

Interpretation Issues With the Patient Health Questionnaire Instructions

Margarita Panayiotou, PhD; Josip Razum, PhD; Gudrun Eisele, PhD; Shirley B. Wang, PhD; Eiko I. Fried, PhD; Zachary D. Cohen, PhD

 Supplemental content

IMPORTANCE Versions of the Patient Health Questionnaire (PHQ) such as PHQ-2, PHQ-8, and PHQ-9, are among the leading assessment tools for depression in research and clinical practice. However, important questions remain about its validity, particularly whether responses reflect symptom frequency or the degree to which symptoms are bothersome.

OBJECTIVE To investigate how participants respond to the PHQ, described as a severity measure by its developers, with instructions about being bothered by symptoms and response options focused on symptom frequency.

DESIGN, SETTING, AND PARTICIPANTS This study used data from a general population sample collected via Amazon Mechanical Turk (MTurk) and a clinical sample with medium to high depression from the Operationalizing Digital PhenoTyping in the Measurement of Anhedonia (OPTIMA) study. Data were collected between 2022 and 2023.

MAIN OUTCOMES AND MEASURES After completing the PHQ-8, participants' interpretation of instructions was assessed via 3 questions: (1) how they would respond to the PHQ sleep item in a hypothetical scenario where they overslept nearly every day but were comfortable with oversleeping; (2) whether they had based their earlier PHQ responses on symptom frequency, being bothered by symptoms, or both; and (3) how they would answer the PHQ in the future, based on these same 3 options.

RESULTS The study sample consisted of a general population sample collected via MTurk (n = 503; mean [SD] age, 40.63 [13.62] years; 253 male, 245 female, 3 transgender, 2 other) and a clinical sample with medium to high depression from the OPTIMA study (n = 349; mean [SD] age, 33.44 [12.23] years; 120 male, 216 female, 5 transgender, 8 other). In the hypothetical oversleeping scenario, only 54.7% (n = 275; MTurk) and 15.5% (n = 53; OPTIMA) of participants interpreted the PHQ as instructed (ie, the frequency with which the problems bothered them). When asked how they had responded to the PHQ, only 21.3% (n = 107; MTurk) and 11.7% (n = 40; OPTIMA) of participants interpreted the instructions as instructed and only 22.3% (n = 112; MTurk) and 9.9% (n = 34; OPTIMA) reported they would do so in the future, indicating stability in their interpretation. The current study also found that the PHQ-8 validity varied depending on how participants interpret its instructions.

CONCLUSIONS AND RELEVANCE This study identified widespread misinterpretation of the PHQ instructions across community and clinical samples, raising doubts about its validity for both research and clinical decision-making.

Author Affiliations: Manchester Institute of Education, University of Manchester, Manchester, United Kingdom (Panayiotou); Faculty of Psychology, University of Iceland, Reykjavik, Iceland (Razum); Ivo Pilar Institute of Social Sciences, Zagreb, Croatia (Razum); Department of Neurosciences, KU Leuven, Leuven, Belgium (Eisele); Department of Psychology, Yale University, New Haven, Connecticut (Wang); Clinical Psychology, Leiden University, Leiden, the Netherlands (Fried); Psychology Department, University of Arizona, Tucson (Cohen).

Corresponding Author: Zachary D. Cohen, PhD, University of Arizona, Tucson, AZ 85721 (cohenzd@arizona.edu).

JAMA Psychiatry. doi:10.1001/jamapsychiatry.2025.3796
Published online December 17, 2025.

The 9-item Patient Health Questionnaire (PHQ-9)¹ and its abridged versions (PHQ-2,² PHQ-8³) are free widely used scales designed to assess *DSM* criteria for major depressive disorder. They are leading tools for depression screening and diagnosis, translated into more than 100 languages, and required by major funders such as the National Institutes of Health and Wellcome Trust as part of their data harmonization efforts.⁴

Despite widespread use, there are concerns about reliability and validity. First, PHQ versions have not been validated for some crucial applications, such as tracking treatment response.⁵ Second, meta-analyses showed lower diagnostic accuracy in studies without developer involvement (sensitivity = 0.48; 95% CI, 0.41-0.91).⁶ Third, similar to other depression scales,⁷ the PHQ-9 was multidimensional in some samples, with the somatic symptoms forming a separate factor.⁸ Finally, qualitative work highlighted problems with interpreting items that ask about multiple things at once (ie, double-barreled and triple-barreled items) and confusing frequency with severity when responding.⁹ The current study extends these findings to quantify the prevalence and clinical importance of interpretation issues with the PHQ instructions in a general population and clinical sample. This is important, as the predominant focus of existing research^{5,6,10} on psychometric testing does not take into consideration that some of the noted PHQ problems could stem from interpretation issues that have not yet been systematically explored.

Methods

Participants

We used data from US general population and clinical samples with written consent (eAppendix 1 in Supplement 1). The general sample was collected in March 2022 via CloudResearch (Prime Research Solutions LLC) on Amazon Mechanical Turk (MTurk). The clinical sample (participants with a PHQ-8 score >9) was part of the Operationalizing Digital PhenoTyping in the Measurement of Anhedonia (OPTIMA) study^{11,12} conducted from October 2022 through December 2023. Both studies, including the present study, were approved by the University of California-Los Angeles (UCLA) institutional review board. This study followed the American Association for Public Opinion Research (AAPOR) reporting guideline.

Measures

In addition to the original PHQ-8³ collected in the MTurk study, both studies used versions of the PHQ-8 adapted from Hadjodis et al,¹³ which separated PHQ double-barreled items (eg, poor appetite or overeating) into 2 and included libido and irritability as additional symptoms. The adapted version used in the OPTIMA study retained the PHQ-8 original instructions (how often one has been bothered by each problem over the past 2 weeks) and the 0 (not at all) to 3 (nearly every day) response scale. The adapted version used in the MTurk study asked about the past week. eAppendix 1 in Supplement 1 includes additional details. Total scores were calculated equivalently to the PHQ-8 by selecting the highest of the separated

Key Points

Question Do responses to the Patient Health Questionnaire (PHQ) reflect symptom frequency or the degree to which symptoms are bothersome?

Findings In this survey study including 503 US general population and 349 clinical samples, most participants misinterpreted the PHQ instructions when judging a hypothetical scenario, when reporting how they responded to items, and how they would respond in the future. Approximately half of the general population (54.7%) and fewer than one-fifth of clinical participants (15.5%) interpreted the PHQ as intended.

Meaning Results of this study suggest that the PHQ is widely misinterpreted, raising concerns about its validity for research and clinical decision-making.

items and omitting the added symptoms. The OPTIMA participants also completed the 15-item Quick Inventory of Depressive Symptomatology (QIDS)¹⁴ measure of depression (eAppendix 1 in Supplement 1).

After completing the PHQ-8, participants answered 3 questions about interpreting its instructions. First, they considered a hypothetical scenario where they overslept nearly every day for a week but were comfortable with oversleeping (ie, not bothered by it, as per the PHQ-8 instructions). Then they responded to the PHQ-8 “sleeping too much” item, where 0 (“not at all”) would reflect “not bothered by oversleeping.” Second, they indicated whether their prior PHQ-8 responses were based on (1) bothered (“I answered mostly based on ‘bothered’” [the frequency with which the problems bothered me]), (2) frequency (“I answered mostly based on ‘how often’” [the frequency of the problems]), or (3) both (“took both into consideration”). Indicating “frequency” or “both” would indicate a misinterpretation of the instructions. Third, they reported how they would answer PHQ-8 questions in the future, after reflecting on the difference between “frequency” and “bothered.”

Statistical Analysis

The frequencies of these responses were explored for each sample. All analyses were conducted using R version 4.5.0 (R Foundation for Statistical Computing).

Results

Interpretation Groups

The study sample consisted of a general population sample collected via MTurk (n = 503; mean [SD] age, 40.63 [13.62] years; 253 male, 245 female, 3 transgender, 2 other) and a clinical sample with medium to high depression from the OPTIMA study (n = 349; mean [SD] age, 33.44 [12.23] years; 120 male, 216 female, 5 transgender, 8 other).

For question 1 (Table, Figure 1 and Figure 2), only 54.7% (n = 275) of participants in the general sample and 15.5% (n = 53) of participants in clinical sample selected the correct response (0); 40% (n = 201) and 75.1% (n = 257) selected the

Table. Participant Characteristics and Findings

Characteristic	Sample	
	General (MTurk)	Clinical (OPTIMA)
No.	503	349 ^a
Age range, y	18-65	18-64
Mean (SD)	40.63 (13.62)	33.44 (12.23)
Gender identity, No.		
Male	253	120
Female	245	216
Transgender	3	5
Other ^b	2	8
Hypothetical scenario, No. (%)		
Bothered by (0 ["not at all"])	275 (54.7)	53 (15.5)
Frequency (3 ["nearly every day"])	201 (40.0)	257 (75.1)
Other	27 (5.4)	32 (9.4)
Total	503 (100)	342 (100)
Reported interpretation, No. (%)		
Bothered by ^c	107 (21.3)	40 (11.7)
Frequency	239 (47.5)	161 (47.1)
Both ^d	157 (31.2)	141 (41.2)
Total	503 (100)	342 (100)
Future reporting, No. (%)		
Bothered by	112 (22.3)	34 (9.9)
Frequency	190 (37.8)	121 (35.4)
Both ^d	201 (40.0)	187 (54.7)
Total	503 (100)	342 (100)

Abbreviations: MTurk, Amazon Mechanical Turk; OPTIMA, Operationalizing Digital Phenotyping in the Measurement of Anhedonia; PHQ, Patient Health Questionnaire.

^a Seven participants had missing data for the interpretation item in the OPTIMA study.

^b Does not identify as male, female, or transgender.

^c This group reflects the correct interpretation of the PHQ instructions.

^d Both = both frequency and bothered by.

maximum score (3). For question 2, only 21.3% (n = 107) in the general sample and 11.7% (n = 40) of participants in the clinical sample responded consistent with the intended interpretation of the PHQ-8 (ie, bothered by the problems). For question 3, only 22.3% (n = 112) in the general sample and 9.9% (n = 34) in the clinical sample indicated that their future PHQ-8 answers would be based on the "bothered by" interpretation.

Psychometric Evidence

We examined whether the PHQ-8 validity varied by interpretation group. First, in 1-way analysis of covariance controlling for gender and age, the interpretation group was unrelated to PHQ-8 total scores in either sample. In OPTIMA, QIDS scores substantially differed by group: the frequency group reported lower depression scores than the group that took both interpretations into account (post hoc comparison Cohen *d* = 0.31).

Second, we assessed convergent validity by correlating, using Pearson *r*, PHQ-8 and QIDS items with overlapping content. Eight of the 11 QIDS items (89%) examined had notable differences in their correlations between the fre-

quency and bothered groups, suggesting that the validity of the PHQ-8 items may vary depending on interpretation group. Details are presented in eAppendix 3 in Supplement 1.

Discussion

Self-reported depression measures have become increasingly common in research and practice as they offer accessible and cost-effective screening. However, the score on a depression measure is only valid if individuals understand and respond to its questions as intended.⁵ In the current study we found inconsistent interpretation of PHQ-8 instructions ("how often have you been bothered by the following problems"), with the lowest proportion of participants selecting the correct response ("the frequency with which the problems bothered me") among the 3 possible interpretations. These findings highlight the need to examine item and instruction comprehension and response processes as a mainstay of measure development. Such data are missing for many clinical instruments, posing a threat to the valid assessment of psychopathology broadly.

Our findings may raise certain concerns: participants tend to skip the instructions of self-report measures¹⁵ and in our case misinterpreted them even after they had explicitly considered them. Furthermore, most patients in the clinical sample misinterpreted the hypothetical scenario, more so than the general sample, possibly due to greater cognitive difficulties.¹⁶ This may not only make group-level findings uninterpretable but may also result in unreliable clinical cutoffs, and consequently flawed treatment decisions.

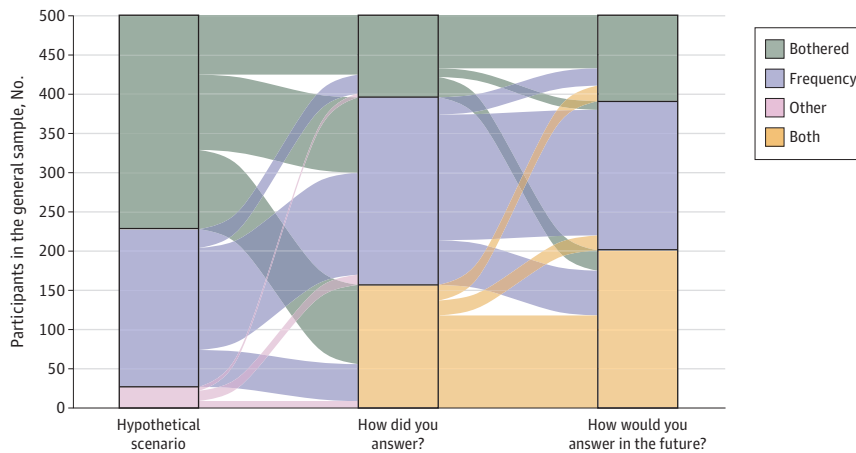
The PHQ is described as a severity measure,¹ yet its instructions emphasize being bothered by symptoms while its response options focus on frequency. While recent work¹⁷ showed that adapting the PHQ to measure either severity or frequency leads to comparable psychometric properties, our findings indicate that the PHQ's combination of both created ambiguity and inconsistent interpretation, with important clinical and validity implications.

Future work must consider clearer alternatives. Omitting "bothered by" from the instructions may improve consistency but would turn the PHQ into a measure of symptom frequency rather than severity. Researchers and clinicians must therefore be mindful of their assessment aims. If the goal is to capture severity, a different measure or a modified response scale (as was done in other PHQ-related measures¹⁸) may be more appropriate.

Limitations

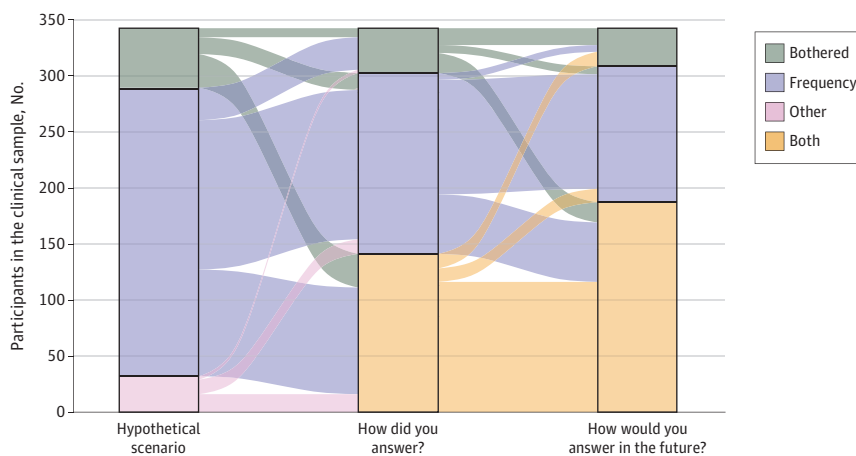
This study has limitations. Our findings require external validation and replication in other samples (eg, outside the post-pandemic era), settings (eg, community, inpatient, and primary care), and languages, with other depression measures, as well as PHQ versions such as the original PHQ-8, given that we used a slightly adapted version. Furthermore, longitudinal designs and consideration of other potential confounders (eg, cognitive functioning) are needed.

Figure 1. Patient Health Questionnaire Interpretation Frequencies in the Amazon Mechanical Turk (MTurk) General Sample



The MTurk sample included 503 participants.

Figure 2. Patient Health Questionnaire Interpretation Frequencies in the Operationalizing Digital Phenotyping in the Measurement of Anhedonia (OPTIMA) Clinical Sample



The OPTIMA sample included 342 participants.

Conclusions

This survey study found substantial misinterpretation. The current findings highlight longstanding problems in self-report mental health assessment, including limited attention to in-

structions, an overemphasis on psychometric testing at the expense of validity, minimal user involvement in measure development, and continued reliance on outdated tools in urgent need of revision. Without these, the PHQ's use rests on a fragile foundation that might not bear the weight of research and clinical decision-making.⁵

ARTICLE INFORMATION

Accepted for Publication: October 3, 2025.

Published Online: December 17, 2025.
doi:10.1001/jamapsychiatry.2025.3796

Author Contributions: Drs Panayiotou and Cohen had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Panayiotou and Razum shared first authorship.

Concept and design: Panayiotou, Razum, Fried, Cohen.

Acquisition, analysis, or interpretation of data: Panayiotou, Razum, Eisele, Wang, Cohen.

Drafting of the manuscript: Panayiotou, Razum,

Eisele, Wang, Fried.
Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Panayiotou, Eisele, Wang.
Administrative, technical, or material support: Cohen.

Supervision: Cohen.

Conflict of Interest Disclosures: Dr Cohen reported grants from Wellcome Leap to additional contributor Michelle G. Craske, PhD, at the UCLA Depression Grand Challenge, at which work on "Deep phenotyping and genetic analysis of anhedonia" (the OPTIMA study) was supported by Wellcome Leap as part of the Multi-Channel Psych

Program during the conduct of the study. No other disclosures were reported.

Data Sharing Statement: See Supplement 2.

Additional Contributions: The authors acknowledge the UCLA Depression Grand Challenge specifically to thank Michelle G. Craske, PhD (co-director of the Depression Grand Challenge and Principal Investigator of the OPTIMA study), for her support of the 2 studies from which data for these analyses were drawn. She did not receive any compensation.

Additional Information: This article was conceptualized at "Measurement is the New Black" (MITNB) consortium meeting at the Dutch

Research Council (NOW)-funded Lorentz Center workshop in Leiden, the Netherlands in November 2024. Work on "Deep phenotyping and genetic analysis of anhedonia" (OPTIMA study) was supported by Wellcome Leap as part of the Multi-Channel Psych Program. During this work, Dr Eisele was supported by a junior postdoctoral fellowship by the Research Foundation Flanders (1223725N). The funder did not give input on the design or conduct of the most relevant components of this study (core data for questions about participants' interpretation of the PHQ), as this was a side project of the senior author (Cohen). The funder also did not play any role in the collection, management, analysis, and interpretation of the data, and played no role in the preparation of the manuscript. However, the funder gave input in the design and conduct of the parent OPTIMA study, in which they played no role in the collection, management, analysis, and interpretation of the data, and played no role in the preparation of manuscripts, including this one. The funder of the OPTIMA study reviewed and approved this manuscript prior to its original submission. Research Foundation Flanders played no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Kroenke K, Spitzer RL, Williams JBW. 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. Löwe B, Kroenke K, Gräfe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J Psychosom Res*. 2005;58(2):163-171. doi:10.1016/j.jpsychores.2004.09.006
3. Kroenke K, Strine TW, Spitzer RL, Williams JBW, 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
4. Wolpert M. Funders agree first common metrics for mental health science. *Linked In page*. 2020. Accessed January 7, 2025. <https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mental-health-science-wolpert/>
5. Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. *Nat Rev Psychol*. 2022;1(6):358-368. doi:10.1038/s44159-022-00050-2
6. Manea L, Boehnke JR, Gilbody S, Moriarty AS, McMillan D. Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. *BMJ Open*. 2017;7(9):e015247. doi:10.1136/bmjopen-2016-015247
7. Fried EI. Are more responsive depression scales really superior depression scales? *J Clin Epidemiol*. 2016;77:4-6. doi:10.1016/j.jclinepi.2016.05.004
8. Lamela D, Soreira C, Matos P, Morais A. Systematic review of the factor structure and measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. *J Affect Disord*. 2020;276:220-233. doi:10.1016/j.jad.2020.06.066
9. Malpass A, Dowrick C, Gilbody S, et al. Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *Br J Gen Pract*. 2016;66(643):e78-e84. doi:10.3399/bjgp16X683473
10. Hlynsson JI, Skúlason S, Andersson G, Carlbring P. Why are we still using the PHQ-9? A historical review and psychometric evaluation of measurement invariance. *Psychiatr Q*. 2025. Published online September 3, 2025. doi:10.1007/s11126-025-10208-9
11. Rotstein NM, Cohen ZD, Welborn A, et al. Investigating low intensity focused ultrasound pulsation in anhedonic depression-A randomized controlled trial. *Front Hum Neurosci*. 2025;19:1478534. Published online March 23, 2025. doi:10.3389/fnhum.2025.1478534
12. Akre S, Cohen ZD, Welborn A, et al. Comparing self reported and physiological sleep quality from consumer devices to depression and neurocognitive performance. *NPJ Digit Med*. 2025; 8(1):92. doi:10.1038/s41746-025-01493-6
13. Haddox D, Cohen DE, Fried EI, Cohen DE. PHQ variants. *Open Science Framework*. 2025. Accessed January 7, 2025. <https://osf.io/w4rj9/overview>
14. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54(5):573-583. doi:10.1016/S0006-3223(02)01866-8
15. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J Exp Soc Psychol*. 2009;45(4):867-872. doi:10.1016/j.jesp.2009.03.009
16. Matcham F, Simblett SK, Leightley D, et al; RADAR-CNS Consortium. The association between persistent cognitive difficulties and depression and functional outcomes in people with major depressive disorder. *Psychol Med*. 2023;53(13):6334-6344. doi:10.1017/S0003291722003671
17. Niileksela CR, Jones NB. Measurement equality of frequency and severity item response options on depression and generalized anxiety scales. *Assessment*. 2023;30(6):2016-2028. doi:10.1177/10731911221134599
18. Kroenke K, Spitzer RL, Williams JB. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. *Psychosom Med*. 2002;64(2):258-266. doi:10.1097/O0006842-200203000-00008