Reliability and cross-group equivalence of the 8-item version of the Patient Health Questionnaire (PHQ-8) to screen for depression in adults with and without chronic pain: a nationally representative U.S. population study

Jennifer S. De La Rosa[1,2]

Greg T. Chism[1,3]

Katie E. Herder[1,4]

Chung Jung Mun [5,6]

Rachel V. Aaron[7]

[1]Comprehensive Center for Pain and Addiction, University of Arizona Health Sciences; [2]Department of Family and Community Medicine, University of Arizona College of Medicine-Tucson; [3]College of Information Science, University of Arizona, Tucson, AZ, USA; [4]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA; [5]Edson College of Nursing and Health Innovation, Arizona State University; Tempe, AZ, USA [6]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA; [7]Department of Physical Medicine and Rehabilitation, Johns Hopkins School of Medicine, Baltimore, MD, USA

Corresponding author: Jennifer S. De La Rosa, jschult1@arizona.edu

Author Note

Jennifer S. De La Rosa, PhD https://orcid.org/0000-0001-6291-5618
Greg T. Chism, PhD https://orcid.org/0000-0002-5478-2445
Katie E. Herder https://orcid.org/0000-0001-5699-8082
Chung Jung Mun https://orcid.org/0000-0002-2439-8100
Rachel V. Aaron https://orcid.org/0000-0002-2182-8566

SUMMARY

Background The 8-item Patient Health Questionnaire (PHQ-8) is the most widely implemented clinical depression screener in the world and is increasingly influential in population surveillance of unmet mental health needs. However, in the context of chronic pain (CP), depression screening could be challenging due to overlapping somatic symptoms. Concerns have been raised that inaccurate attribution of CP symptoms to depression could lead to overestimation of depressive symptom severity or even artifactual inflation of depression prevalence estimates. Results of several small, non-representative studies are inconclusive; suitability of the PHQ-8 to screen for depression in those with CP remains unclear. We aimed to determine the reliability and cross-group equivalence of PHQ-8 in adults with and without CP.

Methods This cross-sectional observational survey study used representative data from the 2019 National Health Interview Survey. Descriptive analyses, confirmatory factor analyses, tests of measurement invariance and cross-group equivalence, and item response theory analyses were completed.

Findings The final sample contained 30,983 respondents. Measurement invariance and cross-group equivalence was observed at the configural, metric, and scalar levels. Prevalence of clinically significant depressive symptoms was 3.5% in adults without CP, 20.1% in those with CP, and 34.8% in high-impact chronic pain (HICP). No evidence consistent with overestimation of depression prevalence or severity in the contexts of CP or HICP was observed.

Interpretation The PHQ-8 measures depressive symptoms equivalently in the context of chronic pain, supporting the continued use of PHQ-8 to screen for depression and to support epidemiological surveillance in adults with and without CP. Empirical evidence suggests that CP-associated disparities are very unlikely to result from measurement issues more likely to be reflective of true differences in prevalence and severity of depressive symptoms in this high-risk group.

Funding Comprehensive Center for Pain & Addiction, University of Arizona Health Sciences.

RESEARCH IN CONTEXT

Evidence before this study

Depression screening tools have been implemented into general health settings worldwide; today, most cases of major depression are identified and addressed in primary or urgent care settings. Chronic pain frequently co-occurs with depression, and chronic pain may be especially prevalent in settings most likely to implement depression screening is part of routine clinical practice.

However, screening for depression among those with chronic pain could be complicated by overlapping somatic symptoms that are part of diagnostic criteria for depression and often occur with chronic pain—e.g. sleep disturbance, fatigue, appetite changes, and psychomotor slowing. Furthermore, concerns about criterion contamination are not exclusive to somatic symptoms: chronic pain is also associated with cognitive and affective symptoms that similarly overlap with diagnostic criteria for depression. Systematic bias in depression screening among those with CP could lead to overestimation of depressive symptom severity in clinical settings or artificial inflation of depression prevalence among those living with chronic pain.

Several studies have examined small, non-representative patient samples; findings are inconclusive. The reliability and equivalence of standard depression screening tools to measure depressive symptoms in the context of chronic pain has remained persistently unclear.

Added value of this study

The PHQ-8 is the most widely implemented depression screening tool in the world, and it holds an increasingly influential role in epidemiological surveillance of unmet mental health needs. If PHQ-8 functions differently in those with chronic pain, clinical decisions, scientific studies, and even epidemiological surveillance tools could be systematically biased. This representative cross-sectional survey study of 30,983 U.S. adults is the first to determine the reliability and cross-group equivalence of the PHQ-8 in those with and without CP: addressing a persistent knowledge gap with important implications for clinical practice, scientific research,

and global public health. Cross-group equivalence and measurement invariance were observed between those with and without CP at the configural, metric, and scalar levels. These results provide clear evidence that PHQ-8 is reliable, consistent, and functions equivalently in adults with and without CP.

Implications of all the available evidence

Empirical evidence did not support with the hypothesis that measurement artifacts contribute to elevated depressive symptoms in those with CP. Findings should enhance confidence among clinicians and researchers that PHQ-8 is appropriate to screen for depressive symptoms in the clinical context of CP and that the interpretation of screening results is straightforward. Findings underscore the verifiable existence of profound mental health disparities experienced by U.S. adults with CP, including increased prevalence of depression, increased severity of depressive symptoms, reduced likelihood that depressive symptoms will be identified, and reduced likelihood of receiving needed mental health support.

1. INTRODUCTION

Depression and chronic pain (CP) commonly co-occur in general and patient populations. Among adults who screen positive for depression, the majority (an estimated 56%) have co-occurring CP,[1] while approximately 40% of adults with CP have co-occurring depression.[2] About half of these—an estimated 20.1% of adults with CP—have clinically significant depressive symptoms that have yet to be effectively addressed.[1] The co-occurrence of CP and depressive symptoms is associated with substantially higher prevalence of functional limitations in work, daily activities, and social participation, compared to those with CP alone or those with depressive symptoms alone.[1] Addressing the mental health needs of those with CP is essential to functional resilience and optimal quality of life,[3] yet CP has been associated with 40% lower odds of receiving needed mental health treatment.[4] Therefore, unmet mental health needs of those with CP is a recognized as a key priority for global population health and a fundamental human right.[5,6]

However, several overlapping somatic symptoms that are part of diagnostic criteria for depression—sleep disturbance, fatigue, appetite changes, and psychomotor slowing—commonly occur with CP. Concerns about criterion contamination between pain and depression are not limited to somatic symptoms, as chronic pain is an unpleasant sensory and emotional experience known to be associated with negative affective and cognitive states.[7] Concerns have been raised that CP symptoms could be inaccurately attributed to depression, potentially resulting in overestimation of depressive symptom severity or even artifactual inflation of depression prevalence estimates among those with CP.

Several studies without non-focal comparison groups have examined the functioning of depression screeners in chronic pain using small, non-representative patient samples. Findings have been inconclusive with some studies demonstrating suitability of depression screening tools in those with CP and others finding support for various types of differential functioning. We

direct the interested reader to an excellent review of prior studies examining issues in the assessment of depression in the context of chronic pain.[8]

The Patient Health Questionnaire (8-item and 9-item versions are equivalent) is the most frequently implemented depression screener worldwide, further, epidemiological estimates based on the PHQ-8 are used to support development of mental health policy, including the allocation of resources to research and health services.[9] Therefore, the possibility that PHQ-8 could be systematically biased in the context of CP warrants rigorous investigation. While it has demonstrated excellent validity and reliability across diverse geographies, settings, and contexts,[10] its equivalence in the context of CP remains unclear. If PHQ-8 functions differently in CP, clinical decisions, scientific research, and prevalence estimates could be biased. This study used representative U.S. population data to evaluate the reliability and equivalence of PHQ-8 to screen for depressive symptoms in the context of CP, addressing a persistent knowledge gap with important implications for clinical practice, research, and global public health.

2. METHODS

2.1 Study design and analysis plan

Preliminary descriptive analyses examined prevalence of positive PHQ-8 screens by CP status, distributions of PHQ-8 scores and item-level responses grouped by CP and depression status **(Figure 1)** and assessed whether CP is associated with differences in the proportional contribution of somatic symptoms in PHQ-8 scores **(Figure 2)**. Next, the factor structure, reliability, cross-group equivalence and measurement invariance of PHQ-8 were evaluated, using the Chi-Square test statistic, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Reliability was assessed using Cronbach's alphas, calculated using standard factor loadings from a weighted factor analysis for categorical items, and McDonald's Omega coefficients calculated using the weighted variance of each item and of the total items **(Table 1)**.

CFI and TLI values greater than 0.95; RMSEA values lower than 0.06; and SRMR values lower than 0.08 were considered indicative of good fit.[11] Cronbach's alphas and McDonald's Omega coefficients greater than 0.80 and less than .90 were considered indicative of acceptable reliability.[12,13]

Third, standardized factor loadings and item-discrimination parameters using graded response models were assessed (**Table 2).** Discrimination parameters greater than or equal to 1.70 were considered "quite satisfactory;" those greater than 1.35 and less than 1.70 were considered "good."[14,15] Item Response Theory (IRT) Test Information Functions (TIFs) and Item Information Functions (IIFs) were evaluated **(Figure 3)**.

Analyses were completed using R version 4.4.2. and several packages were used: *ggplot*2, *survey, lavaan,* and *mirt.*

2.2 Data

The 2019 National Health Interview Survey (NHIS) data (n=31,997) is sampled and weighted to represent 244.6 million adults in the U.S. population.[16] NHIS data are routinely used to estimate prevalence, incidence, trends, and disparities related to depressive symptoms and CP.[17–19]

2.3 Measures

Chronic Pain (CP) was measured using the item: "*In the past 3 months, how often did you have pain? Would you say never, some days, most days, or every day*?" Respondents (Rs) who answered "Most days" or "Every day" were considered to have CP. High-impact CP was defined using the item: "*Over the past 3 months, how often did your pain limit your life or work activities? Would you say never, some days, most days, or every day?"* Rs with CP who answered "Most days" or "Every day" were considered to have **High-Impact Chronic Pain (HICP)**; Rs with CP who answered "never" or "some days" were considered to have **Lower-Impact Chronic Pain (LICP)**. Definitions of CP and HICP aligned with the International

Association for the Study of Pain (IASP) classification of chronic pain for International Classification of Diseases (ICD-11).[7] Though CP and HICP constitute this study's primary analytic focus, their logically nested relationship could be confusing. Lower-Impact Chronic Pain was included in select reporting to facilitate contextualized understanding of mean estimates between CP and HICP and to support representative, non-redundant visualizations of the U.S. adult population.

PHQ-8 Score was measured by summing the 8-item Patient Health Questionnaire (PHQ-8). The PHQ-8 asks respondents to answer: "*Over the last 2 weeks, how often have you been bothered by any of the following problems?"* **PHQ-1 Anhedonia:** "*Little interest or pleasure in doing things"* ; **PHQ-2 Sadness/Blues:** "*Feeling down, depressed, or hopeless"*; **PHQ-3 Sleep:** *"Trouble falling or staying asleep, or sleeping too much*"; **PHQ-4 Fatigue: "***Feeling tired or having little energy*"; **PHQ-5 Appetite:** "*Poor appetite or overeating*"; **PHQ-6 Self-Blame:** "*Feeling bad about yourself – or that you are a failure or have let yourself or your family down*"; **PHQ-7 Concentration: "***Trouble concentrating on things, such as reading the newspaper or watching television*"; and **PHQ-8 Psychomotor Function: "***Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual*." Responses are ordinal: (0) "Not at all," (1) "Several Days," (2) "More than Half the Days," and (3) "Nearly Every Day." **Positive** s**creens** were determined using the standard cut point for PHQ-8 (≥10).[20] **Negative screens** included those with **absent/minimal symptoms** (PHQ scores 0-4) or **subclinical/mild symptoms** (PHQ-8 scores 5-9); **positive screens** include those with **moderate symptoms** (PHQ scores 10-14) or **severe symptoms** (PHQ-8 scores 15-24).[21] Categorization of PHQ-8 items as somatic or cognitive/affective symptoms was aligned with that reflected in prior literature:[22,23] **somatic score** was calculated by summing items 3, 4, 5, and 8— sleep, fatigue, appetite, and psychomotor functioning; **cognitive/affective score** was calculated by summing items 1, 2, 6,

and 7—anhedonia, sadness, self-blame, and concentration. For Rs with non-zero PHQ-8 scores, **somatic symptoms proportion** was calculated as the somatic symptom score divided by the PHQ-8 total score. Rs with PHQ-8 scores of zero were coded as zero for the proportion.

2.4 Missing Data

A total of 1,009 observations (3.15% of data) were missing (responses of “Refused,” “Not Ascertained,” or “Don't Know”) from PHQ-8, CP, or both. Considering study data characteristics of large sample size, small amount of missing data, and minimal expectation of precision and bias loss in the context of observational data, which suggest that deviations from the missing at random (MAR) assumption are unlikely to affect validity of estimates, complete case analysis was used.[24] We evaluated means for key study variables comparing observations with missing data and those without missing data; consistent means provided additional support that the MAR assumption is justifiable.

3. RESULTS

3.1 Preliminary Descriptive Analyses

<u>Comparing Prevalence of Positive PHQ-8 screens by CP status</u>

The prevalence of positive PHQ-8 screens was 6.87% in the U.S. adult general population. U.S. adults with CP or HICP had significantly higher prevalence of positive screens compared to those without CP (NoCP): prevalence of positive screens was 3.5% in NoCP compared to 20.1% in CP— reflecting prevalence of 11.5% in LICP and 34.8% in HICP (all p-values <0.001). Compared to NoCP, positive screens were 5 times more prevalent in CP, and nearly 10 times more prevalent in HICP.

<u>Comparison of PHQ-8 scores by CP status</u>

In the U.S. adult general population, the mean PHQ-8 score was 2.48. Those with CP or HICP had significantly higher scores compared to those without CP. The mean score was 1.77 in NoCP compared to 5.25 in CP: reflecting means of 3.85 in LICP and 7.71 in HICP (all p-values <.001).

Among U.S. adults who screened negative, the mean PHQ-8 score was 1.61; the mean score was 1.33 in NoCP compared to 2.91 in CP—reflecting means of 2.58 in those with LICP and 3.72 in those with HICP (all p-values <.001). Among those with negative screens, the distributions of PHQ-8 scores differed qualitatively across NoCP, CP, and HICP. Absent/minimal depressive symptoms (PHQ-8 scores 0-4) were the modal screening outcome across all groups, however absent/minimal symptoms were more prevalent in NoCP. The prevalence of absent/minimal symptoms was 90.88% in NoCP compared to 72.18% in CP — reflecting prevalence of 77.04% in LICP and 60.58% in HICP (all p-values <.001) **(Figure 1, lower right panel).**

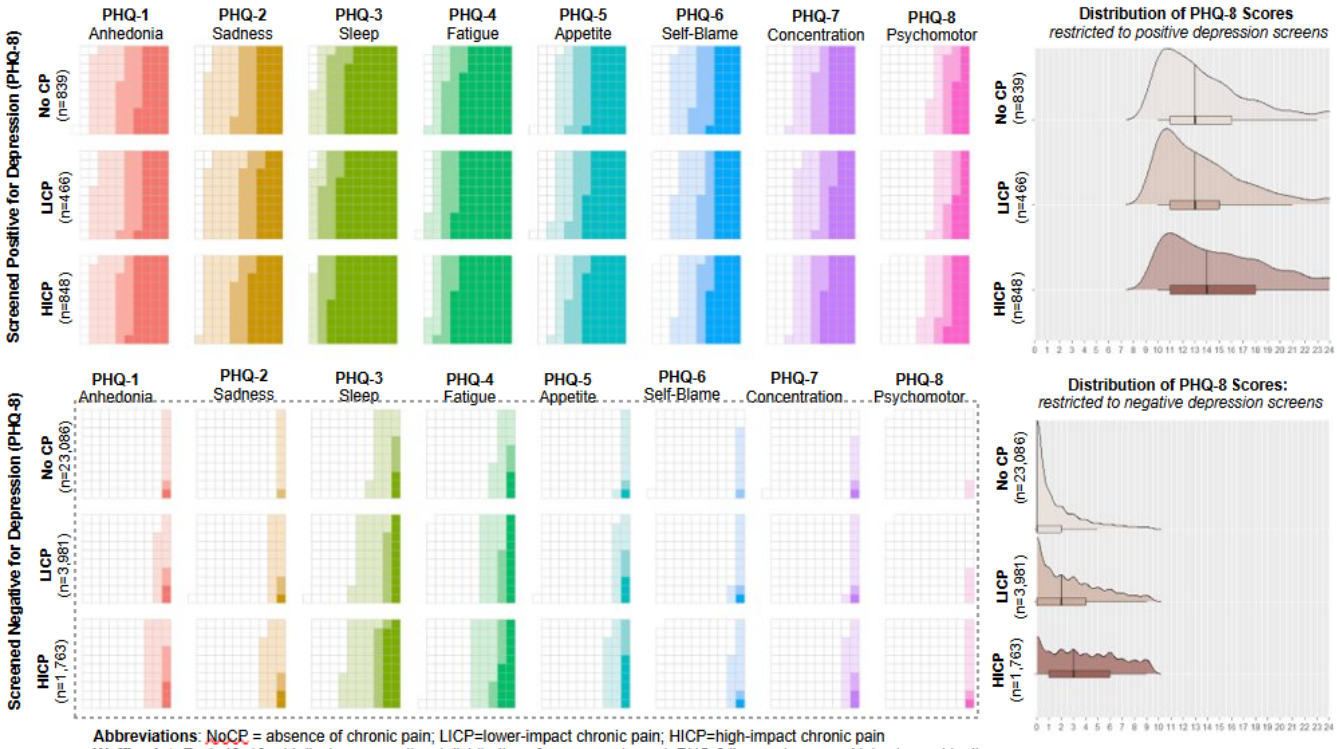
Among U.S. adults who screened positive, the mean PHQ-8 score was 14.29; the mean was 13.83 in NoCP compared to 14.61 in CP — reflecting means of 13.65 in LICP and 15.12 in HICP (all p-values <.001). Among those with positive screens, the distributions of PHQ-8 scores were qualitatively similar across NoCP, CP, and HICP **(Figure 1, upper right panel)**. While statistically significant CP-associated differences in mean PHQ-score were observed among those with positive screens, the magnitude of the difference was considerably attenuated in this group compared to those who screened negative. Moderate depressive symptoms (PHQ-8 scores 10-14) were the modal screening outcome across all groups, however severe symptoms (PHQ-8 scores 15-24) were disproportionately prevalent in CP. The prevalence of severe symptoms was 34.6% in NoCP and 43.7% in those with CP: reflecting prevalence of 33.9% in LICP and 49.3% in HICP.

Comparison of individual PHQ-8 items by CP and screening status

Among U.S. adults who screened negative, CP was associated with statistically and substantively significant increases in mean severity of all 8 items (each ranging from 0 to 3): anhedonia (0.12 vs. 0.28), sadness (0.11 vs. 0.26), sleep (0.33 vs. 0.72), fatigue (0.41 vs. 0.89), appetite(0.14 vs 0.33) self-blame (0.09 vs. 0.18), concentration (0.08 vs. 0.18) and psychomotor functioning (0.03 vs. 0.08); the magnitude of the CP-associated differences was intensified in HICP (all p-values <.001).

However, *among U.S. adults who screened positive,* CP-associated differences in symptom severity were statistically and substantively attenuated. CP was associated with statistically significantly (p<.001) increases in mean severity of sadness (1.65 vs. 1.78), sleep problems (2.24 vs. 2.37), and fatigue (2.33 vs. 2.60); mean differences in severity of anhedonia, appetite, self-blame, concentration, and psychomotor functioning did not reach statistical significance. Statistically significant increases in severity were observed in HICP for seven of the PHQ-8 items (p<.001); only concentration did not reach significance. We visualized response distributions for each item grouped by screening and CP status. (**Figure 1, panels at left**). It is worth noting that in all six combinations of screening status (negative vs. positive) and CP status (NoCP vs. LICP vs. HICP), the mean severity of sleep and fatigue was at least twice that of the mean severity of other items (p<.001 across all CP status and screening).

Figure 1. Comparing item-level responses and total PHQ-8 scores, grouped into mutually exclusive and exhaustive combinations of CP status and depression screening status to facilitate visualization of the total U.S. population.



Abbreviations: NoCP = absence of chronic pain; LICP=lower-impact chronic pain; HICP=high-impact chronic pain
Waffle plot: Each 10x10 grid displays proportional distribution of responses to each PHQ-8 item, sub-grouped into six combinations of depression screening status & CP status. Each small square within grids square represents 1% of subgroup.
Fill: No color = "Not at all", Light color ="Several Days", Medium color = "More than Half the Days", Darkest color ="Nearly Every Day"

Data Source: National Center for Health Statistics, National Health Interview Survey, 2019

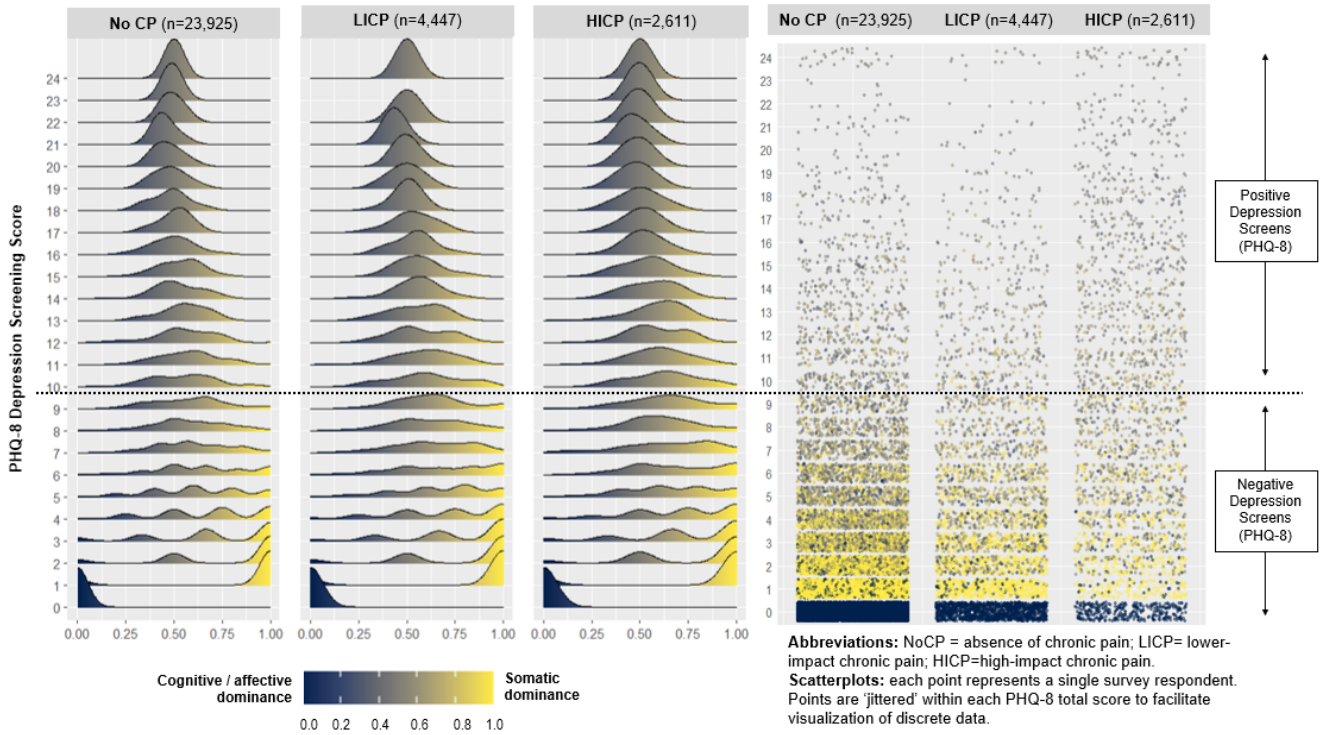
Comparing proportional contribution of somatic symptoms

To facilitate understanding of the contribution of somatic symptoms to PHQ-8 scores, we visualized the distribution of somatic symptoms across NoCP, LICP, and HICP and grouped by PHQ-8 scores **(Figure 2).** *Among those who screened negative,* dominance of either somatic symptoms or cognitive/affective symptoms was frequently observed, however, among those who screened positive, observations reflecting balanced proportional contribution of somatic and cognitive/affective symptoms are increasingly prevalent. This pattern was uniform across NoCP, LICP, and HICP.

Among U.S. adults who screened positive, the mean proportion of somatic symptoms was 0.55. The mean proportion was 0.53 in NoCP, compared to 0.55 in CP: reflecting

proportions of .56 in LICP and 0.55 in HICP. The difference in somatic symptom proportions (0.019) was statistically significant ($p < .008$), given the large sample size, but the magnitude of the difference is trivial.

Figure 2. Smoothed density and scatter plots showing the proportion of somatic symptoms within PHQ-8 scores, grouped by mutually exclusive and exhaustive subgroups to facilitate representative visualization of the U.S. population.



3.2 Confirmatory Factor Analysis (CFA)

The test of measurement invariance determines whether the CFA factor loading matrix is invariant across groups, i.e. whether the PHQ-8 items relate to the single latent trait of depression in the same way. Inspection of model fit indices showed that measurement invariance was supported in CP **(Table 1**), and HICP (**Supplemental Table 1**) at the configural (the overall factor structure), metric (factor loading), and scalar (item intercept) levels, suggesting that the PHQ-8 measures depression and functions in a similar way among U.S.

adults without CP, with CP, and with HICP. The PHQ-8 shows excellent internal consistency and reliability; cross-group equivalence and measurement invariance are observed (**Table 1**).

Table 1. Internal structure, Reliability, Cross-group equivalence and measurement invariance

Internal Structure						**Internal Consistency Reliability**	
	X^2	CFI	TLI	RMSEA (90% CI)	SRMR	Cronbach's alpha coefficient	McDonald's Omega coefficient
No CP n=23,935	885.710	0.994	0.991	0.043 (0.040, 0.045)	0.041	0.82	0.86
CP n=7,083	560.092	0.993	0.991	0.062 (0.057, 0.066)	0.043	0.85	0.89
HICP n= 2,611	309.754	0.991	0.987	0.075 (0.067, 0.082)	0.050	0.85	0.89
Cross-Group Equivalence and Measurement Invariance: Comparing CP vs No CP							
	X^2	CFI	TLI	RMSEA (90% CI)	SRMR		
Configural	608.119	0.985	0.979	0.030 (0.028, 0.032)	0.043		
Metric	799.054	0.981	0.977	0.032 (0.030, 0.034)	0.046		
Scalar	1122.663	0.972	0.971	0.036 (0.034, 0.038)	0.049		
Residual	2866.285	0.927	0.934	0.054 (0.052, 0.056)	0.090		

Fit statistics: $\chi 2$: chi-square test; CFI: comparative fit index; TLI: Tucker–Lewis index; RMSEA: Root Mean Square Error of approximation; SRMR: Standardized Root Mean Square Residual; $\chi 2$ with 40 degrees of freedom and $p < 0.001$ for configural invariance model; $\chi 2$ with 47 degrees of freedom and $p < 0.001$ for metric invariance model; $\chi 2$ with 54 degrees of freedom and $p < 0.001$ for scalar invariance model; $\chi 2$ with 62 degrees of freedom and $p < 0.001$ for residual invariance model.

3.3 CFA Standard Factor Loadings and IRT Discrimination Parameters

CFA and IRT models are conceptually equivalent, yet the interpretation of standard factor loadings and discrimination parameters is complementary;[25] therefore, we report both **(Table 2**). Higher standard factor loadings indicate a greater shared variance between the item and the underlying factor. All standardized factor loadings (λ) exceeded 0.670 across groups, indicating "very good" to "excellent" item performance with substantial shared variance between each PHQ-8 item and the underlying depression factor. The NoCP group demonstrated the strongest factor structure (λ: 0.710-0.875), with seven of eight items achieving excellent performance (λ ≥

0.71). The CP group showed comparable results (λ: 0.677-0.876) with six items reaching excellent performance levels, while the HICP group exhibited the most variability (λ: 0.660-0.884) with five items achieving excellent performance. Sadness demonstrated the highest factor loading (λ: 0.875-0.884) consistently across groups; sadness was the strongest indicator of depression regardless of CP status.

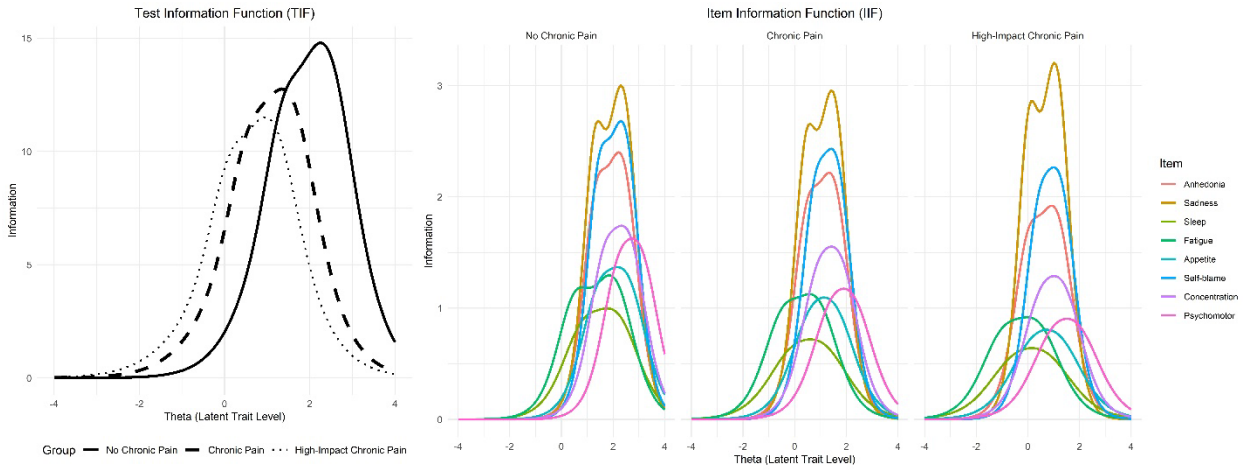
IRT discrimination parameters reflect the ability of each item to discriminate among people with differing *severity* of depression. All items demonstrated “good” to “quite satisfactory” discrimination across groups, with parameters (α) ranging from 1.44 to 3.28. Most items achieved quite satisfactory performance (α ≥ 1.70), across groups, indicating highly effective differentiation of severity between individuals. Sleep achieved relatively lower, though still “good” discrimination ability in CP and HICP, while appetite, fatigue, and psychomotor functioning achieved "good" discrimination in HICP. This suggests that these somatic symptoms may be relatively less discriminant to identify different levels of depression among those with HICP; it is useful to note that this does not suggest problems with comparing groups on the same depression trait dimension.[25]

Table 2. Evaluating Confirmatory Factor Analysis (CFA) standardized item loadings and Item Response Theory (IRT) discrimination parameters

CFA Standardized Item Loadings									
		anhedonia	sadness	sleep	fatigue	appetite	self-blame	concentration	psychomotor
	n	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
No CP	23,925	0.841	0.875	0.71	0.755	0.744	0.839	0.783	0.759
CP	7,063	0.833	0.876	0.755	0.677	0.731	0.834	0.777	0.727
HICP	2,611	0.812	0.884	0.665	0.734	0.685	0.823	0.754	0.690
IRT Discrimination Parameters									
		anhedonia	sadness	sleep	fatigue	appetite	self-blame	concentration	psychomotor
	n	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
No CP	23,925	2.82	3.17	1.81	2.09	2.11	2.98	2.34	2.29
CP	7,063	2.71	3.14	1.53	1.93	1.88	2.83	2.24	1.95
HICP	2,611	2.52	3.28	1.44	1.74	1.61	2.72	2.04	1.71

TIFs show that the PHQ-8 had optimal measurement precision across different severity levels; IIFs are consistent across groups, sadness (PHQ-2) consistently demonstrated the highest ability to detect differing levels of severity **(Figure 3).**

Figure 3. Test information functions (TIFs) and item information functions (IIFs) comparing No Chronic Pain, Chronic Pain, and High-Impact Chronic Pain.



4. DISCUSSION

The PHQ-8 reliably and equivalently measures depressive symptoms in U.S. adults with chronic pain (CP), confirming that the strikingly higher prevalence of depressive symptoms in this population is reflective of true disparities in the prevalence of depression rather than measurement artifacts. In this nationally representative study of 30,983 U.S. adults, CFA, measurement invariance, and IRT analyses demonstrated that PHQ-8 scores are equivalent in those with and without CP. Factor loadings and item discrimination coefficients were uniform and no differential item functioning was observed. Sadness had the highest factor loadings and was the most discriminant across groups, corroborating its importance in differentiating depressive symptoms and their severity regardless of CP status.

This study did not find evidence to suggest somatic (nor any other) symptoms may exert undue influence on the screening outcomes of those with CP compared to others. Consistent with prior studies, those with CP were observed to have elevated severity of several somatic items including fatigue, sleep disturbance, and appetite when compared to other PHQ-8 items. However, comparison of focal and non-focal groups revealed that this finding is not unique to those with CP and that the same cross-item pattern is also readily observed among those without CP.

We evaluated whether CP-associated differences existed in the proportional contribution of somatic symptoms holding PHQ-scores constant **(Figure 2):** Cross-group differences between those with and without CP were not observed. Proportional dominance of somatic symptoms was frequently observed in those who screened negative, while balanced contributions of somatic and cognitive/affective symptoms were more typical among those who screened positive. The proportion of respondents demonstrating dominance of somatic symptoms decreased linearly as PHQ-8 score increased.

Second, CP was associated with increased severity of all individual items and PHQ-8 scores. However, we suggest cautious interpretation of this difference: CP-associated differences in severity are disproportionately driven by those who screened negative. Among those who screened positive, cross-group differences in severity of individual items and PHQ-8 scores were attenuated or eliminated, although significant differences for the majority of items were preserved among those who screened positive, in the context of HICP.

Third, as expected, CP was associated with startling disparities in the prevalence of positive PHQ-8 screens for clinically significant depressive symptoms. Descriptive analyses did not suggest any qualification or nuance related to this finding. The prevalence of positive screens was 3.5% in those without CP, 20.1% in those with CP, and 34.8% in those with high-impact chronic pain (HICP), representing 5- to 10-fold increases in unmet needs vis-a-vis depressive symptoms.

These findings persuasively suggest that disparities in positive screens reflect true differences in the prevalence of clinically significant depressive symptoms and that differences do not result from differential functioning of PHQ-8. Findings should enhance confidence among clinicians and researchers that PHQ-8 is appropriate to screen for depressive symptoms in the clinical context of CP and that the interpretation of screening results is straightforward. Findings underscore the verifiable existence of profound mental health disparities in the patient journeys of U.S. adults with CP, including increased prevalence of depression, increased severity of depressive symptoms, reduced likelihood that depressive symptoms will be identified, and reduced likelihood of receiving needed mental health support.

This study has limitations. First, we acknowledge that some individuals are not reached by NHIS, including U.S. persons with no fixed address, persons living on military bases, and persons living in long-term care facilities or correctional facilities. Second, though the PHQ-8 items administered by NHIS are identical to those administered in clinical practice, it is still possible that screening data collected by NHIS could be less accurate than screening data collected in the context of a clinical encounter.

Contributors

All authors had full access to all study data and take responsibility for the integrity of the data and the accuracy of the data analysis. JD, GC, and RA conceived and designed the study. JD, GC, and KH directly accessed, verified, and analyzed the study data reported in the manuscript. JD, GC, KH, CM, and RA interpreted the data, drafted the manuscript, and critically revised the manuscript for important intellectual content. JD supervised the project.

Data sharing statement

All data from the National Health Interview Survey are publicly available on the website of the National Center for Health Statistics. Code used in this study is available upon reasonable request to the corresponding author.

Declaration of interests

All authors declare that they have no conflict of interest.

Acknowledgements

This study was supported in part by the Comprehensive Center for Pain and Addiction, University of Arizona Health Sciences, Tucson, AZ, USA. The funder had 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; nor in the decision to submit the manuscript for publication.

REFERENCES

1. De La Rosa JS, Brady BR, Ibrahim MM, Herder KE, Wallace JS, Padilla AR, et al. Co-occurrence of chronic pain and anxiety/depression symptoms in U.S. adults: prevalence, functional impacts, and opportunities. Pain. 2024 Mar 1;165(3):666–73.

2. Aaron RV, Ravyts SG, Carnahan ND, Bhattiprolu K, Harte N, McCaulley CC, et al. Prevalence of Depression and Anxiety Among Adults With Chronic Pain: A Systematic Review and Meta-Analysis. JAMA Netw Open [Internet]. 2025 Mar 7 [cited 2025 Aug 7];8(3):e250268. Available from: https://doi.org/10.1001/jamanetworkopen.2025.0268

3. Darnall B. To treat pain, study people in all their complexity. Nature [Internet]. 2018 May 1 [cited 2024 Mar 26];557(7706):7–8. Available from: https://go.gale.com/ps/i.do?p=HRCA&sw=w&issn=00280836&v=2.1&it=r&id=GALE%7CA572639169&sid=googleScholar&linkaccess=abs

4. De La Rosa JS, Brady BR, Herder KE, Wallace JS, Ibrahim MM, Allen AM, et al. The unmet mental health needs of U.S. adults living with chronic pain. PAIN [Internet]. 2024 Dec [cited 2025 Aug 7];165(12):2877–87. Available from: https://journals.lww.com/pain/fulltext/2024/12000/the_unmet_mental_health_needs_of_u_s__adults.22.aspx

5. Darnall BD, Scheman J, Davin S, Burns JW, Murphy JL, Wilson AC, et al. Pain Psychology: A Global Needs Assessment and National Call to Action. Pain Med [Internet]. 2016 Feb 1 [cited 2024 Mar 26];17(2):250–63. Available from: https://doi.org/10.1093/pm/pnv095

6. Brennan F, Carr DB, Cousins M. Pain Management: A Fundamental Human Right. Anesth Analg [Internet]. 2007 Jul [cited 2025 Aug 8];105(1):205. Available from:

https://journals.lww.com/anesthesia-analgesia/fulltext/2007/07000/pain_management__a_fundamental_human_right.37.aspx

7. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. Chronic pain as a symptom or a disease: the IASP Classification of Chronic Pain for the International Classification of Diseases (ICD-11). Pain. 2019 Jan;160(1):19–27.

8. Tenti M, Raffaeli W, Gremigni P. A Narrative Review of the Assessment of Depression in Chronic Pain. Pain Manag Nurs [Internet]. 2022 Apr 1 [cited 2025 Aug 7];23(2):158–67. Available from: https://www.sciencedirect.com/science/article/pii/S1524904221001223

9. Kroenke K. PHQ-9: global uptake of a depression scale. World Psychiatry [Internet]. 2021 [cited 2025 Aug 8];20(1):135–6. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.20821

10. Torre JA de la, Vilagut G, Ronaldson A, Valderas JM, Bakolis I, Dregan A, et al. Reliability and cross-country equivalence of the 8-item version of the Patient Health Questionnaire (PHQ-8) for the assessment of depression: results from 27 countries in Europe. Lancet Reg Health – Eur [Internet]. 2023 Aug 1 [cited 2025 Jan 17];31. Available from: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(23)00078-9/fulltext

11. Hu L tze, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychol Methods. 1998;3(4):424–53.

12. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ [Internet]. 2011 Jun 27 [cited 2025 Aug 8];2:53–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4205511/

13. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98–104.

14. Bichi AA, Talib R. Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. Int J Eval Res Educ IJERE [Internet]. 2018 Jun 1 [cited 2025 Aug 8];7(2):142–51. Available from: https://ijere.iaescore.com/index.php/IJERE/article/view/12900

15. A.Pett M, R.Lackey N, J.Sullivan J. Making Sense of Factor Analysis [Internet]. SAGE Publications, Inc.; 2003 [cited 2025 Aug 8]. Available from: https://methods.sagepub.com/book/mono/making-sense-of-factor-analysis/toc

16. National Center for Health Statistics. Survey Description, National Health Interview Survey, 2019. Hyattsville, MD: National Center for Health Statistics; 2020.

17. Pathak Y, Makk-Frid E. Population Estimates of Self-Reported Depression and Anxiety in the US From a National Survey: Cross-Sectional Survey Study. Interact J Med Res [Internet]. 2025 Apr 16 [cited 2025 Aug 8];14(1):e70626. Available from: https://www.i-jmr.org/2025/1/e70626

18. Duca LM, Helmick CG, Barbour KE, Nahin RL, Von Korff M, Murphy LB, et al. A Review of Potential National Chronic Pain Surveillance Systems in the United States. J Pain. 2022 Sep;23(9):1492–509.

19. Lucas JW, Sohi I. Chronic Pain and High-impact Chronic Pain in U.S. Adults, 2023. NCHS Data Brief [Internet]. 2024 Oct [cited 2025 Jun 9];(518):CS355235. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11726267/

20. 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 [Internet]. 2009 Apr 1 [cited 2023 Feb 25];114(1):163–73. Available from: https://www.sciencedirect.com/science/article/pii/S0165032708002826

21. Villarroel MA. Symptoms of Depression Among Adults: United States, 2019. 2020.

22. Duivis HE, Vogelzangs N, Kupper N, de Jonge P, Penninx BWJH. Differential association of somatic and cognitive symptoms of depression and anxiety with inflammation: Findings from the Netherlands Study of Depression and Anxiety (NESDA). Psychoneuroendocrinology [Internet]. 2013 Sep 1 [cited 2025 Aug 8];38(9):1573–85. Available from: https://www.sciencedirect.com/science/article/pii/S0306453013000073

23. Ehrmann D, Krause-Steinrauf H, Uschner D, Wen H, Hoogendoorn CJ, Crespo-Ramos G, et al. Differential associations of somatic and cognitive-affective symptoms of depression with inflammation and insulin resistance: cross-sectional and longitudinal results from the Emotional Distress Sub-Study of the GRADE study. Diabetologia [Internet]. 2025 Jul 1 [cited 2025 Aug 8];68(7):1403–15. Available from: https://doi.org/10.1007/s00125-025-06369-8

24. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol [Internet]. 2021 Jun [cited 2025 Aug 8];134:79–88. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8168830/

25. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychol Bull. 1993;114(3):552–66.