frontiers | Frontiers in Psychiatry

TYPE Brief Research Report
PUBLISHED 12 December 2024
DOI 10.3389/fpsyt.2024.1483782

OPEN ACCESS

EDITED BY
Edwin de Beurs,
Leiden University, Netherlands

REVIEWED BY
Sami Hamdan,
Academic College Tel Aviv-Jaffa, Israel
Felix Fischer,
Charité University Medicine Berlin, Germany

*CORRESPONDENCE
Sören Kliem
soeren.kliem@eah-jena.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 20 August 2024
ACCEPTED 01 November 2024
PUBLISHED 12 December 2024

CITATION
Kliem S, Sachser C, Lohmann A, Baier D, Brähler E, Gündel H and Fegert JM (2024) Psychometric evaluation and community norms of the PHQ-9, based on a representative German sample.
Front. Psychiatry 15:1483782.
doi: 10.3389/fpsyt.2024.1483782

Psychometric evaluation and community norms of the PHQ-9, based on a representative German sample

Sören Kliem[1*†], Cedric Sachser[2†], Anna Lohmann[1†], Dirk Baier[3], Elmar Brähler[4,5], Harald Gündel[5] and Jörg M. Fegert[2]

[1]Department of Social Welfare, Ernst-Abbe-Hochschule Jena - University of Applied Sciences, Jena, Germany, [2]Department for Child and Adolescent Psychiatry/Psychotherapy, University Clinic for Psychosomatic Medicine and Psychotherapy Ulm, Ulm, Germany, [3]Institute of Delinquency and Crime Prevention, Zurich University of Applied Sciences, Zurich, Switzerland, [4]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany, [5]Department of Psychosomatic Medicine and Psychotherapy, Ulm University Medical Center, Ulm, Germany

Background: The Patient Health Questionnaire (PHQ-9) is a popular tool for assessing depressive symptoms in both general and clinical populations. The present study used a large representative sample of the German adult population to confirm desired psychometric functioning and to provide updated population norms.

Methods: The following psychometric properties were assessed: (i) Item characteristics (item means, standard deviations and inter-item correlations), (ii) Construct validity (correlations of the PHQ-9 sum-score with scores obtained from instruments assessing depression, anxiety and somatization (GAD-7, BSI-18), (iii) Internal consistency (coefficient omega), (iv) Factorial validity (via confirmatory factor analysis of the assumed one factorial model) as well as (v) Measurement invariance (via multi-group confirmatory factor analyses across gender, age, income and education).

Results: The study found that the PHQ-9 had sound psychometric properties in terms of internal consistency and construct validity, and that measurements obtained with the tool could be compared across gender and age.

Limitations: Despite using a representative sample, the response rate was only 42.6%. Furthermore, diagnostic efficiency cannot be assessed as there were no clinical interviews conducted. Conclusion: The updated population based norms, which are presented for the total sample as well as separated by gender and various age-groups, provide a useful reference for clinical practice and epidemiological research.

KEYWORDS

PHQ-9, major depression, self-report questionnaire, population norms, psychometrics, measurement invariance

1 Introduction

Major depression is a common mood disorder that requires brief and comprehensive screening instruments for its detection and assessment in both clinical and research settings (1). The 9-question Patient Health Questionnaire (PHQ-9) scale is such a tool which corresponds to DSM-IV major depressive criteria and was developed as a self-administered questionnaire for use in primary care setting (2). It is widely established for detecting the presence and severity of depression. Various cut-off scores have been suggested that optimize sensitivity, specificity as well as positive and negative predictive values obtained via diagnostic clinical interviews (2–4). A recent individual participant data meta-analysis (5) extending the work of (3), encompassing 100 studies and 44,503 participants, evaluated the accuracy of the PHQ-9 against various diagnostic methods, including semistructured and fully structured interviews, as well as the Mini International Neuropsychiatric Interview (MINI). The analysis revealed that the commonly used cut-off score of ≥10 optimized both sensitivity (0.85, 95% CI 0.79-0.89) and specificity (0.85, 95% CI 0.82-0.87) when compared with semistructured diagnostic interviews.

Studies like these demonstrate that the PHQ-9 is a brief and effective tool for depression screening While it may have limitations compared to more detailed measures like the CIDI (6) or BDI-II (7), particularly in specialized clinical settings. Nevertheless, its ease of use and its alignment with diagnostic criteria of major depressive disorder makes it ideal for routine screening in general practice.

The fact that it is well validated and freely available in many languages also makes it a popular tool in epidemiological studies of mental health and psychological distress [e.g., (1, 8–10)]. It has proven useful across socioeconomic backgrounds (11) and cultures. More recently, the PHQ-9 has also been used extensively assessing the mental health burden related to the COVID-19 pandemic [e.g., (12–15)].

Normative values of a questionnaire are crucial for assessing the level of distress in individuals and groups of patients. While the PHQ-9 has been widely used in various populations and settings, the research literature on normative scores obtained from representative general population samples is very limited [e.g., (16–18)]. The only large normative study conducted with adults from the general population in Germany is based on data from 2003-2008 (19).

As the PHQ-9 uses a sum-score for the interpretation of symptom severity, it is paramount to confirm the factor structure to ensure that this form of aggregation is appropriate (20). The PHQ-9 factor structure has been frequently debated in the literature, with the majority of studies suggesting a one factorial structure and the occasional mention of two or more highly correlated factors (21).

Abbreviations: PHQ-9, Patient Health Questionnaire; GAD-7, Generalized Anxiety scale; BSI GSI, Brief Symptom Inventory Global Severity Index; BSI Somatization, Brief Symptom Inventory Somatization Subscale; BSI Anxiety, Brief Symptom Inventory Anxiety Subscale; BSI Depression, Brief Symptom Inventory Depression Subscale; CFA, confirmatory factor analysis; CI, Confidence Interval; WLSMV, Weighted least square means and variance adjusted estimation; MI, measurement invariance; MGCFA, multiple group factor analysis.

With depression levels varying across several demographic groups, it is furthermore important to confirm measurement invariance in order to access whether findings are comparable across various sub-populations (11, 22). Especially gender differences have been identified as substantial by meta-analyses (23).

Therefore, the aim of this study was to assess the psychometric properties of the PHQ-9 in a large representative sample of the general population and provide updated German population norms.

Moreover, due to the widespread use of the PHQ-9, it is important to capture any potential shifts in item behavior in the general population. The current data obtained from a large representative community sample provides a valuable reference distribution for more meaningful interpretations of data obtained from other populations and settings.

2 Methods

2.1 Procedure

The PHQ-9 was presented as part of a large survey conducted by Leipzig University between December 2020 and March 2021.

The goals of the survey were (a) to assess prevalence rates of a variety of relevant physical or mental disorders and related risk behaviors (descriptive epidemiology), (b) to examine causes and conditions of these disorders (analytic epidemiology), and (c) to analyze psychometric properties and provide German population norms for clinical-psychological instruments. The survey was carried out by the contractor USUMA Markt- und Sozialforschung an independent institute for opinion and social research.

It consisted of two parts. The first part was guided by a trained interviewer and collected extensive demographic as well as household information. Survey contents in this part were based on principles of the German Statistisches Bundesamt (Federal Statistical Office).

The second part consisted of paper-based self-administered questionnaires which the participants filled in independently. Interviewers remained out of view but available for questions. Prior to their participation in the survey all participants obtained a written copy of the confidentiality agreement providing details regarding the handling of their personal data. The study followed the Declaration of Helsinki. Minimum age for participation was 16 years. All participants provided informed consent prior to the interview. For under-aged participants at least one legal guardian was informed about the sampling procedure and the survey contents. All procedures were approved by the Ethics Committee of the Medical Faculty of the University of Leipzig (Az.: 474/20-ek).

2.2 Sample description

As Germany does not keep a central population registry, representativeness of the sample was ensured by using the ADM sampling system F2F. This sampling procedure consists of three steps. In a first step, the area of the Federal Republic of Germany is divided into regions of which 258 are sampled with sampling

probability proportional to the number of households. In a second step, 5676 households are selected based on a random route procedure. Finally, the target person within each household is identified using a Kish selection grid (24). Further details regarding the sampling procedure, COVID measures and sample representativeness can be found in the Supplementary Materials (section A). Details regarding response can be obtained from Figure 1. The following analyses are based on data from $N = 2519$ participants which corresponds to a response rate of 42%. Figure 1 presents a flowchart outlining the sampling procedure and reasons for non-response. Table 1 provides sample descriptives.

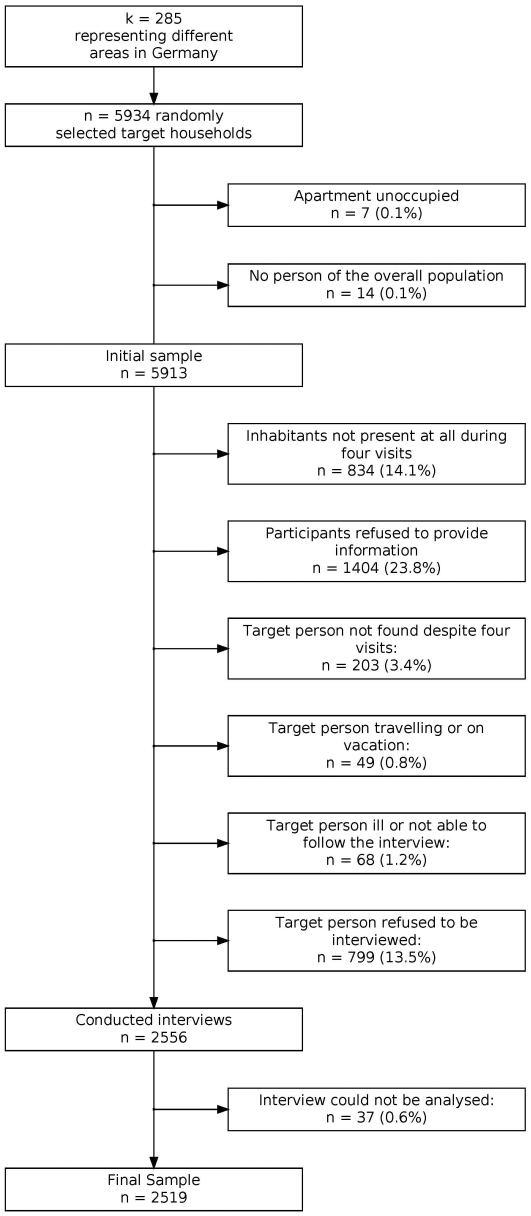

FIGURE 1
Flowchart of sampling procedure and reasons for nonparticipation.

2.3 Instruments

As the survey served multiple epidemiological purposes, only those measures that were used in the validation process are discussed in this paper. In addition to extensive demographic information (see Table 1), health related behavior, such as the number of sick days, doctor visits, and hospital stays, were assessed. The following measures were used for the validation of the scale at hand.

2.3.1 Patient Health Questionnaire (PHQ-9)

The PHQ-9 (2) is a self-report scale, that scores depression symptoms using nine items. Participants indicate symptom frequency over the last two weeks on a four-point Likert scale from 0 (not at all) to 3 (almost every day), providing a total severity score ranging from 0 to 27. In the present study, the German version of the PHQ-9 (25) was used. The PHQ-9 showed high internal consistency in previous general population studies [$\alpha = 0.87$ (19)].

2.3.2 The General Anxiety Disorder Scale (GAD-7)

The GAD-7 (26) is a brief self-report scale with seven items assessing generalized anxiety. Each of the seven items is rated on a scale from 0 (not at all) to 3 (almost every day). The total score of the GAD-7 ranges from 0-21. The GAD-7 showed high internal consistency in previous general population studies ($\alpha = 0.89$ (10); study at hand: study at hand: $\alpha = 0.90$, 95% CI [0.89 - 0.91]; $\omega = 0.92$, 95% CI [0.91 - 0.92]). For a psychometric evaluation of the GAD-7 based on the current data see (Kliem et al., 2024)[1].

2.3.3 The Brief Symptom Inventory (BSI-18)

The BSI-18 (27) is an 18-item short form of the Symptom-Checklist 90-R. It contains three subscales each of which comprises six items: somatization (SOMA), depression (DEPR) and anxiety (ANX). The BSI's sum score of all 18 items can be interpreted as a Global Severity Index (GSI). The BSI-18 has shown high internal consistency in previous studies of the general population ($\alpha = 0.93$ [GSI], 0.82 [SOMA], 0.87 [DEPR], 0.84 [ANX] (28); study at hand: $\alpha = 0.93$, 95%CI [0.92-0.94]; $\omega = 0.94$, 95% CI [0.93- 0.94]).

2.4 Statistical analysis

2.4.1 Missing data

Proportion of missing data on the PHQ-9 items ranged from 0.30% to 0.70%. In order to address missing data, we utilized chained equation modeling as outlined in van Buuren and Groothuis-Oudshoorn (29). The imputation algorithm used the following variables: gender, age, nationality, marital status, living with a partner, educational and income as well as all items from the scales PHQ-9, GAD-7 and BSI-18 to estimate missing data. We corrected for implausible item values by employing predictive mean matching, whereby the closest observable values to the predicted values ($\hat{y}$) were selected. Imputation procedures were implemented using the R package mice (29). Data analysis was carried out on one

1 Kliem S, Sachser C, Lohmann A, Baier D, Brähler E, Fegert J, et al. Psychometric evaluation and community norms of the GAD-7, based on a representative German sample. (2024).

10.3389/fpsyt.2024.1483782

TABLE 1 Demographic characteristics of the study sample.

	Male	Female	Diverse	Total
	(N=1193)	(N=1322)	(N=4)	(N=2519)
Age (years)				
Mean (SD)	50.1 (17.7)	50.5 (18.3)	44.8 (26.5)	50.3 (18.1)
Median [Min, Max]	52.0 [16.0, 96.0]	51.0 [16.0, 96.0]	41.5 [21.0, 75.0]	51.0 [16.0, 96.0]
Age Categories				
16-24	102 (8.5%)	125 (9.5%)	2 (50.0%)	229 (9.1%)
25-34	190 (15.9%)	174 (13.2%)	0 (0%)	364 (14.5%)
35-44	174 (14.6%)	225 (17.0%)	0 (0%)	399 (15.8%)
45-54	195 (16.3%)	216 (16.3%)	0 (0%)	411 (16.3%)
55-64	244 (20.5%)	243 (18.4%)	1 (25.0%)	488 (19.4%)
65-74	190 (15.9%)	200 (15.1%)	0 (0%)	390 (15.5%)
75+	98 (8.2%)	139 (10.5%)	1 (25.0%)	238 (9.4%)
Nationality				
German	1151 (96.5%)	1271 (96.1%)	3 (75.0%)	2425 (96.3%)
not German	42 (3.5%)	48 (3.6%)	1 (25.0%)	91 (3.6%)
Missing	0 (0%)	3 (0.2%)	0 (0%)	3 (0.1%)
Marital Status				
married/living together	547 (45.9%)	527 (39.9%)	2 (50.0%)	1076 (42.7%)
married/separated	40 (3.4%)	25 (1.9%)	0 (0%)	65 (2.6%)
single	398 (33.4%)	357 (27.0%)	2 (50.0%)	757 (30.1%)
divorced	143 (12.0%)	227 (17.2%)	0 (0%)	370 (14.7%)
widowed	62 (5.2%)	181 (13.7%)	0 (0%)	243 (9.6%)
Missing Living with partner	3 (0.3%)	5 (0.4%)	0 (0%)	8 (0.3%)
living with partner	729 (61.1%)	737 (55.7%)	2 (50.0%)	1468 (58.3%)
not living with partner	444 (37.2%)	565 (42.7%)	2 (50.0%)	1011 (40.1%)
Missing Educational Attainment	20 (1.7%)	20 (1.5%)	0 (0%)	40 (1.6%)
No University Entry Qualification	921 (77.2%)	1025 (77.5%)	3 (75.0%)	1949 (77.4%)
University Entrancy Qualification	262 (22.0%)	288 (21.8%)	1 (25.0%)	551 (21.9%)
Missing Monthly per Capita Houshold Income (€)	10 (0.8%)	9 (0.7%)	0 (0%)	19 (0.8%)
Mean (SD)	2050 (997)	1900 (917)	2880 (2100)	1970 (961)
Median [Min, Max]	1750 [125, 7500]	1730 [144, 5300]	1730 [1590, 5300]	1750 [125, 7500]
Missing	22 (1.8%)	41 (3.1%)	1 (25.0%)	64 (2.5%)

imputed data set. More details on item wise missingness can be found in the Supplementary Material. As a sensitivity analysis all major analyses were additionally run on the unimputed data.

2.4.2 Item characteristics

We calculated mean and standard deviations for all items of the PHQ-9 in the total sample and in sub-samples of male and female participants. Cohen's *d* was used to quantify effect sizes for group differences in item means. We also calculated inter-item correlations.

2.4.3 Construct validity

To evaluate construct validity of the PHQ-9, we correlated the scale with the GAD-7 and the three BSI-18 subscales (somatization, anxiety and depression) as well as with the BSI global severity index.

The following hypotheses were formulated: depression levels should be higher in individuals with (a) higher anxiety scores, and (b) higher somatization scores [e.g., Gierk et al. (30); Kliem et al. (31)].

2.4.4 Internal consistency

To account for potential issues arising from unmet assumptions in the calculation of coefficient α (32), we assessed the internal consistency of the PHQ-9 using McDonald's ω, which was computed using the semTools R package (33). This additional measure provides a more robust evaluation of internal consistency.

2.4.5 Factorial validity and measurement invariance

To confirm the one-dimensional structure of the PHQ-9, we conducted confirmatory factor analyses (CFA) using the lavaan package in R statistics (34). Weighted least square means and variance adjusted estimation (WLSMV) were used, as recommended for ordered categorical response options. We also tested measurement invariance (MI) using multiple group factor analysis (MGCFA) following the procedure suggested by Wu and Estabrook (35). We used theta parameterization and identified the model by setting means and variances of latent factors to 0 and 1, respectively, item intercepts to 0, and residual variances to 1. We subsequently tested five models: (i) configural invariance (no constraints apart from those necessary for model identification), (ii) threshold invariance (constraining all thresholds to be equal), (iii) weak invariance (constraint of loadings), (iv) strong invariance (constraining of intercepts), and (v) full invariance (constraining residual variances). Supplementary Figure C1 (in the Supplementary Materials) provides an overview of the structural equation models assessed. Supplementary Table C1 (Supplementary Materials) provides a detailed overview of parameter constraints for each steps of the MGCFA. Chen's (36) cut-off criteria were used, with a change of < -.01 in CFI and a change of ≥.015 in RMSEA indicating non-invariance. As Sass et al. (37) have pointed out, the cut-offs suggested by Chen are often too liberal when using WLSMV estimation. We have hence added sensitivity analyses using MLR estimation. We conducted MGCFA for the PHQ-9 across gender, age (below median age vs. above median age), age * gender, income (below median vs. above median) and educational attainment (no university entrance diploma vs. university entrance diploma). Cases classifying as neither male nor female were not included in the MGCFA for gender and age * gender due to their low number. Due to empty cells in the MGCFA of age * gender in item 9 the two highest answer categories were collapsed for this one item. The semTools package (33) for R statistics was used to conduct MI analyses.

3 Results

3.1 Item characteristics

Supplementary Table C5 (in the Supplementary Materials) displays means and standard deviations for the nine items of the PHQ-9 in the total sample as well as effect sizes for mean differences regarding gender. On the item-level there was a consistent pattern of female participants exhibiting higher mean depression scores as well as higher variability on most PHQ-9 items. Effect sizes (Cohen's d) of these mean differences were very small and ranged from d = -0.05 95% CI [-0.13,0.03] to d = 0 95% CI [-0.08,0.08].

3.2 Construct validity

To determine evidence of construct validity of the PHQ-9, correlation coefficients were calculated with related instruments. In line with our hypotheses, there were high positive correlations between the PHQ-9 and measures of somatization, anxiety and depression as assessed by BSI-18 subscales (see Supplementary Table C7 in the Supplementary Materials). In the same vein, the GAD-7 assessing anxiety showed positive correlations with the PHQ-9.

3.3 Population norms

Table 2 shows cumulative percentiles of PHQ-9 scores for the total sample. Additional norms split by gender as well as age group can be found in the Supplementary Material (see Supplementary Tables C8, C9). Table 3 reports absolute and relative frequencies per severity category. We neither endorse nor have verified this classification but provide it as a mere descriptive to facilitate comparing results across studies.

3.4 Internal consistency

Cronbach's alpha of the PHQ-9 for the full sample was α = 0.90, 95% CI [0.89, 0.91]. McDonald's omega of the PHQ-9 for the full sample was ω = 0.93, 95% CI [0.92, 0.94].

3.5 Factorial validity

A CFA was conducted to assess the unidimensional structure of the PHQ-9. The fit indices indicated reasonable model fit, with a robust CFI of 0.91, a robust TLI of 0.89, and an SRMR of 0.044. However, the robust RMSEA was 0.17 (90% CI [0.153, 0.186]), suggesting some misfit.

To improve the model, we inspected modification indices and introduced residual correlations between items with overlapping content: #1 (Little interest or pleasure) with #2 (Down, depressed, hopeless) and #3 (Sleep problems) with #4 (Tired, little energy). These adjustments improved the fit indices (CFI = 0.96, RMSEA = 0.11), as detailed in Supplementary Table C2.

Despite these modifications, factor score correlations between the original and adjusted models remained high (r > 0.999), confirming the stability of the latent structure and supporting the unidimensionality of the PHQ-9. Strong standardized factor loadings

10.3389/fpsyt.2024.1483782

TABLE 2 Population based norms (cumulative percentiles) of the PHQ-9 scores (total sample).

PHQ-9	Total	Age 16-24	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age 65-74	Age 75+
0	41.0	41.5	50.0	44.9	39.2	39.8	38.2	30.3
1	52.3	50.7	61.0	60.4	52.6	49.6	46.7	41.2
2	64.3	60.3	72.5	73.2	63.0	61.1	60.5	55.9
3	72.9	69.9	77.7	79.7	72.3	70.1	71.8	66.0
4	79.2	74.7	83.5	83.7	78.8	76.2	79.7	75.6
5	83.8	81.2	87.1	85.2	83.9	81.4	84.9	82.4
6	87.5	86.0	90.7	88.5	88.6	84.6	87.2	86.6
7	89.4	88.2	92.0	90.0	89.8	88.1	89.0	88.7
8	92.1	91.7	93.7	91.2	92.7	91.2	93.3	90.3
9	93.9	93.4	95.6	93.5	93.7	93.6	95.6	90.8
10	95.0	95.6	97.8	94.0	94.9	94.3	95.9	91.6
11	95.9	95.6	98.6	95.2	95.4	95.3	97.2	93.3
12	96.4	95.6	99.2	95.2	95.9	95.9	97.7	94.5
13	97.2	96.5	99.5	96.0	97.8	96.3	97.7	96.2
14	97.8	96.9	99.5	96.5	98.5	97.3	98.5	96.6
15	98.3	97.8	99.5	96.7	98.8	98.0	99.0	97.9
16	98.8	98.3	99.7	97.7	99.0	98.8	99.5	98.3
17	99.0	98.3	99.7	98.2	99.3	98.8	99.7	98.3
18	99.0	98.3	99.7	98.2	99.3	98.8	99.7	98.7
19	99.3	98.3	> 99.9	98.5	99.5	99.0	99.7	> 99.9
20	99.5	> 99.9	> 99.9	98.5	99.5	99.2	99.7	> 99.9
21	99.6	> 99.9	> 99.9	99.2	99.8	99.2	99.7	> 99.9
22	99.7	> 99.9	> 99.9	99.5	> 99.9	99.2	99.7	> 99.9
23	99.8	> 99.9	> 99.9	99.7	> 99.9	99.2	99.7	> 99.9
25	99.8	> 99.9	> 99.9	99.7	> 99.9	99.4	> 99.9	> 99.9
26	99.9	> 99.9	> 99.9	99.7	> 99.9	99.6	> 99.9	> 99.9
27	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9

(0.79–0.89) further reinforced this. A SEM path diagram can be found in Supplementary Figure C2 of the Supplementary Materials.

3.6 Measurement invariance

The fit measures obtained in the measurement invariance analyses of the PHQ-9 are presented in Supplementary Table C4 in the Supplementary Materials. Adequate CFI and RMSEA differences were found for all invariance steps and groups. The sensitivity analyses using MLR estimation confirmed measurement invariance regarding age, gender and education. Given the very small CFA and RMSEA differences in the original analysis using WLSMV estimation, measurement invariance regarding age × gender and income is likely, yet inconclusive.

4 Discussion

The present study investigates the psychometric quality of the PHQ-9 using a large and representative sample of the German general population. Based on coefficient ω, the PHQ-9 can be attested a high internal consistency. Furthermore, the analyses showed comparable factor structures using MGCFA in the

10.3389/fpsyt.2024.1483782

TABLE 3 PHQ-9 scores by severity category and gender.

Severity	Total		Men		Women	
	n	%	n	%	n	%
minimal (0-4)	1996	79.24	980	82.15	1014	76.70
mild (5-9)	370	14.69	152	12.74	217	16.41
moderate (10-14)	97	3.85	38	3.19	59	4.46
moderately severe (15-19)	38	1.51	16	1.34	22	1.66
severe (20-27)	18	0.71	7	0.59	10	0.76

subgroups that were compared (gender, age groups). The overall factor structure assumed for the PHQ-9 fitted well for the defined gender and age groups thus indicating that the PHQ-9 can be used for gender or age comparisons. Lastly, the reported correlation between the PHQ-9 and the GAD-7 as well as the BSI-18 lies within the range of previous studies. Overall, the present results are in line with results of previous normative studies (19), suggesting that the PHQ-9 is an efficient, reliable, and valid instrument for assessing depressive symptoms. We provide updated norm tables for clinical practice, which was the main aim of this study. These percentiles are tabulated (see Table 2 and Supplementary Tables C6, C7) for different age ranges and available both gender-specific and gender-unspecific. On the population level, the suggested clinical cut-off of 10 points for the PHQ-9 falls in the range of percentiles 92-98, which reflects above average to well above average values.

The percentiles obtained in the present study are comparable to values reported by previous work in the German general population (19). It is however noteworthy that our norms are almost identical to previous norms "+1" i.e. our percentiles closely align to those of (19) but with a shift of almost 1 point on the PHQ-Sum score. Our values also closely align with more recent percentile-ranks provided by Shin et al. (16) based on general population data from Korea as well as data from Tomitaka et al. (17) from the US general population. (For a more detailed comparison see Supplementary Figure C3 in the online supplements).

The high share of participants indicating virtually no depressive symptoms (41%) might seem counter intuitive given the ongoing COVID-19 pandemic during the survey period. However, a decrease of mood disorder symptoms is in line with other findings from large representative German studies indicating, for example, that contrary to common belief levels of domestic violence (38) decreased and overall mental health increased [e.g., (39)]. There are several large surveys which include the PHQ-9 that were carried out in the German general population [e.g., (40, 41)] as well as in other western general populations [e.g., (12–15, 42)]. While the depressive symptom burden varies considerably among these studies they all report significantly higher levels of depression in the general population than the study at hand. All of these studies utilized large online convenience samples and the surveys were framed as assessing pandemic related mental health burden. We consider these differences in study design to be crucial with respect to self-selection bias and hence believe that our results based on a face-to-face survey using a representative sample is an important contribution, contrasting and potentially balancing the current scientific discourse. Furthermore, these findings highlight the necessity for frequently updated norms and warns against the assumptions of stable prevalence rates and symptom burden. Additionally, our norms serve as an important reference point for longitudinal studies and provide an indication of symptom variability over time on the population level.

4.1 Limitations

Despite the current study being based on high quality data from a large representative sample it is not without limitations. First, the response rate was only 42.6%. However, general population studies commonly have significantly lower response rates than clinical studies and the response rate of this study is comparable to similar surveys [e.g., (31, 43, 44)]. Despite significant efforts to maximize sample representativeness, some degree of non-response remains unavoidable with the current design and there is a potential for bias due to non-response. Unfortunately, the possibility of non-response bias cannot be systematically assessed as no demographic information of non-responders is available. This would only be possible if sampling were based on registry data which is not accessible without government permit in Germany. Furthermore, the current study does not allow any conclusions regarding the diagnostic efficiency of the PHQ-9 as no clinical interviews were conducted. While the presented norms and psychometric properties can serve as valuable reference data for clinical research the generalizability to clinical samples is limited.

4.2 Conclusion

In summary, the German version of the PHQ-9 has shown sound psychometric properties in a large representative population sample. While the percentiles for the sum-score are comparable to similar studies in community samples, the change in symptom burden compared to previous German norm values from over a decade earlier make the present study an important reference. Furthermore, the results from the present study serve as a notable counter point in the interpretation of COVID-19 related findings including the PHQ-9. We suggest updating the norms again in the near future to gain a deeper understanding whether the present findings have to be interpreted as pandemic related or as "the new normal".

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of the Medical Faculty of the University of Leipzig (Az.: 474/20-ek). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SK: Conceptualization, Formal analysis, Methodology, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. CS: Writing – review & editing. AL: Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. DB: Writing – review & editing. EB: Conceptualization, Data curation, Funding acquisition, Project administration, Writing – review & editing. HG: Funding acquisition, Writing – review & editing. JF: Funding acquisition, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research did not receive specific funding, but was performed as part of the employment of the authors. SK&AL Ernst-Abbe Hochschule University of Applied Sciences Jena, CS &JMF University Clinic Ulm, DB Zurich University of Applied Sciences, EB University Mainz. The funders were neither involved in manuscript writing, editing, approval, or decision to publish. Open access funding by Zurich University of Applied Sciences (ZHAW).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1483782/full#supplementary-material

SUPPLEMENTARY FIGURE C1
MGCFA models for Measurement Invariance Analysis.

SUPPLEMENTARY FIGURE C2
One factor CFA model of the PHQ-9.

SUPPLEMENTARY FIGURE C3
One factor CFA model of the PHQ-9 with residual correlation #1 #2.

SUPPLEMENTARY FIGURE C4
Figure C4. One factor CFA model of the PHQ-9 with residual correlation #1 #2 and #3 #4.

SUPPLEMENTARY FIGURE C5
Comparison of PHQ-9 Sum-scores across community studies.

SUPPLEMENTARY TABLE A1
Comparison Sample Distribution vs Population Distribution for Age and Gender.

SUPPLEMENTARY TABLE A2
Comparison Sample Distribution vs Population Distribution for Population of Federal States.

SUPPLEMENTARY TABLE A3
Demographic characteristics of the study sample at hand.

SUPPLEMENTARY TABLE A4
Demographic characteristics of a comparable survey with identical sampling procedure. Item-wise missingness of PHQ-Items. Item-wise Missingness of GAD-Items. Item-wise Missingness of BSI-18 Items.

SUPPLEMENTARY TABLE C1
Parameter Constraints for MGCFA.

SUPPLEMENTARY TABLE C2
Fit indices of One Factor Model and Alternative models with additional residual correlations.

SUPPLEMENTARY TABLE C3
Factor score correlations: PHQ-9 sum score, One Factor Model, Alternative models with additional residual correlations.

SUPPLEMENTARY TABLE C4
Results of measurement invariance analyses.

SUPPLEMENTARY TABLE C5
Means (M), standard deviation (SD), and group differences for the PHQ-9 items.

SUPPLEMENTARY TABLE C6
Means, standard deviations, and correlations with confidence intervals PHQ-9 Items.

SUPPLEMENTARY TABLE C7
Scale correlations: PHQ-9, GAD-7, BSI-18.

SUPPLEMENTARY TABLE C8
Population based norms of the PHQ-9 (male subsample).

SUPPLEMENTARY TABLE C9
Population based norms of the PHQ-9 (female subsample).

SUPPLEMENTARY TABLE D1
Means (M), standard deviation (SD), and group differences for the PHQ-9 items (unimputed data).

SUPPLEMENTARY TABLE D2
Scale correlations: PHQ-9, GAD-7, BSI-18 (unimputed data).

SUPPLEMENTARY TABLE D3
Population based norms of the PHQ-9 (total sample – unimputed data).

SUPPLEMENTARY TABLE E1
Results of measurement invariance analyses (MLR estimator).

References

1. Thase ME. Recommendations for screening for depression in adults. *JAMA*. (2016) 315:349–50. doi: 10.1001/jama.2015.18406

2. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. *J Gen Internal Med*. (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x

3. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. (2019) 365:l1476. doi: 10.1136/bmj.l1476

4. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: A diagnostic meta-analysis of 40 studies. *BJPsych Open*. (2016) 2:127–38. doi: 10.1192/bjpo.bp.115.001685

5. Negeri ZF, Levis B, Sun Y, He C, Krishnan A, Wu Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: Updated systematic review and individual participant data meta-analysis. *BMJ*. (2021) 375:n2183. doi: 10.1136/bmj.n2183

6. Maske UE, Busch MA, Jacobi F, Beesdo-Baum K, Seiffert I, Wittchen HU, et al. Current major depressive syndrome measured with the Patient Health Questionnaire-9 (PHQ-9) and the Composite International Diagnostic Interview (CIDI): Results from a cross-sectional population-based study of adults in Germany. *BMC Psychiatry*. (2015) 15:77. doi: 10.1186/s12888-015-0463-4

7. Titov N, Dear BF, McMillan D, Anderson T, Zou J, Sunderland M. Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cogn Behav Ther*. (2011) 40:126–36. doi: 10.1080/16506073.2010.550059

8. Hinz A, Klein AM, Brähler E, Glaesmer H, Luck T, Riedel-Heller SG, et al. Psychometric evaluation of the Generalized Anxiety Disorder Screener GAD-7, based on a large German general population sample. *J Affect Disord*. (2017) 210:338–44. doi: 10.1016/j.jad.2016.12.012

9. Johansson R, Carlbring P, Heedman Å, Paxling B, Andersson G. Depression, anxiety and their comorbidity in the Swedish general population: Point prevalence and the effect on health-related quality of life. *PeerJ*. (2013) 1:e98. doi: 10.7717/peerj.98

10. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Med Care*. (2008) 46:266–74. doi: 10.1097/MLR.0b013e318160d093

11. Villarreal-Zegarra D, Copez-Lonzoy A, Bernabe-Ortiz A, Melendez-Torres GJ, Bazo-Alvarez JC. Valid group comparisons can be made with the Patient Health Questionnaire (PHQ-9): A measurement invariance study across groups by demographic characteristics. *PloS One*. (2019) 14:e0221717. doi: 10.1371/journal.pone.0221717

12. Jose H, Oliveira C, Costa E, Matos F, Pacheco E, Nave F, et al. Anxiety and depression in the initial stage of the COVID-19 outbreak in a Portuguese sample: Exploratory study. *Healthcare*. (2023) 11:659. doi: 10.3390/healthcare11050659

13. Pieh C, Budimir S, Humer E, Probst T. Comparing mental health during the COVID-19 lockdown and 6 months after the lockdown in Austria: A longitudinal study. *Front Psychiatry*. (2021) 12:625973. doi: 10.3389/fpsyt.2021.625973

14. Shevlin M, McBride O, Murphy J, Miller JG, Hartman TK, Levita L, et al. Anxiety, depression, traumatic stress and COVID-19-related anxiety in the UK general population during the COVID-19 pandemic. *BJPsych Open*. (2020) 6:e125. doi: 10.1192/bjo.2020.109

15. Stocker R, Tran T, Hammarberg K, Nguyen H, Rowe H, Fisher J. Patient Health Questionnaire 9 (PHQ-9) and General Anxiety Disorder 7 (GAD-7) data contributed by 13,829 respondents to a national survey about COVID-19 restrictions in Australia. *Psychiatry Res*. (2021) 298:113792. doi: 10.1016/j.psychres.2021.113792

16. Shin C, Ko YH, An H, Yoon HK, Han C. Normative data and psychometric properties of the Patient Health Questionnaire-9 in a nationally representative Korean population. *BMC Psychiatry*. (2020) 20:194. doi: 10.1186/s12888-020-02613-0

17. Tomitaka S, Kawasaki Y, Ide K, Akutagawa M, Yamada H, Ono Y, et al. Distributional patterns of item responses and total scores on the PHQ-9 in the general population: data from the National Health and Nutrition Examination Survey. *BMC Psychiatry*. (2018) 18:108. doi: 10.1186/s12888-018-1696-9

18. Santos IS, Tavares BF, Munhoz TN, Almeida LSPD, Silva NTBD, Tams BD, et al. Sensibilidade e especificidade do Patient Health Questionnaire-9 (PHQ-9) entre adultos da populacão geral. *Cadernos Saude Publica*. (2013) 29:1533–43. doi: 10.1590/S0102-311X2013001200006

19. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. *Gen Hosp Psychiatry*. (2013) 35:551–5. doi: 10.1016/j.genhosppsych.2013.04.006

20. Stochl J, Fried EI, Fritz J, Croudace TJ, Russo DA, Knight C, et al. On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment*. (2022) 29:355–66. doi: 10.1177/1073191120976863

21. 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) 276:220–33. doi: 10.1016/j.jad.2020.06.066

22. Patel JS, Oh Y, Rand KL, Wu W, Cyders MA, Kroenke K, et al. Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. *Depression Anxiety*. (2019) 36:813–23. doi: 10.1002/da.22940

23. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *psychol Bull*. (2017) 143:783–822. doi: 10.1037/bul0000102

24. Kish L. A procedure for objective respondent selection within the household. *J Am Stat Assoc*. (1949) 44:380–7. doi: 10.1080/01621459.1949.10483314

25. 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) 28:71–7. doi: 10.1016/j.genhosppsych.2005.07.003

26. Spitzer RL, Kroenke K, Williams JBW. Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Internal Med*. (2006) 166:1092. doi: 10.1001/archinte.166.10

27. Derogatis LR, Fitzpatrick M. *The SCL-90-r, the brief symptom inventory (BSI), and the BSI-18*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers (2004).

28. Franke GH, Jaeger S, Glaesmer H, Barkmann C, Petrowski K, Braehler E. Psychometric analysis of the brief symptom inventory 18 (BSI-18) in a representative German sample. *BMC Med Res Method*. (2017) 17:14. doi: 10.1186/s12874-016-0283-3

29. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Software*. (2011) 45:1–67. doi: 10.18637/jss.v045.i03

30. Gierk B, Kohlmann S, Toussaint A, Wahl I, Brünahl CA, Murray AM, et al. Assessing somatic symptom burden: A psychometric comparison of the patient health questionnaire—15 (PHQ-15) and the somatic symptom scale—8 (SSS-8). *J Psychosomatic Res*. (2014) 78:352–5. doi: 10.1016/j.jpsychores.2014.11.006

31. Kliem S, Lohmann A, Klatt T, Mößle T, Rehbein F, Hinz A, et al. Brief assessment of subjective health complaints: Development, validation and population norms of a brief form of the Giessen Subjective Complaints List (GBB-8). *J Psychosomatic Res*. (2017) 95:33–43. doi: 10.1016/j.jpsychores.2017.02.003

32. McNeish D. Thanks coefficient alpha, we'll take it from here. *psychol Methods*. (2018) 23:412–33. doi: 10.1037/met0000144

33. Jorgensen TD, Pornprasertmanit S, Schoemann AM, Rosseel Y. *semTools: Useful tools for structural equation modeling*. (2021), R package version 0.5-5.

34. Rosseel Y. lavaan: An R package for structural equation modeling. *J Stat Software*. (2012) 48:1–36. doi: 10.18637/jss.v048.i02

35. Wu H, Estabrook R. Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*. (2016) 81:1014–45. doi: 10.1007/s11336-016-9506-0

36. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct Equation Modeling: A Multidiscip J*. (2007) 14:464–504. doi: 10.1080/10705510701301834

37. Sass DA, Schmitt TA, Marsh HW. Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Struct Equation Modeling: A Multidiscip J*. (2014) 21:167–80. doi: 10.1080/10705511.2014.882658

38. Kliem S, Baier D, Kröger C. Domestic violence before and during the COVID-19 Pandemic–A comparison of two representative population surveys. *Dtsch Arztebl Int*. (2021) 118:483–4. doi: 10.3238/arztebl.m2021.0267

39. Sachser C, Olaru G, Pfeiffer E, Brähler E, Clemens V, Rassenhofer M, et al. The immediate impact of lockdown measures on mental health and couples' relationships during the COVID-19 pandemic: Results of a representative population survey in Germany. *Soc Sci Med*. (2021) 278:113954. doi: 10.1016/j.socscimed.2021.113954

40. Streit F, Zillich L, Frank J, Kleineidam L, Wagner M, Baune BT, et al. Lifetime and current depression in the German National Cohort (NAKO). *World J Biol Psychiatry*. (2022) 24:865–80. doi: 10.1080/15622975.2021.2014152

41. Erhardt A, Gelbrich G, Klinger-König J, Streit F, Kleineidam L, Riedel-Heller SG, et al. Generalised anxiety and panic symptoms in the German National Cohort (NAKO). *World J Biol Psychiatry*. (2022) 24:881–96 doi: 10.1080/15622975.2021.2011409

42. Hyland P, Shevlin M, Murphy J, McBride O, Fox R, Bondjers K, et al. A longitudinal assessment of depression and anxiety in the Republic of Ireland before and during the COVID-19 pandemic. *Psychiatry Res*. (2021) 300:113905. doi: 10.1016/j.psychres.2021.113905

43. Kliem S, Lohmann A, Mößle T, Brähler E. Psychometric properties and measurement invariance of the beck hopelessness scale (BHS): results from a german representative population sample. *BMC Psychiatry*. (2018) 18:110. doi: 10.1186/s12888-018-1646-6

44. Kliem S, Mößle T, Rehbein F, Hellmann DF, Zenger M, Brähler E. A brief form of the perceived social support questionnaire (f-SozU) was developed, validated, and standardized. *J Clin Epidemiol*. (2015) 68:551–62. doi: 10.1016/j.jclinepi.2014.11.003