Construct validity and differential item functioning of the PHQ-9 among health care workers: Rasch Analysis approach

Surin Jiraniramai
Chiang Mai University Faculty of Medicine

Tinakon Wongpakaran
Chiang Mai University Faculty of Medicine

Chaisiri Angkurawaranon
Chiang Mai University Faculty of Medicine

Wichuda Jirapomcharoen
Chiang Mai University Faculty of Medicine

Nahathai Wongpakaran (nahathai.wongpakaran@cmu.ac.th)
Chiang Mai University Faculty of Medicine

Keywords: PHQ-9, Thai, Rasch Analysis, Differential item functioning, Alcohol consumption

Posted Date: May 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-25004/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background The Patient Health Questionnaire (PHQ-9) is a widely used self-report questionnaire to screen depression. Its psychometric property has been tested in many populations including health care workers. We used Rasch measurement theory to examine the psychometric properties of PHQ-9 regarding item difficulty, item fit and the differences between subgroups of respondents classified by sex, age, education and alcohol user status, based on the same overall location of participants.

Methods In total, 3,204 health care workers of Maharaj Nakorn Chiang Mai Hospital participated and were administered the PHQ-9. Rating scale Rasch measurement modeling was used to examine the psychometric properties of the PHQ-9.

Results The data fitted well to the Rasch model and no violations of the assumption of unidimensionality were observed. All 9 items could form a unidimensional construct of overall depressive severity. Suicidal ideation was the least endorsed while sleep problem was the most. No disordered category and threshold of the rating response were observed. No locally dependent items were observed. No items were found to show differential item functioning across age, sex, education and alcohol consumption. The item-person Wright map showed that the PHQ-9 did not target well with the sample, and a wide gap suggesting few or no items exist to differentiate participants at a certain ability level among the PHQ-9 items.

Conclusion The PHQ-9 can be used as a screening questionnaire for major depressive disorder as its psychometric property was verified based on Rasch measurement model. The findings are generally consistent with related studies in other populations. However, the PHQ-9 may be unsuitable for assessing depressive symptoms among health care workers who have low levels of depression.

Background

Depression is considered the most common mental disorder and a major cause of disability in Thailand [1] and globally [2]. A number of studies have shown that health care workers, defined by WHO as all personnel involved in actions whose principal intent is to promote health [3], are exposed to psychological distress related to their occupation. They have heavy workloads, night work or shift work. These occupational stress factors can lead to burnout, anxiety, sleep problems, psychiatric disorders or even depression [4, 5].

In Thailand, one study among health care workers revealed a high rate of depression at 21.5%[6]. Consequences of depression among health care workers could produce medical errors and affect treatment outcome [7–9]. Therefore, screening and grading depression symptom severity are important approaches to early detect and manage depression among health care workers.

The Patient Health Questionnaire (PHQ-9) has been used as a depression screening tool in various settings, demonstrating good sensitivity and specificity for depressive disorders [10]. It comprises nine
items of diagnostic symptom criteria based on the DSM-IV with Likert scale responses, i.e., 0 (not at all) to 4 (nearly every day) [11]. The PHQ-9 has been translated to many languages including Thai [12].

Even though the PHQ-9 has been widely used because of its brevity; related literature has shown the PHQ-9 exhibited some problems that led to modification when tested for its construct validity. Problems may have stemmed from the sample regarding the scale's unidimensionality (how well items support the same underlying construct of depression), problems of low response endorsement regarding some items, and problems involving response bias despite that the severity of depression was the same (differential item functioning-DIF) [13–15].

PHQ-9 has been investigated for psychometric properties both by traditional classical test theory (CTT) and by item response theory such as Rasch model analysis [13, 16–19]. Most studies reported that all nine items could form a unidimensional construct of overall depressive severity. The advantage of Rasch analysis over CTT was that Rasch measurement transforms ordinal level data (i.e. Likert response categories) to interval-level data, while CTT mistakenly treated the ordinal scale as an interval scale [20–22].

In addition, Rasch analysis was used to evaluate differential item functioning or item bias related to subpopulations. Most studies found that the items of PHQ-9 were free of DIF with regard to age, education, and employment status; however, DIF related to visual impairment was found with “trouble falling asleep” item [16], and DIF related to sex among elderly populations, which was found with “sleep” and “self-blame” items [15].

The PHQ-9 has been investigated for DIF in many subpopulations. While some found that DIF was unrelated to ethnic background [23], language [24], race [25], sex [25, 26] and level of education [19], it found DIF related to age [26, 27], cognitive impairment [28], a history of mania [29] and visual impairment [16]. In addition, age group plays a role concerning DIF. For instance, DIF related to sex was detected in an elderly population [15].

The psychometric quality of the PHQ-9, as evidence shows, varied depending on the studied population. For health care workers, a report of Rasch validation of PHQ-9 is lacking. In addition, one common clinical characteristic among health care workers is alcohol assumption [30, 31]. Alcohol users tend to have irrational beliefs [32], as well as other psychiatric symptoms such as anxiety or depression and cognitive impairment[28]. This casts doubt that alcohol users may view or assess depressive symptoms differently from nonalcohol users due to their distorted beliefs [33]. Whether alcohol consumption is related to DIF has, nevertheless, has yet to be examined.

The present study aimed to investigate the psychometric properties of PHQ-9 among health care worker using the Rasch model. We investigated 1) to see whether the PHQ-9 data in this sample fit the Rasch measurement model, 2) to see how appropriate the PHQ-9 was for this population, i.e., whether item difficulty levels of depressive symptoms sufficiently covered the whole range of an individual's
depression and 3) to examine whether DIF due to age, sex, education and condition of alcohol consumption existed.

**Methods**

Participants and procedures

This cross-sectional health care workers’ study, conducted at Chiang Mai University (CMU) between February and June 2013, was approved by the Ethics Review Committee for Research in Human Subjects, Faculty of Medicine, Chiang Mai University. Health care workers were classified in three groups. The first group consisted of doctors, dentists, nurses and pharmacists (42.1%). The second group was “other health professionals” and other health-related positions (19.4%). The last group was “nonhealth professionals” and mainly consisted of workers (38.5%). A detailed description of the study has been published [34]. Concisely, 5,364 people working for the Faculty of Medicine of CMU, 4,022 people (75.0%) responded to the survey and 3,532 (65.8%) consented to participate in the study. In the end, 3204 participants (59.7% response rate) completed the self-rating online questionnaires concerning PHQ-9 as well as their demographic information. This comprised age, sex, education level and alcohol consumption.

**Measure**

Patient Health Questionnaire (PHQ-9)

PHQ-9 is a self-report tool, consisting of nine questions regarding depressive symptoms based on the DSM-IV criteria for a major depressive episode (Kroenke et al., 2001). The questions included the symptoms of: lack of interest, depressed mood, sleeping difficulties, tiredness, appetite problems, concentration problems, psychomotor agitation/retardation, negative feelings about self and suicidal ideation. The respondent was asked how many symptoms he/she experienced during past the two weeks. Items were administered on a 4-point Likert scale with the response options: 0 “not at all”, 1 “several days”, 2 “more than one half of the days”, and 3 “nearly every day”. The Thai version of the PHQ-9 was shown to have acceptable psychometric properties to screen for major depression in the primary care setting (Lotrakul et al., 2008).

Statistical analysis

Demographic data were described using mean, SD and frequency. The Rasch rating scale model was used to verify the construct validity of the PHQ-9.

Rasch analysis is a mathematical method to calibrate linear logit measures of item difficulty and person ability from ordinal data. To examine the PHQ-9 construct, a firmly established calibration of item measures was needed to make inference about the construct. According to the Rasch model, the probability of an individual’s response counts on both “person ability” and “item difficulty” [35]. Herein, “person ability” refers to as the extent to which the participants experience depression and “item
difficulty” refers to the severity of depression expressed by the item. The response probabilities of each person to each of the individual items, according to the Rasch model, are modeled as a logistic function of the latent depression trait. This model yields person and item depression estimates, as well as estimates of a set between response category thresholds common to all items. Item estimates below 0 (mean) are considered easy for the person to endorse, comparable to a person with a lower level of depression. The opposite meaning is applied when item and person estimates are above 0.

To test whether the data could fit the Rasch model, fit statistics, e.g., information-weighted fit statistics (infit) mean square (MnSq) and outlier-sensitive fit statistics (outfit) MnSq were used. An item with infit or outfit MnSq out of the 0.7–1.5 range was considered a misfit [36]. The performance of the scale was examined using Rasch fit statistics, and the dimensionality of the scale was examined using principal component analysis (PCA) of the standardized residuals. To indicate unidimensionality, there should be an absence of any meaningful pattern in the residuals. The first residual dimension is usually expected to have a value smaller than 2.0, which has been shown to happen entirely due to random variation [37]. In addition, fit statistics <0.6 indicate items overfit the model, usually because they share some components of meaning with other items [22].

Local dependency, referring to the items containing a latent trait other than depression, was tested using the correlation (r) of the Rasch residuals between each pair of item; r ≤ 0.3 was considered acceptable [38].

Item ordering, indicating that a higher severity of a symptom should score a higher category, was examined using category function. The threshold estimates for a 4 –category response option was examined to verify whether participants discriminated between the available ordered response categories. The disordering threshold could be examined in two ways, first) by considering infit and outfit MnSq within 0.7 and 1.3 and second) by the ordering of the ‘observed averages; acceptable response scores should monotonically increase average difficulties (average measure) and step difficulties (step measure).

We used a person-item (Wright) map to plot item difficulty and the individual’s abilities along its continuum on the same axis of the logits allowing the evaluation of the fit of the item difficulties matched to the abilities of the individuals. We examine to what the extent the item positions match the person positions (targeting) using the Wright map. The best targeting of a measurement is when the mean items are at the same measure as the mean persons. Researchers suggests the difference between the mean value of the mean person measure should be within one logit [39]. Floor or ceiling effects could also be visualized using this map.

We tested the differential item functioning (DIF) across sex, age, education and alcohol consumption. Both statistical test and DIF contrast were used, and a DIF contrast >0.64 indicated a substantial DIF [40].
Finally, reliability was evaluated using person separation index (comparable to Cronbach's alpha). Person separation index denotes how well the test is able to differentiate among groups of respondents with different levels of depression. An acceptable value for separation is at least 2. Item reliability was assessed using item separation index. Separation value was less than 3 and item reliability was less than 0.9, implying that the sample is not large enough to endorse construct validity or a difficulty exists with the item hierarchy of the instrument [40].

All analyses were conducted using IBM SPSS for Windows, Version 22 (Chicago, IL, USA), STATA, Version 14 () and Rasch models using WINSTEPS [40].

**Results**

Table 1 shows that the majority of the population was female. Up to 31% consumed alcohol; however, most (80.7%) were considered at low harmful risk. For PHQ-9, over 94% (n = 3005) scored below the cutoff point for clinical depression (a sum score of 10 or over).
Table 1
Sociodemographic characteristics of the subjects

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>N(%) or Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), Mean (SD)</td>
<td>40.2(10.7)</td>
</tr>
<tr>
<td>Sex: Female, n (%)</td>
<td>2471(77.1)</td>
</tr>
<tr>
<td>Marital status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>1380(43.2%)</td>
</tr>
<tr>
<td>Married (lived together, separated, divorced, widowed)</td>
<td>1817 (56.7%)</td>
</tr>
<tr>
<td>Educational status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Lower than Bachelor degree</td>
<td>1135(35.4%)</td>
</tr>
<tr>
<td>Bachelor degree and higher</td>
<td>2069(64.6%)</td>
</tr>
<tr>
<td>Job description</td>
<td></td>
</tr>
<tr>
<td>Doctors, dentists, nurses and pharmacists</td>
<td>1350(42.1%)</td>
</tr>
<tr>
<td>Other health professionals</td>
<td>621(19.4)</td>
</tr>
<tr>
<td>Nonhealth professionals</td>
<td>1233(38.5%)</td>
</tr>
<tr>
<td>Alcohol consumption, n (%)</td>
<td>988 (30.8%)</td>
</tr>
<tr>
<td>PHQ-9 sum score, mean (SD)</td>
<td>4.3(3.2)</td>
</tr>
<tr>
<td>Median, Inter quartile range</td>
<td>4,4</td>
</tr>
<tr>
<td>Min-max</td>
<td>0–23</td>
</tr>
<tr>
<td>Severity of depression, n (%)</td>
<td></td>
</tr>
<tr>
<td>None or minimal (0–4)</td>
<td>1832(57.2)</td>
</tr>
<tr>
<td>Mild (5–9)</td>
<td>1173 (36.2)</td>
</tr>
<tr>
<td>Moderate (10–14)</td>
<td>182 (5.7)</td>
</tr>
<tr>
<td>Moderately severe (15–19)</td>
<td>13(0.4)</td>
</tr>
<tr>
<td>Severe (20–27)</td>
<td>4(0.1)</td>
</tr>
</tbody>
</table>

SD = standard deviation, PHQ = Patient health questionnaire

Table 2 shows the proportion of each category and fit statistics. All items were shown to have fit statistics in the required range, 0.7–1.5. As suggested by Linacre, the standardized statistics is basically overly sensitive to misfit a large sample. However, Zstd > 2.00 could be ignored, when mean-squares were
acceptable. The overall fit of the data to the model was good, indicating that overall, the 9-item scale formed a valid measure.

Table 2
Item fit for PHQ-9

<table>
<thead>
<tr>
<th>Item description</th>
<th>Measure or logits (SE)</th>
<th>Infit Mean square</th>
<th>Infit Zstd</th>
<th>Outfit Mean square</th>
<th>Outfit Zstd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Little interest</td>
<td>-1.55(0.4)</td>
<td>0.72</td>
<td>-9.90</td>
<td>0.71</td>
<td>-9.90</td>
</tr>
<tr>
<td>2. Feeling down</td>
<td>-0.87(0.4)</td>
<td>0.79</td>
<td>-8.57</td>
<td>0.78</td>
<td>-8.71</td>
</tr>
<tr>
<td>3. Trouble sleeping</td>
<td>-1.70(0.4)</td>
<td>1.27</td>
<td>9.17</td>
<td>1.29</td>
<td>9.72</td>
</tr>
<tr>
<td>4. Feeling tired</td>
<td>-0.87(0.4)</td>
<td>1.06</td>
<td>2.11</td>
<td>1.05</td>
<td>1.65</td>
</tr>
<tr>
<td>5. Poor appetite</td>
<td>-0.94(0.4)</td>
<td>1.22</td>
<td>7.65</td>
<td>1.18</td>
<td>6.12</td>
</tr>
<tr>
<td>6. Feeling bad about self</td>
<td>0.98(0.5)</td>
<td>1.01</td>
<td>0.51</td>
<td>0.90</td>
<td>-1.79</td>
</tr>
<tr>
<td>7. Trouble concentrating</td>
<td>-0.24(0.4)</td>
<td>0.96</td>
<td>-1.39</td>
<td>0.97</td>
<td>-1.06</td>
</tr>
<tr>
<td>8. Moving slowly</td>
<td>1.28(0.5)</td>
<td>1.04</td>
<td>1.32</td>
<td>0.87</td>
<td>-2.04</td>
</tr>
<tr>
<td>9. Better off dead</td>
<td>3.91(1.0)</td>
<td>1.13</td>
<td>1.48</td>
<td>0.70</td>
<td>-1.70</td>
</tr>
</tbody>
</table>

infit = information-weighted fit statistics, outfit = outlier-sensitive fit statistics, Zstd = z-score standardized, SE = Standard error, PHQ = Patient Health Questionnaire

“Trouble sleeping” was the most endorsed item (logit= -1.70), while “Better off dead” was the least endorsed (logit = 3.91).

In terms of dimensionality, the analysis showed that the raw unexplained variance was 48.2%, while the eigenvalue of the unexplained variance was in 1st contrast = 1.58 which was less than 2.0 indicating the PHQ-9 was less likely to have another dimension. In addition, we found the Pearson's correlation was 1.000 for item clusters 1–2, 1–3, and 2–3 indicating no different dimension. No pair of item had a residual correlation > 0.2. The most were items 2, “Feeling down” and 1, “Little interest“ ( r = .12). This lacked substantial residual correlations between items, indicating the criteria of local independence was met.

PHQ-9 was free of DIF. However, one item showed a minimal level of DIF by sex, with male participants rating item #6 (Feeling bad about self) as easier than female participants (.66 and 1.08, respectively). Participants who had alcohol consumption rated the item “Feeling bad about self as 0.70 logits more easily than those who did not consume alcohol (1.06 logits). The DIF contrast for all were; however, less than the cut-off .64 logits.
Figure 1 shows the person-item map for the PHQ-9. The person-item map indicated items for persons with lower ability estimates were missing. (3.31 logits was the highest person ability estimate) and presented evidence for a floor effect of the PHQ-9. The mean person ability was \( -3.83 \) logits (SD 1.85). The PHQ-9 had a wider range but was not a better match for the sample. Most individual’s level of depression did not match any item of the PHQ-9. Specifically, no items were available to accurately measure individuals with abilities between \(-2.22\) and \(-5.85\) logits. This implied that the items were too difficult for the abilities of the respondents. In addition, redundancy of the items was observed, that is, pairs of items, e.g., “Feeling tired” and “Feeling down”, were shown on the map to be located at the same difficulty level, indicating that they exhibited a similar level of depression.

The three most difficult items were, “Better off dead,” “Moving slowly,” and “Feeling bad about self” Conversely, the three least difficult items were “Trouble sleeping”, “Little interest,” and “Poor appetite”.

Table 3 shows the summary statistics for the 4 rating-scale categories. The “frequency of use” of categories 0 and 1 were much more than any of categories 2 and 3. This implied that most were “less able” persons (low level of depression), consistent with the fact that mean person ability (-3.83) was lower than mean item difficulty (0). The observed person measures that increased from category 0 to category 3 represented low (-4.80) to high ability (0.60) denoting that no collapse of rating categories was necessary. The outfit MnSq for category 3 was slightly high (1.52), indicating an idiosyncratic use of category 3. An adjustment of the description for “category 3 (nearly every day)” would probably improve the functioning of the entire rating scale.

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency of use</th>
<th>Percent</th>
<th>Observed person measure</th>
<th>Infit MnSq</th>
<th>Outfit MnSq</th>
<th>Andrich threshold</th>
<th>Category measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16679</td>
<td>58</td>
<td>-4.80</td>
<td>0.99</td>
<td>0.98</td>
<td>NONE</td>
<td>(-4.20)</td>
</tr>
<tr>
<td>1</td>
<td>10685</td>
<td>37</td>
<td>-2.11</td>
<td>0.96</td>
<td>0.83</td>
<td>-3.08</td>
<td>-1.14</td>
</tr>
<tr>
<td>2</td>
<td>1277</td>
<td>4</td>
<td>-0.49</td>
<td>1.07</td>
<td>1.08</td>
<td>0.87</td>
<td>1.55</td>
</tr>
<tr>
<td>3</td>
<td>195</td>
<td>1</td>
<td>0.60</td>
<td>1.33</td>
<td>1.52</td>
<td>2.22</td>
<td>(3.47)</td>
</tr>
</tbody>
</table>

infit = information-weighted fit statistics, outfit = outlier-sensitive fit statistics, MnSq = mean square, PHQ = patient Health Questionnaire

Figure 1 shows the category probability curves for an item of the PHQ-9. No evidence of disordered thresholds with the 4-category response was observed. The person separation was 1.54 and the reliability, 0.70. The internal consistency was good when used (Cronbach Alpha = .80). The item separation was 31.24 and item reliability = 1.00.

**Discussion**
The present study aimed to evaluate the validity of the PHQ-9 in a large sample of health care workers using Rasch analysis, as well as possible item bias due to sex, age, education and particularly alcohol consumption. Our results were in line with related studies that the PHQ-9 fitted the assumption of the Rasch measurement model, i.e., unidimensionality and local independence, indicating that all items contributed to the same depression underlying construct [13, 27, 41], but contrasted with other related studies [15, 17, 18]. This inconsistency may have contributed to the different characteristics of the studied sample, especially none of studies had been conducted before among a health care worker sample.

In terms of item hierarchy, our findings concurred with related studies. The easy items were, “Trouble sleeping”, “Little interest” and “Poor appetite”, while the most difficult item was “Better off dead”. Basically, suicidal ideation was related to severe depression [12, 42]. It appeared that this item was difficult to be endorsed in the general population including among health care workers because health care workers experienced mild levels of depression compared with a clinical subject with depression who is intended to be the real target for the PHQ-9. The item ordering may vary from sample to sample, especially among clinical subjects [13, 16, 43]. In line with related studies, the present results showed that items, “Feeling tired” and “Feeling down” appeared to be redundant and one could be removed when the same accuracy of the reduced version is warranted [17, 18].

Notably, item 9 loosely related to latent dimension depression, leaving the question whether or not incorporating suicidality is useful in such a screening scale for depression [15, 18, 44]. Not only did it receive the lowest endorsement, one study showed that item 9 illustrated both item misfit and disordered threshold [45]. From this reason item 9 was removed, and PHQ-8 became adopted for a screening tool of depression [46, 47]. In terms of category, most participants tended to endorse “0 (not at all)” or “1 (several days)” due to their low level of depression, category “3 (nearly every day)” was thus less endorsed to the extent that it created a mild misfit category. Despite that; however, this can be ignored as the effect is not detrimental to the scale [36].

That PHQ-9 does not target well in a sample who generally have low level depression was expected and consistent with one related study in a nonclinical population [15]. Because the PHQ-9 adopted symptoms originating from DSM, the items are designed for depressed people. Well-targeting is more usually found in studies in a clinical sample than a nonclinical population [17]. This is to confirm that PHQ-9 functions as intended among clinical but not general people including health care worker. Therefore, PHQ-9 can still be used a screening tool for major depression, but is not to be applied to measure depression (as outcome measure) among health care workers. The big gap in the Wright map from Rasch analysis results suggested that easier items are needed if the PHQ-9 is used to assess depression level.

Other measurements, containing more items, for example, the 20-item Center of Epidemiological Study of Depression scale (CES-D) and 21-item Beck Depression Inventory (BDI-II), were found to have better targeting in adult populations, whereas the 15-item geriatric depression rating scale showed better
targeting in elderly populations [15, 48]. This could be because those scales have more items that are able to cover broader latent construct of a subject's depressive severity.

In terms of reliability, the person separation was not excellent, albeit acceptable (.74), while the Cronbach alpha was .80. This may have contributed to the number of items being relative low - more items may be needed [40]. Pearson separation is basically lower than Cronbach's alpha because the Rasch-based reliability of separation statistics is based on a linear, interval-level scale when a good model-data fit is observed, whereas alpha is based only on the assumption of linear measures [49]. Here again, it could be suggested that more items are needed for PHQ-9 to be used as a measuring tool for this population.

The present study found no significant DIF due to sex, age, education and alcohol consumption indicating that the PHQ-9 could be used in this population without modifying. This may have contributed to the no to low level of depression for the whole sample. DIF is usually found in some specific demographic or clinical sample, e.g., visual impairment, the elderly, primary care with depression or ethnic background [15, 16, 18, 19, 41]. However, to emphasize no DIF was due to alcohol consumption, that is, whether or not, a participant used alcohol, revealed no bias concerning the PHQ-9 items.

Strengths and Limitations

Our study revealed some limitations. First, we did not use other measurements to concurrently validate the PHQ-9. Second, avoiding response bias was difficult. Because the respondents were health workers, some might underreport their real symptoms for fear of stigma. Our study; however, indicated some strengths. To the best of our knowledge, this was the first study to report the validity of the PHQ-9 using Rasch analysis with this substantial sample size of health workers. Thus, it would be likely to conclude that no DIF was observed among this population, which was somewhat comparable to the general population.

Conclusion

The PHQ-9 was, demonstrated by Rasch measurement model, shown to be a unidimensional structure with ordered response categories evaluating a single construct of depression with sufficient person and item reproducibility. However, the low person separation value and poor targeting showing on the person-person-item map suggested PHQ-9 might not be appropriate to measure depressive level among health care workers who seem to have low levels of depression. Easier items should be added to the scale to capture those with low depression but who may need clinical attention.

Abbreviations

CTT        Classical test theory
DIF        Differential Item Functioning
INFIT information-weighted fit statistics,
MNSQ Mean square
OUTFIT outlier-sensitive fit statistics
PCA Principal Component Analysis
PHQ Patient Health Questionnaire
Zstd Z-score standardized

Declarations

Ethics approval and consent to participate

This study was approved by the research ethics committee of the Faculty of Medicine, Chiang Mai University. All patients provided written informed consent to the study.

Consent for publication

Consent for publication is not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

Chaisiri Angkurawaranon is an Associate Editor of BMC Family Practice. All the authors declare that they have no competing interests.

Funding

This research was supported by the Faculty of Medicine Research Fund of Chiang Mai University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

SJ, TW, CA, WJ, and NW participated in the concept and design of the study. SJ, CA, and WJ collected data. NW and TW performed the statistical analyses. SJ, TW, CA, WJ, and NW drafted and edited the
manuscript. All authors made substantial contributions to interpret data and revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank all the participants who participate in the study.

References


20. Tennant A, Conaghan PG: **The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?** *Arthritis Care Res* 2007, **57**.


41. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL: Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *J Gen Intern Med*
2006, 21:547-552.


Figures
Figure 1

Person-Item Wright Map Legend: The persons are on the left of the dashed line, and items are located on the right of the dashed line. More able (depressed) persons are located at the top of the map. More difficult (severe) items are located at the top of the map. Each “#” represents 33 persons. Each “.” represents 1-32 persons (M = mean; S = 1 standard deviation from the mean; T = 2 standard deviations from the mean).
Figure 2

Person-Item Wright Map Legend: The curves for the PHQ-9 illustrating the range over which each of the 4 categories is most likely to be chosen. The red, blue, pink and blank curves on the graph represent the 0, 1, 2 and 3 and 4 PHQ-9 rating categories.