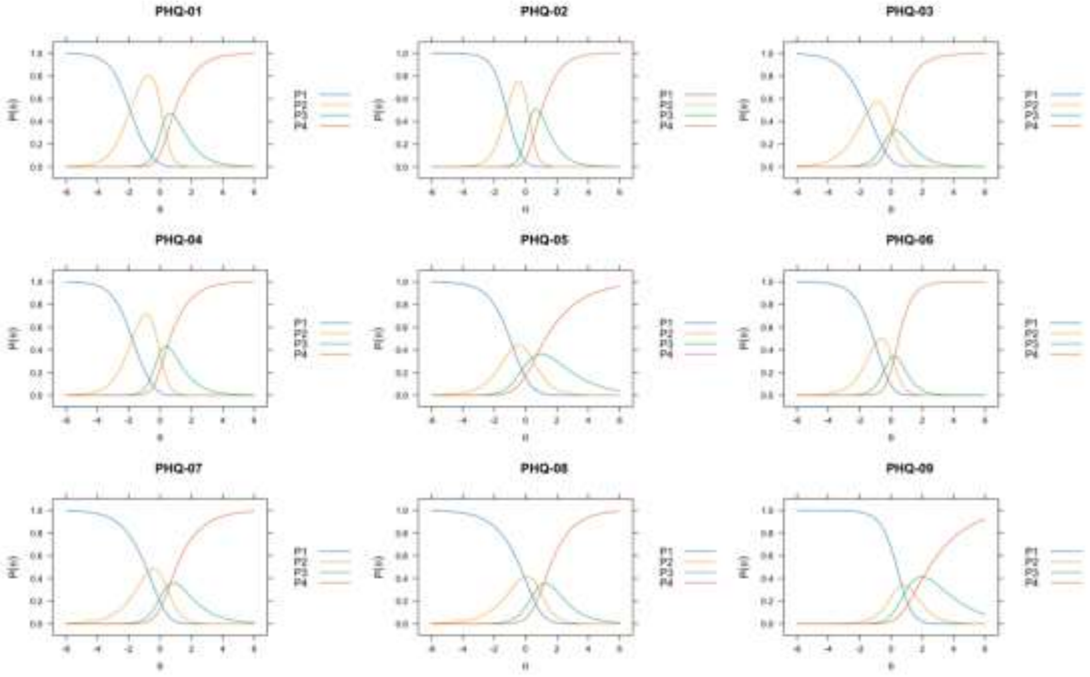
Appendix 2. Correlation matrices between items and total score using Pearson

Items		1	2	3	4	5	6	7	8	9
1	PHQ01	-								
2	PHQ02	.647	-							
3	PHQ03	.542	.520	-						
4	PHQ04	.601	.587	.585	-					
5	PHQ05	.433	.449	.537	.516	-				
6	PHQ06	.549	.662	.481	.533	.444	-			
7	PHQ07	.536	.506	.492	.515	.409	.544	-		
8	PHQ08	.438	.477	.446	.449	.432	.481	.519	-	
9	PHQ09	.448	.557	.411	.402	.385	.507	.431	.416	-
10	Total	.769	.804	.752	.773	.693	.782	.745	.698	.674

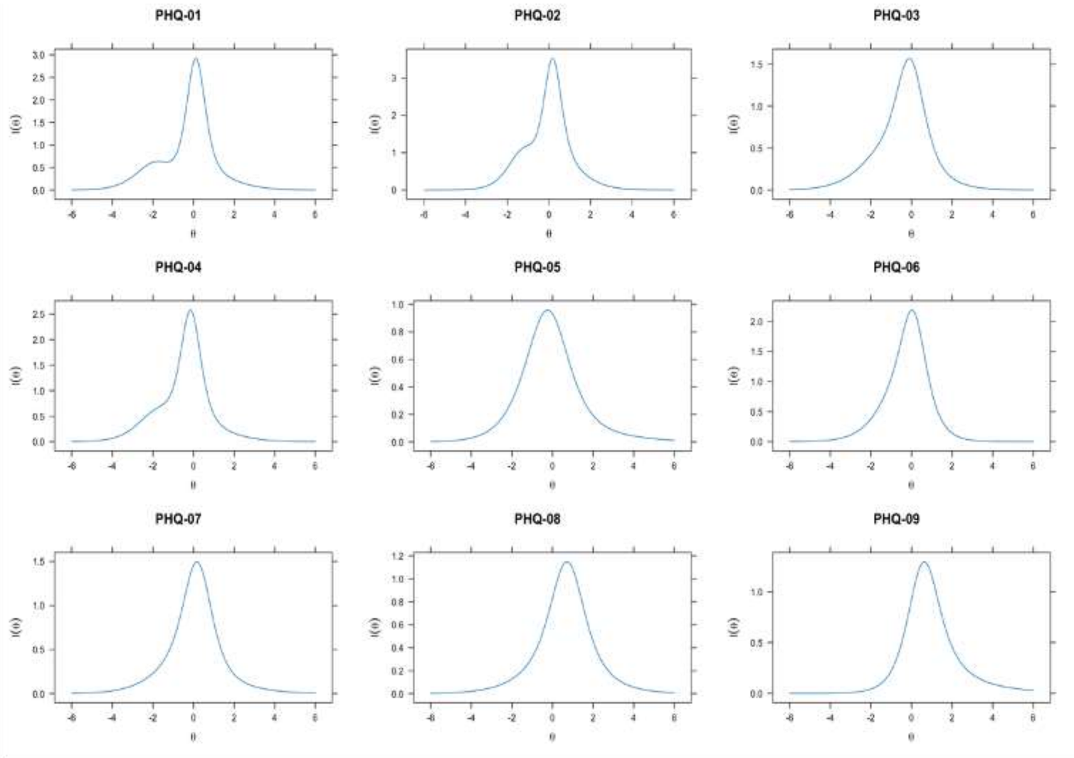
Appendix 3. Correlation matrices between Q3 residue

	Items	1	2	3	4	5	6	7	8
1	PHQ01	-							
2	PHQ02	-.063	-						
3	PHQ03	-.114	-.265	-					
4	PHQ04	-.046	-.179	.004	-				
5	PHQ05	-.172	-.224	.115	.021	-			
6	PHQ06	-.213	.007	-.277	-.258	-.151	-		
7	PHQ07	-.087	-.263	-.096	-.128	-.117	-.062	-	
8	PHQ08	-.178	-.188	-.079	-.136	-.016	-.069	.077	-
9	PHQ09	-.162	-.015	-.120	-.192	-.068	-.001	-.104	-.110

Appendix 4. Option response function (ORF) in each item



Appendix 5. Item information function (IIF) in each item



Appendix 6. Item wording dan item fit using S-χ^2 inpackage mirt

	Item	No corrections			FDR		
		χ^2	df	p.s	χ^2	df	p.s
1	Kurang tertarik atau bergairah dalam melakukan apapun	34.433	39	.678	34.433	39	.763
2	Merasa murung, muram, atau putus asa	45.034	37	.171	45.034	37	.463
3	Sulit tidur atau mudah terbangun, atau terlalu banyak tidur	46.439	44	.372	46.439	44	.670
4	Merasa lelah atau kurang bertenaga	28.165	40	.920	28.165	40	.920
5	Kurang nafsu makan atau terlalu banyak makan	59.017	51	.206	59.017	51	.463
6	Kurang percaya diri — atau merasa bahwa Anda adalah orang yang gagal atau telah mengecewakan diri sendiri atau keluarga	44.460	44	.452	44.460	44	.678
7	Sulit berkonsentrasi pada sesuatu, misalnya membaca koran atau menonton televisi	46.185	48	.547	46.185	48	.704
8	Bergerak atau berbicara sangat lambat sehingga orang lain memperhatikannya. Atau sebaliknya — merasa resah atau gelisah sehingga Anda lebih sering bergerak dari biasanya.	68.940	48	.025	68.940	48	.229
9	Merasa lebih baik mati atau ingin melukai diri sendiri dengan cara apapun.	59.860	44	.056	59.860	44	.251