Neuropsychiatric Disease and Treatment

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Construct Validity and Differential Item Functioning of the PHQ-9 Among Health Care Workers: Rasch Analysis Approach

This article was published in the following Dove Press journal:
Neuropsychiatric Disease and Treatment

Surin Jiraniramai[1,*]
Tinakon Wongpakaran[2,*]
Chaisiri Angkurawaranon[1]
Wichuda Jiraporncharoen[1]
Nahathai Wongpakaran[2]

[1]Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Kingdom of Thailand; [2]Department of Psychiatry, Faculty of Medicine, Chiang Mai University, Chiang Mai, Kingdom of Thailand

*These authors contributed equally to this work

Purpose: 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.

Patients and Methods: In total, 3204 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.

Keywords: PHQ-9, Thai, Rasch analysis, differential item functioning, alcohol consumption

Introduction

Depression is considered the most common mental disorder and a major cause of disability in Thailand[1–3] and globally.[4] 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,[5] 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.[6,7] Health systems and healthcare workers worldwide are encountering tremendous stress because of the growing Coronavirus Disease 2019

Correspondence: Nahathai Wongpakaran
Geriatric Psychiatry Unit, Department of Psychiatry, Faculty of Medicine, Chiang Mai University., 110 Intawaroros Road, T. Sriphum, A. Muang, Chiang Mai, 50200, Kingdom of Thailand
Tel +66 53 935422 ext 320
Fax +66 53 935426
Email nahathai.wongpakaran@cmu.ac.th

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/NDT.S271987

(COVID-19) pandemic,[8] especially those on the frontline, migrant workers, and workers in contact with the public.[9] Reports have shown that screening for depression among these healthcare workers is increasing,[10] and healthcare workers significantly experience anxiety and depression symptoms.[11]

In Thailand, one study among health care workers revealed a high rate of depression at 21.5%.[12] Consequences of depression among health care workers could produce medical errors and affect treatment outcome.[13–15] Therefore, screening and grading depression symptom severity are important approaches to early detect and manage depression among health care workers.

Several tools have been used for screening depression including the Hospital Anxiety and Depression Scale (HADS),[16] Center for Epidemiologic Studies Depression Scale (CES-D),[17] self-rating depression Scale (SDS), Depression, Anxiety and Stress Scale-21 items (DASS-21).[18] While some have included anxiety, CES-D and The Patient Health Questionnaire (PHQ-9) measure only depressive symptoms. Besides, the different number of items used for each scale, the outcome as the prevalence of depression may differ across measurements. In comparison among tools, a study conducted among patients with cancer found that HADS had the widest measurement range, whereas the DASS-D had the narrowest. Based on the cutoff, PHQ-9 was easier to meet the criteria for mild depression than CES-D and DASS-21.[19]

PHQ-9, one of the most used tools, has screened for depression in various settings, demonstrating good sensitivity and specificity for depressive disorders.[20] It comprises nine items of diagnostic symptom criteria based on the DSM-IV with Likert scale responses, ie, 0 (not at all) to 4 (nearly every day).[21] The PHQ-9 has been translated into many languages including Thai.[22]

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).[23–25]

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.[23,26–29] Most studies reported that all nine items could form a unidimensional construct of overall depressive severity.

Rasch analysis is the science of developing, examining, and analyzing the performance and quality of measurement instruments that are completed by individuals.[30] The Rasch model offers procedures for constructing and revising instruments and documenting measurement properties of instruments (eg, reliability, construct validity). Contrasting with traditional classical test theory, the Rasch model allows investigators to make critical corrections when using raw test score data by transforming raw data (nonlinear, ordinal level data) to interval-level data, which then can be evaluated through the use of parametric tests.[31–33]

One way that the Rasch model assesses the quality of a measurement instrument is to evaluate the "fit" of items to the Rasch model based on the fact that items at the more difficult end of the variable should be harder to correctly answer than items at the easy end of the continuum. In addition, the unique person-item (Wright) map created by Rasch analysis displays the linear scale graphically, to suggest good test-item targeting.[34]

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,[26] and DIF related to sex among elderly populations, which was found with "sleep" and "self-blame" items.[25]

The PHQ-9 has been investigated for DIF in many subpopulations. While some found that DIF was unrelated to ethnic background,[35] language,[36] race,[37] sex[37,38] and level of education,[29] it found DIF related to age,[38,39] cognitive impairment,[40] a history of mania[41] and visual impairment.[26] In addition, age group plays a role concerning DIF. For instance, DIF related to sex was detected in an elderly population.[25]

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.[42,43] Alcohol users tend to have irrational beliefs,[44] as well as other psychiatric symptoms, such as anxiety or depression and cognitive impairment.[40] This casts doubt that alcohol users may view or assess depressive symptoms differently from nonalcohol users due to their distorted beliefs.[45]

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 workers 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, ie, 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.

Patients and Methods

Participants and Procedures

Health care workers were classified into 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%). This cross-sectional health care workers' study, conducted at Chiang Mai University (CMU), was approved by the Ethics Review Committee for Research in Human Subjects, Faculty of Medicine, Chiang Mai University Hospital. Ethics approval was obtained from the Faculty of Medicine, Chiang Mai University (No. 069/2012). It was a single site study from a survey of health care workers from Chiang Mai University Hospital. The first recruitment period was from January to February 2013 where about 56.3% of all health care workers responded to the survey. The second enrollment period was from March to May 2013. Additional recruitment strategies included posters and turning the desktop wallpaper of all computer-operated computers to provide information about the survey. The second recruitment period helped bring the proportion of overall respondents to 75%, and 3532 (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. A Thai National ID number was used to assign a study identification number and to ensure no duplicate responses. A detailed description of the study has been published.[46]

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.[22]

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".[47] 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, eg, 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.[48] 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.[49] In addition, fit statistics <0.6 indicate items overfit the model, usually because they share some components of meaning with other items.[33]

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 items; r ≤0.3 was considered acceptable.[50]

Item ordering, indicating that a higher severity of a symptom should score a higher category, was examined using the category function. The threshold estimates for a 4-category response option were 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 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 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 suggest the difference between the mean value of the mean person measure should be within one logit.[30] 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.[51]

Finally, reliability was evaluated using the 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 the item separation index. Separation value was less than 3 and the 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.[51]

All analyses were conducted using IBM SPSS for Windows, Version 22 (Chicago, IL, USA), STATA, Version 14 (College Station, TX: StataCorp LP) and Rasch models using WINSTEPS, Version 4.5.4 (Winsteps® Rasch Measurement, 2017).

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 cut-off point for clinical depression (a sum score of 10 or over).

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.

"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 items had a residual correlation >0.2. The most were items 2, "Feeling down" and 1, "Little interest" (r = 0.12). 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 (0.66 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 0.64 logits.

Table 1 Sociodemographic Characteristics of the Subjects

Characteristic	N(%) or Mean (SD)
Age (years), Mean (SD)	40.2 (10.7)
Sex: Female, n (%)	2471 (77.1)
Marital status, n (%)	
Single	1380 (43.2%)
Married (lived together, separated, divorced, widowed)	1817 (56.7%)
Educational status, n (%)	
Lower than Bachelor degree	1135 (35.4%)
Bachelor degree and higher	2069 (64.6%)
Job description	
Doctors, dentists, nurses and pharmacists	1350 (42.1%)
Other health professionals	621 (19.4)
Nonhealth professionals	1233 (38.5%)
Alcohol consumption, n (%)	988 (30.8%)
PHQ-9	
Sum score, mean (SD)	4.3 (3.2)
Median, Inter quartile range	4,4
Min-max	0–23
Severity of depression, n (%)	
None or minimal (0–4)	1832 (57.2)
Mild (5–9)	1173 (36.6)
Moderate (10–14)	182 (5.7)
Moderately severe (15–19)	13 (0.4)
Severe(20–27)	4 (0.1)

Abbreviations: SD, standard deviation; PHQ, Patient health questionnaire.

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, eg, "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

Table 2 Item Fit for PHQ-9

Item Description	Measure or Logits (SE)	Infit Mean Square	Infit Zstd	Outfit Mean Square	Outfit Zstd
1. Little interest	−1.55(0.4)	0.72	−9.90	0.71	−9.90
1. Feeling down	−0.87(0.4)	0.79	−8.57	0.78	−8.71
1. Trouble sleeping	−1.70(0.4)	1.27	9.17	1.29	9.72
1. Feeling tired	−0.87(0.4)	1.06	2.11	1.05	1.65
1. Poor appetite	−0.94(0.4)	1.22	7.65	1.18	6.12
1. Feeling bad about self	0.98(0.5)	1.01	0.51	0.90	−1.79
1. Trouble concentrating	−0.24(0.4)	0.96	−1.39	0.97	−1.06
8. Moving slowly	1.28(0.5)	1.04	1.32	0.87	−2.04
9. Better off dead	3.91(1.0)	1.13	1.48	0.70	−1.70

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

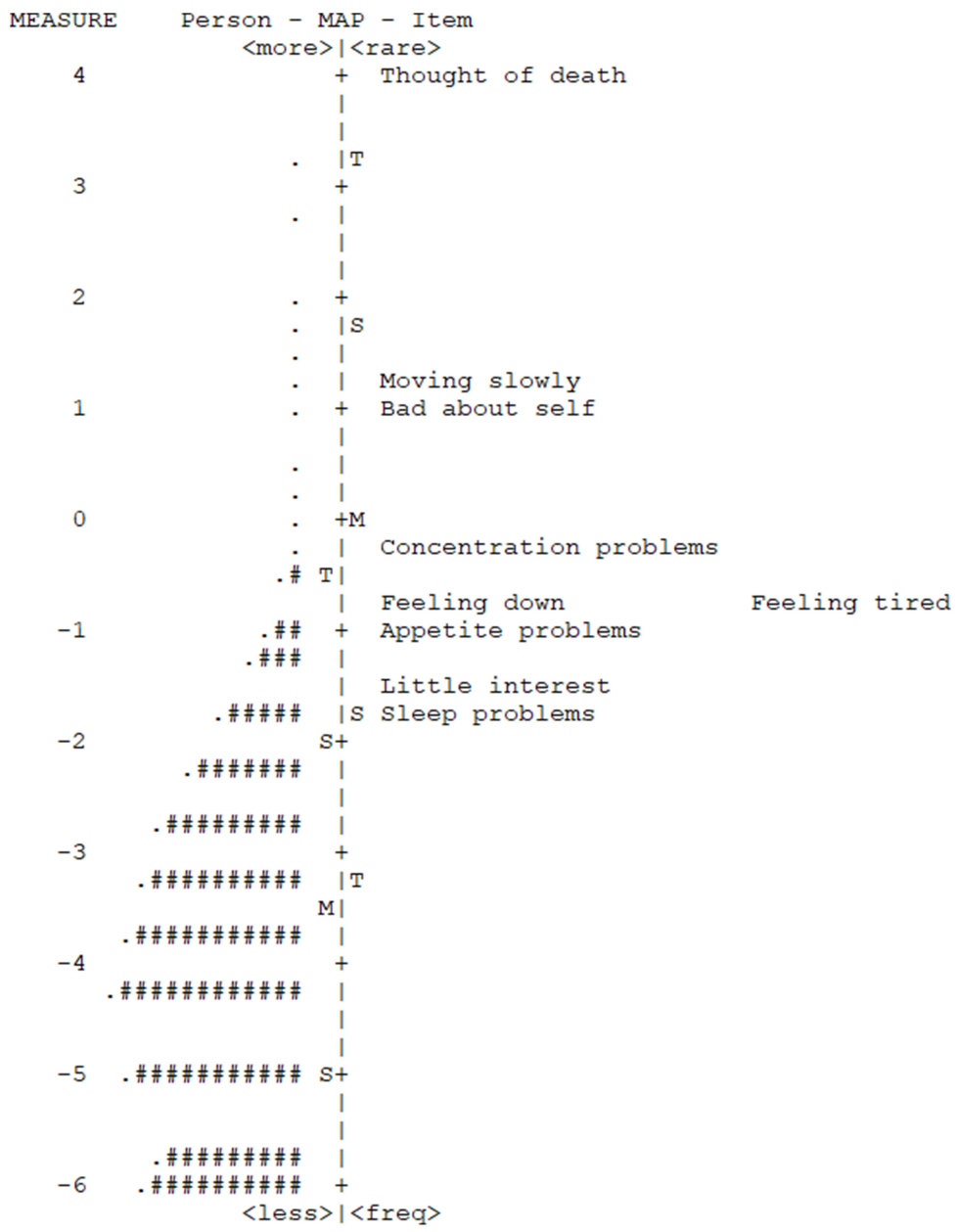

Figure 1 Person-item Wright Map.
Notes: The persons are on the left of the dashed line, and items are located on the right of the dashed line. More able (depressed) persons are located at the top of the map. More difficult (severe) items are located at the top of the map. Each "#" represents 33 persons.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

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.

Figure 2 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 = 0.80). 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

Table 3 Summary Statistics for the 4-Rating Scale Categories of PHQ- 9

Category	Frequency of Use	Percent	Observed Person Measure	Infit Mnsq	Outfit Mnsq	Andrich Threshold	Category Measure
0	16,679	58	−4.80	0.99	0.98	NONE	(−4.20)
1	10,685	37	−2.11	0.96	0.83	−3.08	−1.14
2	1277	4	−0.49	1.07	1.08	0.87	1.55
3	195	1	0.60	1.33	1.52	2.22	(3.47)

Abbreviations: infit, information-weighted fit statistics; outfit, outlier-sensitive fit statistics; Mnsq, mean square; PHQ, patient Health Questionnaire.

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, ie, unidimensionality and local independence, indicating that all items contributed to the same depression underlying construct,[23,39,52] but contrasted with other related studies.[25,27,28] This inconsistency may have contributed to the different characteristics of the studied sample, especially since none of the 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.[22,53] 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.[23,26,54] 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.[27,28]

Notably, item 9 is loosely related to latent dimension depression, leaving the question whether or not incorporating suicidality is useful in such a screening scale for depression.[25,28,55] Not only did it receive the lowest endorsement, one study showed that item 9 illustrated both item misfit and disordered threshold.[56] From this reason, item 9 was removed, and PHQ-8 became adopted for a screening tool of depression.[57,58] 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.[48]

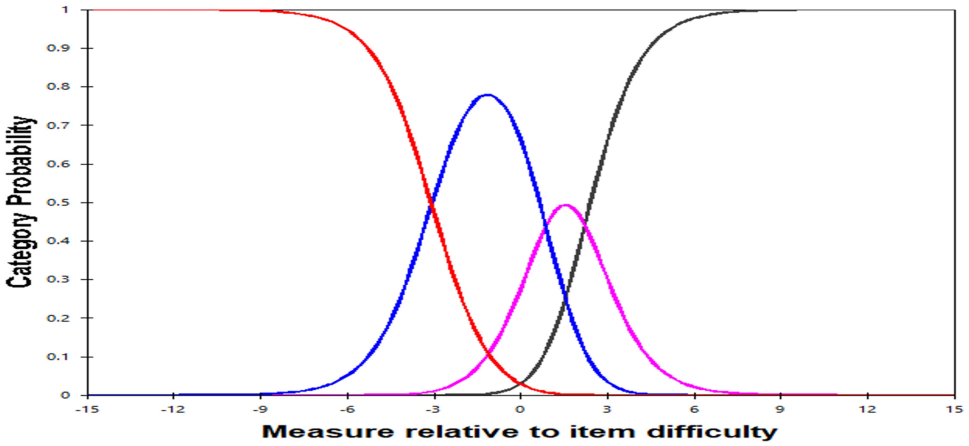

Figure 2 Categorical probability curves for PHQ-9 with no threshold disorder.
Notes: The curves for the PHQ-9 illustrating the range over which each of the 4 categories is most likely to be chosen. The red, blue, pink and blank curves on the graph represent the 0, 1, 2 and 3 and 4 PHQ-9 rating categories.

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.[25] 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.[27] This is to confirm that PHQ-9 functions as intended among clinical but not general people including health care workers. Therefore, PHQ-9 can still be used as 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. Depression is developed by the influence of biopsychosocial factor. Clinically, it may form at least part of a continuum of affective disorder, rather than a discrete disorder.[59] A mild form of depression precedes moderate and may become a severe form of depression when an individual's coping strategy fails. A screening that detects a mild form of depression is useful, but PHQ-9 is deemed unable to capture well enough among this population.

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.[19,25] This could be because those scales have more items that are able to cover the broader latent construct of a subject's depressive severity.

In terms of reliability, the person separation was not excellent, albeit acceptable (0.74), while the Cronbach alpha was 0.80. This may have contributed to the number of items being relatively low – more items may be needed.[51] 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.[60] 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, eg, visual impairment, the elderly, primary care with depression or ethnic background.[25,26,28,29,52] 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. Rasch analysis provides an opportunity to improve the measurement so that it would have better matched with the target population. PHQ-9 might not be appropriate enough to assess the depressive symptoms among nonclinical subjects such as healthcare workers. Other screening tools for depression, especially those with more items (eg, CES-D, DASS-21) should be trialed in further study.

Data Sharing Statement

The dataset is available upon reasonable request to Chaisiri Angkurawaranon (email: chaisiri.a@cmu.ac.th).

Ethics Approval and Consent to Participate

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the

Institutional Review Board (or Ethics Committee) of the Faculty of Medicine, Chiang Mai University. All patients provided written informed consent to the study.

Acknowledgments

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

Author Contributions

SJ, TW, CA, WJ, and NW participated in the concept and design of the study. SJ, CA, and WJ collected data. All authors contributed to data analysis, drafting or revising the article, have agreed on the journal to which the article will be submitted, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

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.

Disclosure

All the authors declare that they have no competing interests in this work.

References

1. Kongsuk T, Supanya S, Kenbubpha K, Phimtra S, Sukhawaha S, Leejongpermpoon J. Services for depression and suicide in Thailand. *WHO South-East Asia J Public Health*. 2017;6(1):34–38. doi:10.4103/2224-3151.206162
2. Wongpakaran T, Wongpakaran N, Pinyopornpanish M, et al. Baseline characteristics of depressive disorders in Thai outpatients: findings from the Thai study of affective disorders. *Neuropsychiatr Dis Treat*. 2014;10:217–223. doi:10.2147/NDT.S56680
3. Wongpakaran N, Wongpakaran T. Prevalence of major depressive disorders and suicide in long-term care facilities: a report from northern Thailand. *Psychogeriatrics*. 2012;12(1):11–17. doi:10.1111/j.1479-8301.2011.00383.x
4. Bauer M, Whybrow PC, Angst J, Versiani M, Möller HJ, Disorders WF. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, Part 2: maintenance treatment of major depressive disorder and treatment of chronic depressive disorders and subthreshold depressions. *World J Biol Psychiatry*. 2002;3(2):69–86. doi:10.3109/15622970209150605
5. WHO. Health Workers. WHO. Health Workers, Available from: https://www.who.int/whr/2006/06_chap1_en.pdf. Accessed January 29, 2019.
6. Hegney DG, Craigie M, Hemsworth D, et al. Compassion satisfaction, compassion fatigue, anxiety, depression and stress in registered nurses in Australia: study 1 results. *J Nurs Manag*. 2014;22(4):506–518. doi:10.1111/jonm.12160
7. Kim K, Lee S, Choi YH. Relationship between occupational stress and depressive mood among interns and residents in a tertiary hospital, Seoul, Korea. *Clin Exp Emerg Med*. 2015;2(2):117–122. doi:10.15441/ceem.15.002
8. Ng QX, De Deyn MLZQ, Lim DY, Chan HW, Yeo WS. The wounded healer: a narrative review of the mental health effects of the COVID-19 pandemic on healthcare workers. *Asian J Psychiatr*. 2020;54:102258. doi:10.1016/j.ajp.2020.102258
9. Giorgi G, Lecca LI, Alessio F, et al. COVID-19-related mental health effects in the workplace: a narrative review. *Int J Environ Res Public Health*. 2020;17(21):7857. doi:10.3390/ijerph17217857
10. Soltani S, Tabibzadeh A, Zakeri A, et al. COVID-19 associated central nervous system manifestations, mental and neurological symptoms: a systematic review and meta-analysis. *Rev Neurosci*. 2021. doi:10.1515/revneuro-2020-0108
11. Moitra M, Rahman M, Collins PY, et al. Mental health consequences for healthcare workers during the COVID-19 pandemic: a scoping review to draw lessons for LMICs. *Front Psychiatry*. 2021;12:602614. doi:10.3389/fpsyt.2021.602614
12. Kaewporndawan T, Chaiudomsom C. The prevalence and associated factors of depression among residents in training at Faculty of Medicine, Siriraj Hospital. 2014;41–50.
13. Deckard G, Meterko M, Field D. Physician burnout: an examination of personal, professional, and organizational relationships. *Med Care*. 1994;32(7):745–754. doi:10.1097/00005650-199407000-00007
14. Felton JS. Burnout as a clinical entity—its importance in health care workers. *Occup Med (Chic Ill)*. 1998;48(4):237–250. doi:10.1093/occmed/48.4.237
15. Ruotsalainen J, Serra C, Marine A, Verbeek J. Systematic review of interventions for reducing occupational stress in health care workers. *Scand J Work Environ Health*. 2008;34(3):169–178. doi:10.5271/sjweh.1240
16. Beneria A, Arnedo M, Contreras S, et al. Impact of simulation-based teamwork training on COVID-19 distress in healthcare professionals. *BMC Med Educ*. 2020;20(1):515. doi:10.1186/s12909-020-02427-4
17. Awano N, Oyama N, Akiyama K, et al. Anxiety, depression, and resilience of healthcare workers in Japan during the coronavirus disease 2019 outbreak. *Intern Med*. 2020;59(21):2693–2699. doi:10.2169/internalmedicine.5694-20
18. Maduke T, Dorroh J, Bhat A, Krvavac A, Are RH. We coping well with COVID-19?: a study on its psycho-social impact on front-line healthcare workers. *Mo Med*. 2021;118(1):55–62.
19. Lambert SD, Clover K, Pallant JF, et al. Making sense of variations in prevalence estimates of depression in cancer: a co-calibration of commonly used depression scales using rasch analysis. *J Natl Compr Canc Netw*. 2015;13(10):1203–1211. doi:10.6004/jnccn.2015.0149
20. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA*. 1994;272(22):1749–1756. doi:10.1001/jama.1994.03520220043029
21. Kroenke K, Spitzer RL, Williams JB, The PHQ-9. validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606–613. doi:10.1046/j.1525-1497.2001.016009606.x
22. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry*. 2008;8:46. doi:10.1186/1471-244X-8-46
23. Christensen KS, Oernboel E, Zatzick D, Russo J. Screening for depression: rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. *J Psychosom Res*. 2017;97:18–22. doi:10.1016/j.jpsychores.2017.03.117
24. Barthel D, Barkmann C, Ehrhardt S, Schoppen S, Bindt C, Group ICS. Screening for depression in pregnant women from Côte d'Ivoire and Ghana: psychometric properties of the Patient Health Questionnaire-9. *J Affect Disord*. 2015;187:232–240. doi:10.1016/j.jad.2015.06.042.

25. Forkmann T, Gauggel S, Spangenberg L, Brähler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis. *J Affect Disord*. 2013;148(2–3):323–330. doi:10.1016/j.jad.2012.12.019
26. Gothwal VK, Bagga DK, Sumalini R. Rasch validation of the PHQ-9 in people with visual impairment in South India. *J Affect Disord*. 2014;167:171–177. doi:10.1016/j.jad.2014.06.019
27. Horton M, Perry AE. Screening for depression in primary care: a Rasch analysis of the PHQ-9. *BJPsych Bull*. 2016;40(5):237–243. doi:10.1192/pb.bp.114.050294
28. Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. *J Affect Disord*. 2010;122(3):241–246. doi:10.1016/j.jad.2009.07.004
29. Zhong Q, Gelaye B, Fann JR, Sanchez SE, Williams MA. Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: a Rasch item response theory analysis. *J Affect Disord*. 2014;158:148–153. doi:10.1016/j.jad.2014.02.012
30. Boone WJ, Staver JR, Yale MS. *Rasch Analysis in the Human Sciences*. Kindle Edition ed. Springer; 2014.
31. 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;2007:57. doi:10.1002/art.23108
32. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. *Int J Nurs Stud*. 2009;2009:46. doi:10.1016/j.ijnurstu.2008.10.007
33. Bond TG. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences/Authored by Trevor G. Bond and Christine M. Fox*. Routledge, Taylor & Francis Group; 2015.
34. Boone WJ. Rasch analysis for instrument development: why, when, and how? *CBE Life Sci Educ*. 2016;15(4):rm4. doi:10.1187/cbe.16-04-0148
35. Galenkamp H, Stronks K, Snijder MB, Derks EM. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*. 2017;17(1):349. doi:10.1186/s12888-017-1506-9
36. Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD, Canadian Scleroderma Research G. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. *PLoS One*. 2012;7(12):e52028–e52028. doi:10.1371/journal.pone.0052028
37. Uebelacker LA, Strong D, Weinstock LM, Miller IW. Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychol Med*. 2009;39(4):591–601. doi:10.1017/S0033291708003875
38. Cameron IM, Crawford JR, Lawton K, Reid IC. Differential item functioning of the HADS and PHQ-9: an investigation of age, gender and educational background in a clinical UK primary care sample. *J Affect Disord*. 2013;147(1):262–268. doi:10.1016/j.jad.2012.11.015
39. Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can clinicians use the PHQ-9 to assess depression in people with vision loss? *Optom Vis Sci*. 2009;86(2):139–145. doi:10.1097/OPX.0b013e318194eb47
40. Boyle LL, Richardson TM, He H, et al. How do the PHQ-2, the PHQ-9 perform in aging services clients with cognitive impairment? *Int J Geriatr Psychiatry*. 2011;26(9):952–960. doi:10.1002/gps.2632
41. Weinstock LM, Strong D, Uebelacker LA, Miller IW. Differential item functioning of DSM-IV depressive symptoms in individuals with a history of mania versus those without: an item response theory analysis. *Bipolar Disord*. 2009;11(3):289–297. doi:10.1111/j.1399-5618.2009.00681.x
42. Hattingh HL, Hallett J, Tait RJ. 'Making the invisible visible' through alcohol screening and brief intervention in community pharmacies: an Australian feasibility study. *BMC Public Health*. 2016;16(1):1141. doi:10.1186/s12889-016-3805-3
43. Khadjesari Z, Newbury-Birch D, Murray E, Shenker D, Marston L, Kaner E. Online health check for reducing alcohol intake among employees: a feasibility study in six workplaces across England. *PLoS One*. 2015;10(3):e0121174. doi:10.1371/journal.pone.0121174
44. Camatta CD, Nagoshi CT. Stress, depression, irrational beliefs, and alcohol use and problems in a college student sample. *Alcohol Clin Exp Res*. 1995;19(1):142–146. doi:10.1111/j.1530-0277.1995.tb01482.x
45. Bridges K, Harnish R. Role of irrational beliefs in depression and anxiety. *Health*. 2010;2:862–877. doi:10.4236/health.2010.28130
46. Angkurawaranon C, Wisetborisut A, Jiraporncharoen W, et al. Chiang Mai University Health Worker Study aiming toward a better understanding of noncommunicable disease development in Thailand: methods and description of study population. *Clin Epidemiol*. 2014;6:277–286. doi:10.2147/CLEP.S65338
47. Walker ER, Engelhard G, Thompson NJ. Using Rasch measurement theory to assess three depression scales among adults with epilepsy. *Seizure*. 2012;21(6):437–443. doi:10.1016/j.seizure.2012.04.009
48. Wright BD, Linacre JM. Reasonable mean-square fit values. *Rasch Measurement Trans*. 1994;370–371.
49. Raîche G. Critical eigenvalue sizes in standardized residual principal component analysis. *Rasch Meas Transact*. 2005;19:1012.
50. Christensen KB, Makransky G, Horton M. Critical Values for Yen's Q(3): identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas*. 2017;41(3):178–194. doi:10.1177/0146621616677520
51. Linacre JM *Winsteps® Rasch measurement computer program User's Guide*. Winsteps.com. 2017.
52. 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(6):547–552. doi:10.1111/j.1525-1497.2006.00409.x
53. Van Orden K, Conwell Y. Suicides in late life. *Curr Psychiatry Rep*. 2011;13(3):234–241. doi:10.1007/s11920-011-0193-3
54. Woldetensay YK, Belachew T, Tesfaye M, et al. Validation of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression in pregnant women: afaan Oromo version. *PLoS One*. 2018;13(2):e0191782. doi:10.1371/journal.pone.0191782
55. Razykov I, Ziegelstein RC, Whooley MA, Thombs BD. The PHQ-9 versus the PHQ-8–is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the Heart and Soul Study. *J Psychosom Res*. 2012;73(3):163–168. doi:10.1016/j.jpsychores.2012.06.001
56. Adams RJ, Wu ML, Wilson M. The Rasch rating model and the disordered threshold controversy. *Educ Psychol Meas*. 2012;72(4):547–573. doi:10.1177/0013164411432166
57. McMahon AB, Arms-Chavez CJ, Harper BD, LoBello SG. PHQ-8 minor depression among pregnant women: association with somatic symptoms of depression. *Arch Womens Ment Health*. 2017;20(3):405–409. doi:10.1007/s00737-017-0715-z
58. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. 2009;114(1–3):163–173. doi:10.1016/j.jad.2008.06.026
59. Ng QX, Lim DY, Chee KT. Reimagining the spectrum of affective disorders. *Bipolar Disord*. 2020;22(6):638–639. doi:10.1111/bdi.12960
60. Wind SA, Gale JD. Diagnostic opportunities using rasch measurement in the context of a misconceptions-based physical science assessment. *Sci Educ*. 2015;99(4):721–741. doi:10.1002/sce.21172