Assessing the Construct Validity of the Patient Health Questionnaire-9 in Measuring Severity of Major Depressive Disorder

Himanshu Agrawal, MBBS; Timothy McAuliffe, PhD; Alexandra Johnson, MD; Quamaine Bond, MD; Leah Flanagan, MD; Rita Sieracki, MLS

ABSTRACT

Introduction: Developed by Spitzer et al in 1999, the Patient Health Questionnaire (PHQ-9) is a 9-item, self-administered instrument used to screen for depression. This screening tool has been validated by more than 25 studies. However, although the PHQ-9 asks respondents to rate the frequency of symptoms, most validation studies focus on symptom severity. The objective of this study was to assess the construct validity of PHQ-9 for detecting the severity of depression, as defined by the Diagnostic and Statistical Manual of Mental Disorders.

Methods: In part 1 of this study, 1408 outpatients across 14 family practice clinics were asked whether they scored the PHQ-9 based on symptom frequency, symptom severity, or a combination of both. In part 2, 87 mental health clinicians were asked how they typically interpret PHQ-9 scores.

Results: Of the 87 clinician responses, 79.3%, reported interpreting PHQ-9 scores based on severity of depressive symptoms, 4.6% based on frequency, and 16.1% based on a combination of severity and frequency. In striking contrast, among the 1408 patient responses, only 10.7% reported completing it based solely on severity, and 28.6% reported completing it based on frequency alone. An additional 60.7% reported completing the PHQ-9 based on a combination of frequency and severity.

Conclusions: The findings of this study indicate that the language used in the PHQ-9 may be interpreted differently by patients and clinicians. These differences may lead to conflation of symptom frequency and severity, resulting in misinterpretation of the PHQ-9 scores. The authors recommend revisions to the language of this screening tool to determine whether such changes improve alignment between patient responses and clinician interpretation.

INTRODUCTION

The Patient Health Questionnaire (PHQ)[1] is a self-administered version of the Primary Care Evaluation of Mental Disorders (PRIME MD)[2] diagnostic instrument for common mental disorders. The PHQ-9[1] is the 9-item depression module from the full PHQ that assesses depressive symptoms based on criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM IV).[3] The instrument asks, "Over the last 2 weeks, how often have you been bothered by any of the following problems? Use 'X' to indicate your answer," and directs participants to rate each item as follows: "0" (not at all), "1" (several days), "2" (more than half the days), or "3" (nearly every day). It is widely used in the United States[1,4-7] and globally,[8-11] and is utilized extensively within the authors' affiliated ambulatory clinics.[12-14] Numerous studies have concluded that the PHQ-9 is a valid[1,4-8] and reliable[5,9,10,15-18] screening instrument for assessing depression severity across various patient populations and geographies. Many studies correlate PHQ-9 scores of 5, 10, 15, and 20 with mild, moderate, moderately severe, and severe depression, respectively.[4] A score of 10 or higher may warrant treatment.[4]

Jiraniramai et al[6] noted that the "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." Regarding construct validity – that is, whether the PHQ-9 precisely assesses what it is intended to measure – McCord and Provost[4] noted that "the high level of heterogeneity in just these 9

• • •

Author affiliations: Department of Psychiatry and Behavioral Medicine, Medical College of Wisconsin (MCW), Milwaukee, Wisconsin (Agrawal, McAuliffe); School of Medicine, MCW, Milwaukee, Wisconsin (Johnson, Bond, Flanagan); Medical College of Wisconsin Libraries, MCW, Milwaukee, Wisconsin (Sieracki).

Corresponding author: Himanshu Agrawal, MBBS, DF-APA; Department of Psychiatry and Behavioral Medicine, Medical College of Wisconsin, 1155 N Mayfair Road, Milwaukee, WI 53226; email hagrawal@mcw.edu; ORCID ID 0000-0002-5343-7260

items reflects one of the most significant problems with the categorical diagnostic paradigm…there are 18351 different patterns that produce exactly a score of 10…there are 101 different patterns with which two different patients can each earn a score of 10 without sharing a single symptom."

Although Levis et al[16] report that the "PHQ-9 is often used to estimate depression prevalence, it overestimates major depression prevalence substantially," they conclude that the heterogenic nature of presenting symptoms poses a statistical challenge in individual study analyses. Therefore, they recommend that "estimates of depression prevalence should be based on validated diagnostic interviews designed for determining case status; users should evaluate published reports of depression prevalence to ensure that they are based on methods intended to classify major depression."

The actual language of the PHQ-9 asks patients to rate the frequency of symptoms, not severity. In fact, symptom severity is not explored until the end of the questionnaire ("If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?"), with options ranging from "not difficult at all," to "extremely difficult." While symptom frequency over time may correlate with illness severity, these concepts are not interchangeable.

We hypothesized that this discrepancy may obscure the reliability and validity of the PHQ-9 as a tool for determining depression severity, potentially leading to overtreatment or undertreatment of depressive symptoms.

METHODS

Database Searches and Study Selection

A systematic literature search was designed and conducted by a medical librarian (RS), with input from the research team (HA, LF). The search strategy included Medical Subject Headings (MeSH) terms and keywords related to "Patient Health Questionnaire-9" or "PHQ-9" combined with "symptom assessment," "validity," or "systematic reviews." Searches were conducted using MEDLINE (Ovid), PsycINFO, and Scopus on July 27, 2021, using advanced search techniques relative to each database. Results were limited to English-language publication and managed in EndNote for duplicate removal. The complete search strategy is presented in the Appendix.

The Medical College of Wisconsin Human Research Protections Program approved all study procedures. Following Institutional Review Board (IRB) approval, informed consent forms and the survey (see Appendix) were distributed across 14 MCW/Froedtert outpatient clinics in the greater Milwaukee, Wisconsin, metropolitan area. These included 2 specialty behavioral health clinics and 12 primary care clinics.

Selection of Participants

From April 2022 through October 2022, patients visiting these clinics were asked to complete a voluntary, anonymous 1-question survey at check-out. After obtaining informed consent, research staff confirmed that participants had not completed the survey at a previous visit. Participants were asked if they had completed the PHQ-9 during that visit (an image of the PHQ-9 questionnaire was provided for reference). Participants were then asked, "When you answered the questions, did you score your symptoms based on severity of symptoms (how much you experience them); frequency of symptoms (how often you experience them); or both frequency and severity (a combination of how much and how often you experience them)."

For the clinician component of the study, we recruited participants via a national listserv of mental health clinicians. The post was an invitation to take a voluntary, anonymous 1-question survey that asked, "When you look at the scores of a PHQ-9, do you interpret the score as severity of depressive symptoms, as frequency of depressive symptoms, or as some combination of severity and frequency of depressive symptoms?"

Definition of Validity

The term "construct validity" is widely used and carries diverse connotations. The definition originates with Cronbach and Meehl,[19] who in 1955 discussed the challenges in quantifying construct validity and identified several contributing factors, including validation procedures, group differences, the nomological network, and the application of scientific methodology to complex testing. Campbell[20] posits that construct validity is crucial for understanding what a test measures and ensuring that it accurately reflects its theoretical construct. In a subsequent paper,[21] Campbell and Fiske differentiated between convergent validity (the degree to which theoretically related measures are actually related) and divergent validity (the degree to which a test does not correlate too highly with measures from which it should differ). Messick[22] emphasizes that valid inference in psychological assessment involves a comprehensive evaluation, including a unified concept of validity, theoretical rationale, and social consequences.

After considering these factors, we utilized the definition of validity provided by the Standards for Educational and Psychological Testing,[23] often considered the gold standard: "Validity is the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed."

RESULTS

Each project member collected survey responses from assigned clinics and entered them into a secure central data bank. Incomplete responses were excluded

Patient Survey

A total of 1408 participants who had filled out the PHQ-9 com-

Table 1. Patient Health Questionnaire-9 (PHQ-9) Scoring – Overall[a]

	N	%	95%CI
Severity	151	10.7	9.2–12.5
Frequency	402	28.6	26.2 – 31.0
Both	855	60.7	58.1 – 63.3
TOTAL	1408	100.0	

[a]Responses are to the question, "When you look at the scores of a PHQ-9, do you interpret the score as severity of depressive symptoms, as frequency of depressive symptoms, or as some combination of severity and frequency of depressive symptoms?

An examination by site (clinic) of the basis on which patients scored their symptoms showed significant variation between sites.

Table 2. Patient Health Questionnaire-9 (PHQ-9) Scoring Basis by Site

Site	Severity N (%)	Frequency N (%)	Both N (%)	Total N
Clinic A	22 (14.5)	40 (26.3)	90 (59.2)	152
Clinic B	40 (6.9)	156 (26.8)	387)66.4)	583
Clinic C	3 (7.0)	14 (32.6)	26 (60.5)	43
Clinic D	0 (0.0)	0 (0.0)	2 (100.0)	2
Clinic E	0 (0.0)	0 (0.0)	2 (100.0)	2
Clinic F	19 (22.1)	33 (38.4)	34 (39.5)	86
Clinic G	8 (9.2)	17 (19.5)	62 (71.3)	87
Clinic H	22 (27.8)	32 (40.5)	25 (31.6)	79
Clinic I	7 (8.6)	17 (21.0)	57 (70.4)	81
Clinic J	11 (6.2)	49 (27.5)	118 (66.3)	178
Clinic K	10 (16.9)	27 (45.8)	22 (37.3)	59
Clinic L	3 (21.4)	3 (21.4)	8 (57.1)	14
Clinic M	1 (4.8)	8 (38.1)	12 (57.1)	21
Clinic N	5 (23.8)	6 (28.6)	10 (47.6)	21
TOTAL	151 (10.7)	402 (28.6)	855 (60.7)	1408

There is a statistically significant association between Clinic and how patients scored their symptoms on the PHQ-9 (P <.001).

Table 3. Clinician Survey Responses[a]

	N	%	95% CI
Severity of depressive symptoms	69	79.3	69.3–87.3
Frequency of depressive symptoms	4	4.6	1.3–11.4
Some combination of severity and frequency of depressive symptoms	14	16.1	9.1–25.5
TOTAL	87	100.0	

[a]Responses are to the question, "When you look at the scores of a PHQ-9, do you interpret the score as severity of depressive symptoms, as frequency of depressive symptoms, or as some combination of severity and frequency of depressive symptoms?"

Table 4. Patient and Clinician Basis of Scoring and Interpreting PHQ-9[a]

	Patient N (%)	Clinician N (%)
Severity of symptoms	151 (10.7)	69 (79.3)
Frequency of symptoms	402 (28.6)	4 (4.6)
Some combination of severity and frequency of symptoms	855 (60.7)	14 (16.1)
Total	1408	87

[a]$\chi^2 = 307.3$; df = 2; P < .001 (chi-square test of independence).

pleted the follow-up survey. Three sites were excluded: Clinic O and Clinic P (each contributed only 2 completed surveys) and Clinic Q (because collaboration could not be established).

All follow-up surveys (n = 1408) were deemed valid. Overall, 10.7% of patients scored their symptoms based on severity, 28.6% based on frequency, and 60.7% based on a combination of both frequency and severity. We calculated 95% confidence intervals (CIs) for all proportions (Table 1). The proportion of patients scoring symptoms based on severity was 0.10 (95% CI, 0.09-0.12).

Additionally, the basis for scoring symptoms differed significantly by clinic (P < .001; Table 2).

Clinician Interpretations of PHQ-9 Scores

Of the100 clinicians surveys received, 99 were valid. Twelve clinicians reported that they did not use the PHQ-9. Of the 87 clinicians who reported using the instrument, 79% (95% CI, 69%-87%) interpreted PHQ-9 scores based as reflecting severity of depressive symptoms, 4% (95% CI, 1%-11%) as reflecting frequency, and 16% (95% CI, 9%-25%) as reflecting a combination of both severity and frequency (Table 3).

These results indicate a marked discrepancy: while up to 87% of clinicians interpreted PHQ-9 scores as reflecting symptom severity, up to 89% of patients scored their symptoms based on frequency (combining "frequency" and "both") rather than severity alone (Table 4).

DISCUSSION

In this study of 1408 outpatients, 151 participants (11%) reported completing the PHQ-9 based solely on symptom severity; 402 (28%) based on frequency, and 855 (60%) reported using a combination of both frequency and severity.

The language used in the PHQ-9 ("Over the last 2 weeks, how often have you been bothered by any of these problems") indicates that the tool measures frequency, not severity. However, studies examining the construct validity of the PHQ-9 have focused on the severity of symptoms, not the frequency. Even if PHQ-9 scores were shown to accurately reflect depression severity, this would suggest positive predictive value rather than construct validity. In other words, although the PHQ-9 explicitly asks about frequency, if the resulting score accurately reflects severity, it does so despite the instrument's wording, not because of it. Furthermore, only 11% of the participants in this study reported that the PHQ-9 facilitated their rating of the severity of their depression.

Contrary to prior studies that compared PHQ-9 with other depression questionnaires (eg, Beck's Depression Inventory,[24]Hamilton Depression Rating scale[25]), when we took a direct approach to examining construct validity – by asking patients whether they completed the PHQ-9 based on severity – only 12% of the 1408 participants indicated that they had. By comparison, 28% completed the PHQ-9 solely based on

frequency, and a majority (60%) reported using a combination of frequency and severity, a construct that remains subjectively defined by respondents.

These findings also raise concerns about the sensitivity and specificity of the PHQ-9 in detecting depression severity. Because clinicians tend to interpret PHQ scores as indicators of symptom severity, several clinical scenarios may arise. For example, if a patient completes the PHQ-9 based on frequency rather than severity, a low score may be inaccurately interpreted as indicating low severity, which may put the patient at risk of undertreatment. Conversely, if the patient completes the PHQ-9 based on frequency of symptoms, a high score may be interpreted as high severity, which may put the patient at risk of overtreatment or unnecessary treatment.

Based on these findings, we suggest that the language used in the PHQ-9 may be interpreted differently by patients versus clinicians. Although prior studies have demonstrated consistency between the PHQ-9 scores and other standard depression instruments, the construct validity of the PHQ-9—its ability to measure what it is intended to measure—may be improved. We propose revising the language of the PHQ-9 to explicitly assess severity rather than frequency (see Appendix). As Campbell and Fiske[21] state, construct validity should be strengthened rather than removed, as it provides a comprehensive framework for evaluating test validity. Accordingly, we recommend that a similar study be conducted using the revised questionnaire to assess whether construct validity improves.

It should be noted that any revised questionnaire should be used solely as a screening tool and must be accompanied by a comprehensive clinical evaluation, including consideration of DSM diagnostic criteria , such as exclusion of symptoms attributable to other effects/conditions. Use of a revised questionnaire as a standalone diagnostic tool could create many issues. For example, a patient who scores a "3" on Question 9 (suicidal ideation) and minimal scores on other items could receive a total score that underrepresents the seriousness of the situation.

Limitations

This study had several limitations. It was conducted only among adult outpatients, limiting its generalizability to other settings (eg, inpatient hospitals) or pediatric populations. Additionally, the study was confined to a specific geographic area (the greater Milwaukee metropolitan area), which may limit applicability to other geographic and cultural populations.

Although most participants reported completing the PHQ-9 based on a combination of frequency and severity, the survey asks respondents to elaborate on what they meant by this combination. As a result, these responses are subjective and open to a wide variation in interpretation. The study could have been strengthened by exploring participants' interpretations of this combined approach.

Finally, the finding that the basis on which patients scored their symptoms differed by clinic complicates interpretation of the overall proportions. Without additional data on clinic-level factors that could explain the differences, this association should be interpreted cautiously.

CONCLUSIONS

The findings of this study indicate that the language used in the PHQ-9 is interpreted differently by patients and clinicians. These differences may lead to conflation of symptom frequency and severity, potentially resulting in misinterpretation of PHQ-9 scores. The authors recommend revising the language of this screening tool and conducting further study to determine whether such revisions reduce the discrepancies identified in this analysis.

Financial disclosures: None declared.

Funding/support: None declared.

Appendix: Available at www.wmjonline.org.

REFERENCES

1. 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

2. Spitzer RL, Williams JBW, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care: The PRIME-MD 1000 study. *JAMA*. 1994;72(22):1749–1756.

3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*. American Psychiatric Association; 1994.

4. McCord DM, Provost RP. Construct validity of the PHQ-9 depression screen: correlations with substantive scales of the MMPI-2-RF. *J Clin Psychol Med Settings*. 2020;27(1):150-157. doi:10.1007/s10880-019-09629-z

5. Dajpratham P, Pukrittayakamee P, Atsariyasing W, Wannarit K, Boonhong J, Pongpirul K. The validity and reliability of the PHQ-9 in screening for post-stroke depression. *BMC Psychiatry*. 2020;20(1):291. doi:10.1186/s12888-020-02699-6

6. Jiraniramai S, Wongpakaran T, Angkurawaranon C, Jiraporncharoen W, Wongpakaran N. Construct validity and differential item functioning of the PHQ-9 among health care workers: Rasch analysis approach. *Neuropsychiatr Dis Treat*. 2021;17:1035-1045. doi:10.2147/NDT.S271987

7. Martin A, Rief W, Klaiberg A, Braehler E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *Gen Hosp Psychiatry.* 2006;28(1):71-77. doi:10.1016/j.genhosppsych.2005.07.003

8. Arrieta J, Aguerrebere M, Raviola G, et al. Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in Rural Chiapas, Mexico: a cross-sectional study. *J Clin Psychol.* 2017;73(9):1076-1090. doi:10.1002/jclp.22390

9. Indu PS, Anilkumar TV, Vijayakumar K, et al. Reliability and validity of PHQ-9 when administered by health workers for depression screening among women in primary care. *Asian J Psychiatr.* 2018;37:10-14. doi:10.1016/j.ajp.2018.07.021

10. Sebera F, Vissoci JRN, Umwiringirwa J, Teuwen DE, Boon PE, Dedeken P. Validity, reliability and cut-offs of the Patient Health Questionnaire-9 as a screening tool for depression among patients living with epilepsy in Rwanda. *PLoS One.* 2020;15(6):e0234095. Published 2020 Jun 12. doi:10.1371/journal.pone.0234095

11. Cameron IM, Crawford JR, Lawton K, et al. Assessing the validity of the PHQ-9, HADS, BDI-II and OIDS-SR16 in measuring severity of depression in a UK sample of primary care patients with a diagnosis of depression: study protocol. *Prim Care Community Psychiatr.* 2008;13(2):67-71.

12. Molinaro J, Banerjee A, Lyndon S, et al. Reducing distress and depression in cancer patients during survivorship. *Psychooncology*. 2021;30(6):962-969. doi:10.1002/pon.5683.

13. Ozieh MN, Garacci E, Walker RJ, Palatnik A, Egede LE. The cumulative impact of social determinants of health factors on mortality in adults with diabetes and chronic kidney disease. *BMC Nephrol.* 2021;22(1):76. doi:10.1186/s12882-021-02277-2

14. Karafin MS, Singavi A, Hussain J, et al. Predictive factors of daily opioid use and quality of life in adults with sickle cell disease. *Hematology*. 2018;23(10):856-863. doi:10.1080/10245332.2018.1479997

15. Richardson T, Wrightman M, Yeebo M, Lisicka A. Reliability and score ranges of the PHQ-9 and GAD-7 in a primary and secondary care mental health service. *J Psychosoc Rehabil Ment Health*. 2017;4(2):237-240. doi:10.1007/s40737-017-0090-0

16. Levis B, Benedetti A, Ioannidis JPA, et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. *J Clin Epidemiol.* 2020;122:115-128.e1. doi:10.1016/j.jclinepi.2020.02.002

17. Poongothai S, Pradeepa R, Ganesan A, Mohan V. Reliability and validity of a modified PHQ-9 item inventory (PHQ-12) as a screening instrument for assessing depression in Asian Indians (CURES-65). *J Assoc Physicians India*. 2009;57:147-152.

18. Nallusamy V, Afgarshe M, Shlosser H. Reliability and validity of Somali version of the PHQ-9 in primary care practice. *Int J Psychiatry Med*. 2016;51(6):508-520. doi:10.1177/0091217417696732

19. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull.* 1955;52(4):281-302. doi:10.1037/h0040957

20. Campbell DT. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *Am Psychol*. 1960;15(8):546-553. doi: 10.1037/h0048255

21. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56(2):81-105. doi: 10.1037/h0046016

22. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50(9):741-749. doi: 10.1037/0003-066X.50.9.741

23. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Validity. In: American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* American Educational Research Association; 2014.

24. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry*. 1961;4:561–571. doi:10.1001/archpsyc.1961.01710120031004

25. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23(1):56–62. doi:10.1136/jnnp.23.1.56

advancing the art & science of medicine in the midwest

WMJ

WMJ (ISSN 2379-3961) is published through a collaboration between The Medical College of Wisconsin and The University of Wisconsin School of Medicine and Public Health. The mission of *WMJ* is to provide an opportunity to publish original research, case reports, review articles, and essays about current medical and public health issues.

Visit www.wmjonline.org to learn more.