BMC Psychology

RESEARCH Open Access

Disclosing depressive symptoms: perceived sensitivity of PHQ-9 items in a general population sample

Kristín Hulda Kristófersdóttir[1*], Vaka Vésteinsdóttir[1], Hafrún Kristjánsdóttir[2], Thorlakur Karlsson[3] and Fanney Thorsdottir[1]

Abstract

Background The Patient Health Questionnaire-9 (PHQ-9) is one of the most widely used tools for screening and assessing depression. However, previous research has yielded inconsistent results regarding its factor structure, with studies suggesting either a one- or two-factor model. One possible explanation is socially desirable responding (SDR), which may arise if some items are perceived as more sensitive than others. This study examines whether such differences in perceived sensitivity exist and how they correspond to patterns reported in prior factor-analytic research.

Methods A total of 273 participants completed 36 paired comparisons of the PHQ-9 items, indicating which symptoms they would find more uncomfortable to disclose. Additionally, absolute judgments were collected, where participants rated each item as either uncomfortable or not uncomfortable to disclose. Data were analyzed using a paired comparisons model rooted in Thurstone's law of comparative judgment to estimate the relative sensitivity of each item and whether they were more or less likely to be judged as (not) uncomfortable to disclose. Kendall's coefficients of consistence and agreement were calculated to evaluate the internal consistency of participants' responses and the level of agreement between them.

Results Results showed that cognitive/affective symptoms, such as feelings of worthlessness and depressed mood were perceived as more sensitive than somatic symptoms like fatigue and sleep disturbances. Notably, the sensitivity estimates obtained in this study align closely with prior factor analytic findings that have supported a two-factor model distinguishing cognitive/affective and somatic symptoms.

Conclusions These findings suggest that social desirability may contribute to the underreporting of certain depression symptoms, potentially helping to explain inconsistencies in the PHQ-9's factor structure. Researchers and clinicians should consider the impact of SDR when interpreting PHQ-9 scores to enable more accurate assessments of depression symptom severity.

Keywords PHQ-9, Depression assessment, Socially desirable responding, Item sensitivity, Paired comparisons, Factor structure

*Correspondence:
Kristín Hulda Kristófersdóttir
kristinhk@hi.is
[1]Faculty of Psychology, School of Health Sciences, University of Iceland, Sæmundargata 12, Reykjavík 102, Iceland
[2]Department of Sport Science, School of Social Sciences, Reykjavík University, Reykjavík, Iceland
[3]Brandr, Reykjavík, Iceland

The Patient Health Questionnaire-9 (PHQ-9) [1] is one of the most widely used instruments for assessing depressive symptoms in both clinical practice and research [2]. Its nine items correspond to the diagnostic criteria for major depressive disorder (MDD) in the Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV) [3], and the instrument is routinely scored by summing responses across items to obtain a single index of depression severity. The sum score guides clinical decision-making, screening cut-offs, and the monitoring of treatment progress in primary care and mental health services worldwide [2, 4].

The measurement practices surrounding the PHQ-9 can be traced to its conceptual origins. Although the DSM diagnostic system is atheoretical in the sense that it does not specify the causal mechanisms that generate depressive symptoms, it implicitly treats MDD as a single syndrome [5]. This syndromal perspective, together with long-standing assumptions in clinical psychiatry that symptoms of mental disorders are the downstream effects of a common pathological process[1] (e.g., neurochemical imbalance, genetic risk, or brain abnormalities), has reinforced the idea that depressive symptoms arise from a single underlying cause [7–10].

Within this context, it is unsurprising that PHQ-9 items are routinely summed into a single total score and interpreted as reflecting overall depression severity. This scoring practice is often supported by evidence that a unidimensional CFA model provides an acceptable fit to item covariances.[2] Indeed, if a common-cause or unidimensional structure fully accounted for item relationships, factor analyses should consistently recover a dominant general factor across samples and settings. Yet empirical research on the factor structure of the PHQ-9 has produced mixed results. In the following sections, we review these findings and highlight emerging trends and inconsistencies that call for alternative explanations.

Debates on the PHQ-9 factor structure

Some studies using exploratory and confirmatory factor analysis (EFA and CFA) report that a single factor best explains the observed relations between the PHQ-9 items [13–29], while other studies have concluded that the PHQ-9 consists of two correlated factors: a somatic factor (e.g., fatigue, sleep disturbances) and a cognitive/affective (or non-somatic) factor (e.g., depressed mood, feelings of worthlessness) [30–40].

Closer examination of this literature reveals several recurring patterns. EFAs generally identify a single dominant factor accounting for 40–70% of the total variance, with a smaller secondary factor explaining roughly 4–11%. When eigenvalues are reported, the first is 5 to 11 times larger than the second, and the second is not much larger than the eigenvalues associated with the remaining factors. Yet, interpretations vary: some view this as evidence of unidimensionality [16, 20, 21, 24, 25], while others argue for two distinct dimensions [40].

Similarly, studies comparing one- and two-factor CFA models in the same data often find that both structures provide acceptable fit, with the two-factor model generally providing a comparable or better fit than the one-factor model. However, the two factors are typically highly correlated ($r > 0.80$).[3] In light of such high intercorrelation, some researchers conclude that the PHQ-9 is best viewed as a unidimensional instrument, arguing that separate factors would be redundant [14–16, 21, 22, 28], while others have placed greater weight on the better statistical fit of the two-factor model and the potential clinical utility of distinguishing the symptom clusters [32, 37–39].

Finally, although there are some common trends in item assignments within two-factor representations of the PHQ-9, inconsistencies persist regarding which items load on which factor in studies employing EFA.[4] Violations of simple structure are also not uncommon, as studies have found cross-loadings [20, 24, 33, 41] or items with low loadings (< 0.30) on both factors [20, 40]. Furthermore, several different two-factor models have shown good model fit in studies employing CFA [42, 43], highlighting variability in the conceptualization and empirical support for these models. In the next section, we explore what might underly this variability in PHQ-9 factor-analytic results.

Potential reasons for the variability in factor-analytic findings

Several explanations have been proposed to account for the heterogeneity observed in factor-analytic studies of the PHQ-9. Some researchers have interpreted the better fit of two-factor models as evidence that the PHQ-9

[1] A notable example is the serotonin theory of depression, which attributes depressive symptoms to a deficiency in serotonin. Although widely endorsed in both clinical and popular discourse, a recent review found no consistent evidence supporting this theory [6].

[2] Note that statistical unidimensionality does not necessarily imply that depression is a singular underlying condition, nor does it preclude the presence of meaningful multidimensional structure [11, 12].

[3] A single study stands out for reporting a notably lower factor correlation ($r = .30$) than typically found in other studies [31], possibly due to its unique sample characteristics, which consisted primarily of cancer patients receiving palliative care.

[4] The cognitive/affective factor typically includes Items 1 (anhedonia), 2 (depressed mood), 6 (feelings of worthlessness), and 9 (thoughts of death or self-harm), while the somatic factor generally includes Items 3 (sleep disturbances), 4 (fatigue), and 5 (appetite changes). The placement of Items 7 (concentration difficulties) and 8 (psychomotor agitation or retardation) varies; in studies using EFA, they have sometimes had higher loadings on the cognitive/affective factor, and sometimes on the somatic factor. This pattern of item assignments also appears in studies assessing two-factor CFA models, which are generally based on prior EFA findings.

captures two distinct, though related, symptom dimensions of depression [32, 42, 44–46]. This interpretation has motivated suggestions to calculate and report separate somatic and cognitive/affective subscale scores [30, 33, 41, 45, 47]. However, although this distinction has emerged empirically, we are not aware of any theoretical framework that conceptualizes depression as being composed of precisely these two symptom dimensions. Without such a foundation, assigning meaning to these factors risks reifying statistical artifacts as psychologically distinct constructs [11].

Others have attributed the variability to sample-related factors, including differences in clinical status, comorbid health conditions, age, or cultural background [18, 28, 39, 48]. While plausible, these explanations are not strongly supported by empirical evidence. For example, a systematic review by Lamela et al. [42] found no clear relationship between sample characteristics and the resulting PHQ-9 factor structure, suggesting that other mechanisms may contribute to the observed variability.

An alternative explanation for the inconsistent factor-analytic findings for the PHQ-9 is the marked heterogeneity of depression itself [5]. People who meet diagnostic criteria for MDD can differ substantially in the specific symptoms they experience, even when their overall severity is similar [49]. Such heterogeneity may lead certain symptoms to co-occur more frequently across subsets of respondents (e.g., if some predominantly report somatic symptoms, whereas others mainly endorse cognitive/affective symptoms). These patterns of covariance can produce the appearance of multiple factors in factor analyses, even if they do not correspond to distinct underlying constructs, as some researchers have theorized.

However, although heterogeneity in depressive symptomatology may be invoked to interpret two-factor solutions reported across studies, the recurrence of similar item groupings conducted in samples differing in clinical status, demographic composition, and cultural context [42] likewise warrants consideration of alternative methodological explanations for the observed factor structures.

One such mechanism involves method effects, which reflect systematic influences on item responses that are unrelated to the construct of interest [50]. These effects violate the common factor model's assumption of equal item intercepts across respondents, undermining the validity of observed scores. That is, observed differences may reflect biases in item interpretation or response tendencies rather than (or in addition to) true variations in the trait of interest. Such biases can diminish model fit and increase the dimensionality of observed scores [51].

This issue has been specifically noted in the context of positively and negatively worded items, where items assumed to measure the same construct load onto separate factors [52–55]. While the PHQ-9 does not include positively and negatively worded items, another response bias, particularly relevant to depression measures, may similarly distort factor-analytic results: socially desirable responding.

Socially desirable responding and depression items

One specific form of response bias relevant to the PHQ-9 is socially desirable responding (SDR). SDR refers to individuals' tendency to respond in a way that presents themselves favorably to others or conforms to social norms and expectations, rather than providing completely honest or accurate information. Importantly, SDR is a function of the item content; if the content elicits a response that is perceived as socially undesirable, people are more likely to respond in a socially desirable way [56].

Given the sensitive nature of the PHQ-9 items, they may be particularly vulnerable to SDR. There is extensive evidence that shows that people with depression are viewed negatively by the public [57–59], by themselves [60, 61], and perceive others as having negative attitudes towards depression [62, 63]. This societal stigma may contribute to reluctance in disclosing depressive symptoms on self-report scales. For example, Hunt et al. [64] found that individuals reported significantly more symptoms when the purpose of the depression instrument was masked, suggesting that social desirability concerns can suppress honest responding on self-report measures.

The sensitive nature of the PHQ-9 items could also evoke a perceived threat of disclosure. That is, people may worry about potential consequences if their responses were to become known to others [65]. These anticipated consequences can be social, such as fear of disapproval, making social desirability a specific subtype of broader disclosure threat concerns. Individuals may also underreport symptoms to avoid worrying friends, family, or colleagues, or to prevent unwanted interventions. Evidence of such response behavior has been documented by Robinson et al. [66], who found that some primary care patients deliberately omitted or underscored PHQ-9 items to avoid triggering further clinical action. Disclosing mental health issues may also carry perceived professional consequences in fields involving public safety or high-stakes decision-making, such as aviation [67], law enforcement [68], or healthcare [69], where individuals often fear loss of licensure, restricted duties, or career setbacks.

Given the stigma surrounding depression and the perceived threat of disclosure, it is reasonable to assume that the PHQ-9 items are susceptible to underreporting simply due to their sensitive nature. However, some of the PHQ-9 items are more likely to result in underreporting

because they describe symptoms that are inherently more sensitive than others. That is, the content of the items is not equally undesirable. For example, Item 9 (Thoughts that you would be better off dead, or of hurting yourself in some way) is unquestionably more undesirable than Item 4 (Feeling tired or having little energy). If certain items are perceived as more socially undesirable, they are more likely to be underreported. This could introduce systematic error variance among the more sensitive items, inflating their observed associations while attenuating their relations with less sensitive items. Such patterns may influence the factor structure recovered in statistical models, potentially contributing to the appearance of distinct dimensions that partly reflect variation in item sensitivity rather than or in addition to differences in underlying symptom domains.

Evidence from qualitative studies supports the notion that people are tempted to underreport some of the items [66, 70–72]. For example, Malpass et al. [71] found a discrepancy between people's responses to Item 9 and subsequent qualitative interviews that highlight potential reluctance to endorse thoughts of death or self-harm. Furthermore, signs of SDR have been identified using cognitive interviews in Items 1 (anhedonia), 2 (depressed mood), 6 (feelings of worthlessness), and 9 (thoughts of death or self-harm) [70]. Notably, these items tend to be assigned to the cognitive/affective factor in many PHQ-9 factor-analytic studies, raising the possibility that item sensitivity may contribute to their observed interrelations.

The potential impact of SDR on PHQ-9 scores is, therefore, a topic that deserves more attention. Not only does it have implications for validity research examining the factor structure of the PHQ-9, but also for studies using the instrument for other research purposes, and for its use in clinical practice. The consequences of systematic underreporting because of SDR in research settings include underrepresentation of the true prevalence and severity of depressive symptoms in the studied population, and researchers drawing inaccurate conclusions about the effectiveness of treatment interventions. Furthermore, SDR could inflate or deflate observed associations between PHQ-9 scores and other variables of interest or alter the strength of those associations. Possible consequences in clinical practice include people getting suboptimal or ineffective treatment that does not address the actual severity or nature of the depressive symptoms. Also, if clinicians are not aware of the true extent of a patient's depressive symptoms, timely and appropriate interventions may be delayed, potentially worsening the condition. Finally, if prevalence rates of depression are underestimated, healthcare resources may be misallocated, leading to insufficient support for individuals in need.

Despite the theoretical and practical relevance of SDR for understanding variability in the PHQ-9 's factor structure, no prior research has systematically investigated whether specific items differ in their sensitivity. This study addresses that gap by using the method of pair comparisons to assess the perceived sensitivity of each PHQ-9 item. We hypothesize that cognitive/affective items will be perceived as more sensitive than somatic items. By identifying systematic differences in item sensitivity, this study provides a foundation for future work exploring the impact of SDR on PHQ-9 scores.

Method

Participants

Participants were recruited from an online research panel consisting of approximately 1,500 Icelandic residents aged 18 years or older (53% women). Panel members had been randomly selected from Registers Iceland and had agreed to participate in methodological online surveys. Overall, 344 individuals opened the survey link, and 273 completed the entire survey. The final sample consisted of 59% women ($n = 162$) and 40% men ($n = 110$), with ages ranging from 19 to 81 years ($M = 53$, $SD = 14$). One participant did not want to reveal their gender, and three did not provide their age. Participants did not receive any reimbursements for their participation.

Instrument

The PHQ-9 [1] is an instrument that was designed to screen for depression and to assess its severity. It consists of nine items, each rated on a 4-point ordinal scale ranging from 0 (*Not at all*) to 3 (*Nearly every day*), based on the frequency of symptoms over the past two weeks. Total scores range from 0 to 27, with higher scores indicating greater severity of depressive symptoms. For screening purposes, a score of 10 or higher is generally used as a depression screening cutoff.

The translation process for the original Icelandic version has been described elsewhere [70]. However, since that version contains several translation errors, an updated version was developed, in which these errors were corrected. Furthermore, given that Icelandic is a gendered language, the PHQ-9 was administered in one of three gender-specific versions (masculine, feminine, neuter) depending on how participants responded to the gender question at the beginning of the survey. Participants who preferred not to disclose their gender were asked how they would like to be addressed.

Procedure

A survey invitation was sent out via email to the panel members in August 2024. The survey was open for two weeks, during which two reminder emails were sent to

those who had not yet responded. Data were collected using the online survey platform QuestionPro.

On the first page of the survey, participants were provided with information about the study and were asked to provide their consent to participate. Those who consented were then asked about their birth year and gender on the following page, after which they were shown the task instructions (see Additional file 1).

Participants had to confirm that they had read the instructions carefully to continue to the next page, after which they then proceeded to three practice questions designed to familiarize them with the task of making paired comparisons. This step was intended to minimize the likelihood of people not understanding the task when presented with the PHQ-9 paired comparisons.

Upon completing the practice questions, participants were informed that they would next answer 36 questions in the same format. They were asked to confirm their understanding; those who indicated they understood proceeded directly to the PHQ-9 comparisons, while those who did not received the instructions again before continuing.

In each paired comparison (including practice comparisons), participants were asked: "Imagine that you have experienced the symptoms below in the past two weeks. Which symptom would you feel more uncomfortable disclosing?" This wording was chosen instead of the more direct question "Which symptom is more *socially undesirable*?" to better capture the intended construct (i.e., overall perceived sensitivity) while accommodating the possibility that respondents might find the concept of social desirability abstract or unfamiliar, particularly in contexts where the term is not commonly used. Below the comparison question were two depression items presented side by side, and participants selected one before moving to the next pair of items on the following page. For the practice questions, four items from the Edinburgh Postnatal Depression Scale (EPDS) [73] were randomly paired and presented in three different ways.

The PHQ-9 comparisons (36 in total) involved each of the nine items being compared to every other item. To control for potential order effects, Ross's algorithm was used, ensuring each item appeared approximately equally spaced throughout the list of pairs and equally often as the first or second member of a pair [74, 75]. To further balance potential fatigue effects, as recommended by Ross, half of the participants received this optimally spaced and balanced ordering in reverse.

After completing the paired comparisons, participants were presented with all the PHQ-9 items again on the same page and asked to make absolute judgments about each item's sensitivity. For each item, they selected either "Uncomfortable" or "Not uncomfortable" to indicate whether they would generally find it uncomfortable to disclose. The order of the items and response options was randomized.

Statistical analysis

All analyses were conducted using the R Statistical Software v4.4.0 [76]. The *eba* package [77] was used to compute the information on internal consistency and agreement between participants.

Internal consistency and agreement between participants

The number of circular triads made by each participant was calculated to assess the internal consistency of the comparative judgments. A circular triad is formed when a respondent is inconsistent in their comparisons. For example, if Item A is preferred to Item B, and Item B is preferred to Item C, then Item A should also be preferred to Item C. If, however, Item C is preferred to Item A, a circular triad would be formed. The number of circular triads (T) was used to calculate Kendall's coefficient of consistence (*zeta*) [78, 79] for each participant by subtracting the proportion of circular triads present in their responses from 1, giving the proportion of consistent responses ($zeta = 1 - \frac{T}{T_{max}}$). When the number of items (m) being judged is odd (as with the nine PHQ-9 items), the maximum possible number of circular triads (T_{max}) is found with $\frac{m^3-m}{24}$, giving 30 possible triads [79]. The *zeta* coefficient ranges from 0, indicating completely inconsistent judgments ($T = T_{max}$), to 1, indicating completely consistent judgments ($T = 0$).

Many factors could explain a low value of the coefficient of consistence. The participant may not be interested in the task, resulting in careless responding, or they might not have read or understood the task. Another possibility is that it might be too difficult to distinguish the items on the psychological continuum in question (the items are closely spaced), or that the items do not fall along the single dimension that we are trying to scale [80]. A high value, on the other hand, indicates that respondents make consistent judgments (not made at random) and that they are capable of sufficiently differentiating between the items.

To minimize the influence of highly inconsistent response patterns, we excluded participants who produced 15 or more circular triads. This threshold reflects cases where inconsistencies occurred in at least half of the 30 possible triads. It corresponds to a *zeta* coefficient of ≤ 0.50, which we interpret as indicating, at most, moderate consistency given the theoretical range of the coefficient.

To assess the agreement in participants' comparative judgments, Kendall's coefficient of agreement (u) was computed [78, 79]. The coefficient ranges between

0 (minimum agreement)[5] to 1 (maximum agreement). A chi-square test was also conducted to assess the null hypothesis that the agreement between participants was by chance.

The method of pair comparisons

The method of pair comparisons involves presenting respondents with two stimuli at a time (in this case, items) and asking them to indicate which one is of greater quantity than the other in some defined respect [81]. From these responses, it is possible to derive a single value for each item that locates it on a linear scale. The method is rooted in Thurstone's law of comparative judgment [82], which posits that each stimulus elicits a *discriminal process*, a psychological mechanism through which individuals react to the stimulus. These discriminal processes are conceptualized as values on a psychological continuum, where the distribution of reactions to any given stimulus is assumed to follow a normal distribution. The mode of this distribution is taken as the *scale value* of the given stimulus, and the standard deviation, Thurstone referred to as the *discriminal dispersion.* Thurstone's law assumes that when two stimuli are compared, the probability of one being preferred over the other depends on the difference between their underlying scale values [80, 81, 83].

A key advantage of paired comparisons is that they rely on relative rather than absolute judgments. As Nunnally and Bernstein [84] note (pp. 51), "one of psychology's truisms is that people are almost invariably better (more consistent and/or accurate) at making comparative responses than absolute responses," largely because comparative judgments avoid the frame-of-reference problem. Asking "How sensitive is the content of this item?" (requesting an absolute response) could raise other questions, such as "Compared to what?", "In what context?", and "Who would see the response?". With paired comparisons, the focus is instead on the relative sensitivity between two items at a time, thereby minimizing context-dependent variability. By evaluating every possible pair, this method allows for a more nuanced estimation of each item's position than conventional absolute ratings.

Although relative judgements offer these advantages, the resulting sensitivity estimates do not indicate whether an item is perceived as sensitive in an absolute sense. To address this, we also collected absolute ratings in which participants classified each item as either "uncomfortable" or "not uncomfortable" to disclose. These classifications were used to anchor the relative scale values to a meaningful zero point: the position where 50% of respondents would judge an item as uncomfortable and 50% as not uncomfortable to disclose [81]. Positive scale values thus indicate greater susceptibility to SDR (i.e., people generally judged the item's symptoms as uncomfortable to disclose), whereas negative scale values indicate lower susceptibility (i.e., people generally judged the item's symptoms as not uncomfortable to disclose).

Four methods were considered to estimate the scale values and their standard deviations: Thurstone's Case V and Case III models [81], the normal approximation of the binomial distribution (the binomial model), and the beta density function (the beta model) [85]. After considering the limitations of each model and assessing their goodness of fit to the data, the binomial model was selected to represent the final scale values. Details of the model selection procedure are provided in Additional file 2. The scale values were calculated following procedures outlined by Guilford (pp. 160–174) [81] and Wild and Cabral [85]. The unit of the resulting sensitivity/social undesirability scale was converted to two standard deviations for easier interpretation.

Results

Internal consistency and agreement between participants

The number of circular triads and Kendall's coefficient of consistence were computed for each participant to assess how consistent they were in their judgments. Eight respondents were excluded from further analysis due to low internal consistency ($zeta \leq 0.50$). Descriptive statistics for the internal consistency of the remaining participants ($n = 265$) are presented in Table 1. In general, people were highly consistent in their judgments, as 36% of participants had no circular triads and 83% had 5 or fewer circular triads, resulting in a mean zeta coefficient of 0.91.

The most frequent circular triad involved Items 1 (anhedonia), 2 (depressed mood), and 8 (psychomotor agitation or retardation), with 12% of participants ranking these items inconsistently.[6] This suggests that participants found it most difficult to consistently judge the sensitivity of these particular symptoms in relation to each other. Other item combinations within triads were made by less than 8% of the participants.

Examining the frequency of individual items across all triads revealed that Items 6 (feelings of worthlessness), 9

[5] More specifically, the value for minimum agreement equals $-1/(n-1)$, if n (the number of participants) is even, and $-1/n$, if n is odd.

Table 1 Descriptive statistics for the number of circular triads and Kendall's coefficient of consistence ($n = 265$)

	Mean	SD	Median	Min	Max
Circular Triads	2.55	3.25	1.00	0.00	14.00
Zeta	.91	.11	.97	.53	1.00

[6] The ranking patterns were either a) Item 1 > Item 8, Item 8 > Item 2, Item 2 > Item 1 ($n = 21$), or b) Item 1 > Item 2, Item 2 > Item 8, Item 8 > Item 1 ($n = 11$).

(thoughts of death or self-harm), and 4 (fatigue) appeared the least often, being included in 21–22% of all circular triads. This indicates that these items were easier for participants to rank consistently in terms of sensitivity. The frequency of the remaining items across the triads in the sample ranged from 30% (Item 7, concentration difficulties) to 40% (Item 8, psychomotor agitation or retardation).

Kendall's coefficient of agreement was calculated to assess the consistency between participants' rankings in the paired comparisons. The results showed a u value of 0.48, indicating a moderate level of agreement. The chi-square test further indicated that the observed agreement was unlikely to be compatible with the null hypothesis of no agreement ($\chi^2 = 4605.6$, $df = 36.4$, $p < 0.001$).

Scale values

Results from the paired comparisons can be found in Table 2 and Fig. 1. Cognitive/affective symptoms (Items 2, 6, and 9) were judged as substantially more sensitive (i.e., uncomfortable to disclose) than somatic symptoms such as appetite changes (Item 5), sleep disturbances (Item 3), and fatigue (Item 4), which most respondents judged as not sensitive. These cognitive/affective items could, therefore, be more susceptible to SDR compared to the somatic ones. Items reflecting anhedonia (Item 1) and psychomotor changes (Item 8) fell closest to the neutral point, indicating that respondents were divided on whether these symptoms would be uncomfortable to disclose.

The standard deviations of the scale values (see Sigma in Table 2) were relatively similar across most items (approximately 1.2–1.4), suggesting comparable variability in judgments. The lowest values were observed for Items 6 and 9 (0.9 and 0.8, respectively), reflecting stronger agreement that feelings of worthlessness and thoughts of death or self-harm are especially sensitive symptoms. This pattern can also be seen in panel a) of Fig. 1, where the theoretical distribution curves for these items are narrower with higher peaks.

Panel a) of Fig. 1 also shows that there is considerable overlap among the scale value distributions of certain items. For example, Items 3 and 4 share nearly identical curves, suggesting that sleep disturbances and fatigue are perceived very similarly in terms of sensitivity. Likewise,

Table 2 Estimated PHQ-9 item sensitivity scale values and their estimated standard deviations ($n = 265$)

Item	Absolute prop	Scale Value	Sigma
9. Thoughts that you would be better off dead or of hurting yourself in some way	.97	3.54	0.84
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	.91	2.61	0.91
2. Feeling down, depressed, or hopeless	.73	1.12	1.18
1. Little interest or pleasure in doing things	.54	0.23	1.25
8. 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	.49	0.15	1.34
5. Poor appetite or overeating	.38	−0.54	1.39
7. Trouble concentrating on things, such as reading the newspaper or watching television	.22	−1.65	1.27
3. Trouble falling or staying asleep, or sleeping too much	.09	−2.43	1.20
4. Feeling tired or having little energy	.11	−2.54	1.16

The items are arranged in descending order according to their scale value. Positive values indicate higher susceptibility to SDR (symptoms judged uncomfortable to disclose), and negative values indicate lower susceptibility to SDR (symptoms judged not uncomfortable to disclose). Absolute Prop. = The proportion of participants judging the item as uncomfortable to disclose instead of not uncomfortable. Sigma = Estimated standard deviations for the scale values

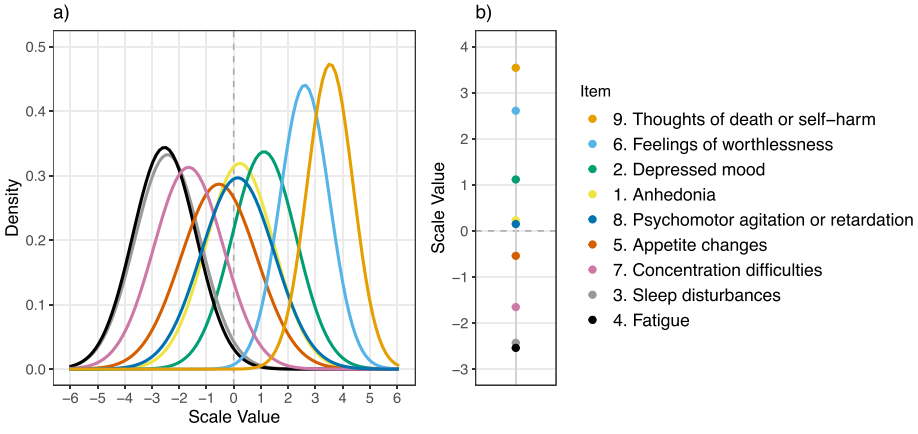

Fig. 1 Estimated PHQ-9 item sensitivity scale values and their theoretical distribution curves ($n = 265$). Panel **a)** shows the theoretical distribution curves for the PHQ-9 sensitivity scale values, while panel **b)** displays the estimated scale values in descending order along the y-axis. Positive values indicate higher susceptibility to SDR (symptoms judged uncomfortable to disclose), and negative values indicate lower susceptibility to SDR (symptoms judged not uncomfortable to disclose)

the distribution for Item 1 nearly overlaps that of Item 8, indicating that respondents rated the symptom of anhedonia similarly to psychomotor agitation and retardation. In contrast, Items 3, 4, and 7, which are positioned at the lower end of the scale, and Items 6 and 9, situated at the higher end of the scale, show almost no overlap. These two groups of items are largely separated by the zero point on the scale (except for Item 7, whose distribution extends slightly further into the positive range), showing they were consistently judged as either uncomfortable or not uncomfortable to disclose.

Discussion

This study investigated whether PHQ-9 items differ in perceived sensitivity and whether such differences across items may contribute to inconsistencies in factor-analytic findings. While our data do not provide direct evidence that sensitivity leads to biased reporting, the consistent differences in how items were perceived raise the possibility that SDR may contribute to observed factor structures. In particular, items frequently assigned to a cognitive/affective factor in prior research were perceived as more sensitive, suggesting a potential link between item sensitivity and factor structure. To our knowledge, this possibility has been largely overlooked in prior psychometric research on the PHQ-9.

Our results suggest that the appearance of two factors in some prior studies may not necessarily indicate distinct substantive dimensions of depression, as has sometimes been hypothesized [32, 38, 42, 44–46] but could instead reflect differences in how honestly respondents report certain symptoms. Items that have consistently loaded onto the cognitive/affective factor, specifically, Item 2 (depressed mood), 6 (feelings of worthlessness), and 9 (thoughts of death or self-harm), were rated as more uncomfortable to disclose in the current study. Their heightened sensitivity may lead to systematic underreporting, which could artificially inflate the correlations among these more sensitive items. As a result, these items may appear to form a coherent factor, even if the grouping does not reflect a substantive construct. This interpretation aligns with qualitative findings indicating that people may be tempted to underreport these items due to their sensitive nature [66, 70–72].

In contrast, somatic symptoms like sleep difficulties (Item 3), fatigue (Item 4), and appetite changes (Item 5) were perceived as less sensitive. These items may, therefore, be more likely to be reported accurately, resulting in weaker correlations with the more sensitive cognitive/affective items when SDR is present. This pattern could help explain why somatic and cognitive/affective factors have emerged in some prior factor analyses.

However, the influence of SDR likely varies across studies, as it is a context-sensitive bias [86]. Its impact may depend on factors such as the administration setting, perceived anonymity, and sample characteristics. For example, people may feel more pressure to underreport depressive symptoms in clinical settings or face-to-face interviews than in anonymous online surveys [87]. Structural validity evidence for a one-factor PHQ-9 model has, indeed, been stronger in paper-and-pencil and online formats than in interviews [43]. Additionally, because depression stigma seems to vary across cultures [88], SDR may be more pronounced in some cultural contexts than others. When SDR is stronger, systematic underreporting of the more sensitive symptoms can reduce their correlations with the remaining items, potentially yielding stronger support for a two-factor solution in EFA or a better-fitting two-factor CFA model. When SDR is weaker, these items correlate more strongly with the rest of the scale, producing a structure that appears more unidimensional. Such variation may help explain why some studies conclude in favor of a two-factor representation of the PHQ-9, while others support a single dominant factor, despite the consistent item content of the PHQ-9.

This raises an important question: why are cognitive/affective symptoms, in particular, perceived as more sensitive than somatic symptoms? One possibility is that they feel more self-revealing or stigmatizing, potentially signaling emotional instability or a psychiatric condition. In contrast, somatic symptoms may be seen as more normative and less personally exposing. While speculative, this may help explain why certain items are more prone to SDR.

Item-level patterns

Looking more closely at the results for individual items, some interesting patterns are worth highlighting. For instance, concentration difficulties (Item 7), despite being a cognitive symptom, were generally not perceived as uncomfortable to disclose, ranking even lower in sensitivity than the somatic symptom of appetite change (Item 5). This may reflect the growing normalization of attention-related issues in public discourse, particularly concerning conditions such as ADHD [89]. As a result, concentration problems may be seen as less socially undesirable to report than other cognitive/affective symptoms of depression, such as depressed mood or feelings of worthlessness. This interpretation aligns with prior factor-analytic research, in which Item 7 has been inconsistently assigned to either the somatic or cognitive/affective factor. Furthermore, cognitive interview data suggest that respondents tend to endorse this item for reasons unrelated to depression, and that the examples provided in the item ("…such as reading the newspaper or watching television") can cause various comprehension issues [70].

Participants were in the least agreement about the sensitivity of Items 1 (anhedonia) and 8 (psychomotor agitation or retardation). These items fell near the midpoint of the sensitivity continuum, suggesting that people are equally likely to find them either uncomfortable or not uncomfortable to disclose. This ambiguity indicates that these symptoms may be less clearly defined in terms of their sensitivity compared to more overtly sensitive symptoms, such as suicidal ideations or feelings of worthlessness. Given that Items 1 and 8 were most frequently involved in circular triads, in conjunction with Item 2 (depressed mood), further supports the idea that participants struggled to consistently differentiate the sensitivity of these items. In particular, Item 8 appeared most frequently within all circular triads observed in the study, highlighting that respondents found it especially challenging to place it consistently on the continuum. This may be due to the complexity of Item 8, which is not only the most wordy item in the PHQ-9 but also represents symptoms many people may not be as familiar with. Unlike more familiar symptoms such as fatigue or mood changes, psychomotor symptoms might not be as easily recognized or understood, making their sensitivity more difficult to assess. In line with this, Item 8 has consistently shown suboptimal psychometric properties in previous studies examining the scalability of the PHQ-9 items [90–92], and its assignment to either the somatic or cognitive/affective factor in two-factor models has not been consistent across studies. The scale value for Item 8 in this study may, therefore, not only reflect the sensitivity of its contents but also the complexity and ambiguity associated with the psychomotor symptoms themselves and the lengthy wording of the item.

Since Item 5 (appetite change) has consistently been assigned to the somatic factor in two-factor representations of the PHQ-9 items, alongside Items 3 (sleep disturbances) and 4 (fatigue), we anticipated its scale value to align more closely with these items. However, while it received a negative scale value, indicating that it was generally not perceived as uncomfortable to disclose, its position was relatively closer to the point of indifference than Items 3 and 4. This discrepancy might stem from the item's double-barreled and polar-opposite phrasing, which includes both "poor appetite" and "overeating." It is plausible that respondents perceived one aspect of the item (e.g., overeating) as much more (or less) sensitive than the other (e.g., poor appetite). Consequently, it is unclear which of the two symptoms dominated participants' judgments when comparing this item to the others. This ambiguity may have introduced greater variability in responses, potentially leading to higher levels of disagreement between participants than would otherwise be expected.

Practical implications

Our findings indicate that not all PHQ-9 items are perceived as equally sensitive. This has important implications for the interpretation of self-reported depression symptoms. If some items are viewed as more uncomfortable to disclose, this could influence how respondents answer them, potentially leading to biased reporting.

In clinical settings, this means that more sensitive symptoms, such as those reflected in Items 2 (depressed mood), 6 (feelings of worthlessness), and 9 (thoughts of death or self-harm), may be systematically underreported. Clinicians using the PHQ-9 for screening or treatment planning should be aware that patients may minimize or omit these symptoms. If certain items are consistently underreported, the severity of depressive symptoms may be underestimated, leading to delayed or suboptimal treatment. This is especially concerning for symptoms like suicidal ideation (Item 9), where underreporting could have very serious consequences.

This also complicates using the PHQ-9 for tracking symptom changes over time. If SDR remains constant or even increases as treatment progresses (e.g., due to perceived expectations of improvement), it may obscure real changes in symptom severity. As a result, clinicians could be misled into thinking a treatment is more effective than it is.

In research settings, SDR may attenuate or disattenuate observed associations with other variables and undermine the validity of conclusions drawn from PHQ-9 data. The use of cognitive/affective and somatic subscale scores, as some have suggested [30, 33, 41, 45, 47], may be particularly problematic if one subset of items is more or less likely to be underreported. On a broader scale, systematic underreporting may lead to underestimated prevalence rates of depression and symptom severity, potentially influencing how mental health resources are allocated.

Altogether, these findings emphasize the importance of recognizing SDR as a potential source of measurement bias in the PHQ-9. While further work is needed to determine the extent of actual misreporting and its impact, both clinicians and researchers should interpret PHQ-9 scores with caution.

Methodological considerations

The current findings highlight broader methodological concerns in the psychometric literature on the PHQ-9. Despite recurring patterns in prior factor-analytic studies (e.g., the emergence of a strong first factor, a weaker second factor, and high factor intercorrelations), the conclusions drawn from these patterns have varied. A closer look suggests that methodological decisions and reporting practices may have played a role in this inconsistency.

In EFAs, critical details such as eigenvalues, scree plots, factor correlations, item communalities, or full factor loading matrices are often missing, limiting the transparency and interpretability of results. Moreover, researchers have often relied on simplistic criteria, such as retaining all factors with eigenvalues above 1.0.

In CFAs, model selection is often guided by minor improvements in global fit indices without sufficient theoretical justification. While such indices can provide useful information, they do not distinguish between substantively meaningful structures and patterns that may arise from response biases, such as SDR. This problem is compounded by a broader reporting issue in psychometrics known as *model-HARKing* [93], where models derived from exploratory procedures (e.g., post-hoc model modifications or iterative searches for better fit) are presented as if they were confirmatory or theory-driven from the outset. The problem lies not in exploration itself, but in the lack of transparency about its role. This tendency may have contributed to the proliferation of competing factor models for the PHQ-9 [42, 43].

Moving forward, greater transparency in reporting of analytic decisions, stronger integration of theory, and systematic evaluation of response biases are essential for advancing our understanding of the PHQ-9 ‘s structural validity.

Future directions

A key next step is to test whether perceived item sensitivity translates into actual underreporting in PHQ-9 responses. To that end, we have collected a new dataset from the same online panel in which participants completed the PHQ-9 together with a measure of SDR. This follow-up analysis will examine whether item sensitivity predicts SDR-related method variance in PHQ-9 responses and how accounting for SDR affects the scale ‘s internal structure. A separate manuscript reporting these results is currently in preparation.

More broadly, future research should incorporate methods that explicitly address SDR, such as administering SDR scales alongside the PHQ-9 [94, 95]. Identifying individuals with a strong tendency toward SDR could allow for targeted scoring adjustments or sensitivity analyses. Alternatively, the scale values derived from the current study could serve as item-level weights to apply corrective adjustments to items most susceptible to underreporting. Future research should evaluate whether such corrections improve the PHQ-9 ‘s structural consistency.

In addition to detecting or adjusting for SDR, researchers might also focus on directly modifying the PHQ-9 to reduce the likelihood of SDR. Rewording sensitive items or adding instructions that normalize disclosure may help minimize SDR [96]; however, the evidence on this is mixed [65, 97, 98]. Changing administration protocols may also be beneficial, such as emphasizing confidentiality (when possible), by creating conditions that encourage more honest responding [65, 99].

Strengths and limitations

While this study offers valuable insights into the sensitivity of depression symptoms, several limitations should be acknowledged. First, the sample consisted of Icelandic adults, limiting the generalizability of findings to other cultural contexts. Given that mental health stigma, and by extension, the perceived sensitivity of depression symptoms, likely vary across cultures [88, 100], future research should examine more diverse populations to capture these differences more accurately.

Second, despite its strengths, the paired comparison method has some constraints. Requiring participants to judge all possible pairings of the PHQ-9 items (36 comparisons in total) may have led to fatigue and disengagement. The method also scales poorly as the number of items increases,[7] but with nine items the number of comparisons was still manageable in this study. Furthermore, the task is likely unfamiliar to many people, which may have further increased inconsistent judgments. Although efforts were made to minimize this by including practice comparisons and optimizing the order of item pairs, we still saw that some people were highly inconsistent and had to be left out of the analysis. Interestingly, these eight most inconsistent participants were mostly men over the age of 70, which raises an important consideration for future research: it may be beneficial to explore how different demographic factors, such as age and gender, influence the consistency of responses in paired comparison tasks.

Third, the binary classifications used to anchor the sensitivity scale may be susceptible to context effects. Respondents may imagine different disclosure settings or audiences that influence whether they judge an item as uncomfortable to disclose, introducing potential measurement error in the anchoring judgments. However, because the paired comparison task involves repeated and focused reflection on the discomfort associated with each item, participants are likely to develop a general sense of which items they find uncomfortable. As such, asking for binary classifications *after* completing the comparisons may yield more internally consistent and meaningful responses than if those judgments were made in isolation.

[7] The number of comparisons is obtained with $m(m-1)/2$, where m is the number of items being compared. For example, a 20-item scale would require $20(20-1)/2 = 190$ comparisons! Note that it is possible to implement an incomplete design with planned missingness where each respondent is only administered a subset of all possible comparisons [74].

Lastly, while the study provides evidence that certain PHQ-9 items are perceived as more sensitive than others, it did not directly assess whether this sensitivity results in increased SDR in practice. As such, the link between item sensitivity and actual misreporting behavior remains speculative, and the implications discussed above rely on the assumption that more sensitive items are more likely to be underreported. While theory and prior research support the notion that item sensitivity increases the likelihood of SDR [56, 65, 101], sensitivity alone might not be sufficient to lead to misreporting. Other factors, such as mode of administration and perceived anonymity, also appear to influence the extent of SDR [102, 103].

Despite these limitations, the paired comparisons approach is an advancement over traditional methods of assessing item sensitivity or social desirability. By focusing on comparative rather than absolute judgments, the study minimized context-dependent biases and allowed for a more accurate evaluation of the relative sensitivity of the PHQ-9 items. This study also fills a gap in the literature by being the first to systematically assess the perceived sensitivity of the PHQ-9 items. These results can inform strategies to mitigate SDR to the PHQ-9, as described above.

Abbreviations

CFA	Confirmatory factor analysis
EFA	Exploratory factor analysis
MDD	Major depressive disorder
PHQ-9	Patient Health Questionnaire-9
SDR	Socially desirable responding
SDSV	Social desirability scale value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40359-026-04067-7.

Additional file 1. Procedure: Survey Instructions. Description: This file contains the instructions presented to the participants at the beginning of the survey, before completing the paired comparison task.

Additional file 2. Statistical Analysis: Model Selection. Description: This file outlines the model selection procedure used to estimate item-level sensitivity scale values from the paired comparison data.

Acknowledgements

Not applicable.

Authors' contributions

K.H.K. conceptualized the study, designed the methodology, supervised data collection, conducted the statistical analyses, and wrote the manuscript. F.T. conceptualized the study, contributed to the methodological design, supervised the project, and provided critical input throughout the research process and manuscript revisions. V.V., H.K., and T.K. contributed to the study design and reviewed and provided feedback on the manuscript. V.V. and H.K. also provided guidance on the theoretical framework. All authors reviewed and approved the final manuscript.

Funding

No funding was received for this study.

Data availability

The datasets generated and analyzed during the current study, along with the R code used to conduct the analyses, are publicly available on the Open Science Framework: https://osf.io/mw7y9/.

Declarations

Ethics approval and consent to participate

The study protocol received a positive review from the Science Ethics Committee of Icelandic Universities (approval ID: SHV2024-060). Additional review was sought from the National Bioethics Committee in Iceland, which determined that formal approval was not required. All participants provided consent after being presented with information about the study's purpose, procedures, and their rights as research participants. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 June 2025 / Accepted: 27 January 2026
Published online: 03 February 2026

References

1. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.
2. Kroenke K. PHQ-9: global uptake of a depression scale. World Psychiatry. 2021. https://doi.org/10.1002/wps.20821.
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). 4th ed. 1994.
4. DeJesus RS, Vickers KS, Melin GJ, Williams MD. A system-based approach to depression management in primary care using the patient health questionnaire-9. Mayo Clin Proc. 2007. https://doi.org/10.4065/82.11.1395.
5. Fried EI. Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. Front Psychol. 2015. https://doi.org/10.3389/fpsyg.2015.00309.
6. Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA. The serotonin theory of depression: a systematic umbrella review of the evidence. Mol Psychiatry. 2023. https://doi.org/10.1038/s41380-022-01661-0.
7. Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. Nat Rev Psychol. 2022. https://doi.org/10.1038/s44159-022-00050-2.
8. Deacon BJ. The biomedical model of mental disorder: a critical analysis of its validity, utility, and effects on psychotherapy research. Clin Psychol Rev. 2013. https://doi.org/10.1016/j.cpr.2012.09.007.
9. Bekhuis E, Hartman TCO, Boschloo L, Lucassen PL. A new approach to psychopathology: the example of depression. Br J Gen Pract. 2019. https://doi.org/10.3399/bjgp19X701717.
10. Helén I. The depression paradigm and beyond: the practical ontology of mood disorders. Sci Technol Stud. 2011. https://doi.org/10.23987/sts.101299.
11. Fried EI. Lack of theory building and testing impedes progress in the factor and network literature. Psychol Inq. 2020. https://doi.org/10.1080/1047840X.2020.1853461.
12. Heinrich M, Zagorscak P, Eid M, Knaevelsrud C. Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck Depression Inventory–II. Assessment. 2020. https://doi.org/10.1177/1073191118803738.
13. Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, et al. Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in rural Chiapas, Mexico: a cross-sectional study. J Clin Psychol. 2017. https://doi.org/10.1002/jclp.22390.
14. Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD, (CSRG) Canadian Scleroderma Research Group. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. PLoS ONE. 2012. https://doi.org/10.1371/journal.pone.0052028.

15. Boothroyd L, Dagnan D, Muncer S. PHQ-9: one factor or two? Psychiatry Res. 2019. https://doi.org/10.1016/j.psychres.2018.12.048.
16. Familiar I, Ortiz-Panozo E, Hall B, Vieitez I, Romieu I, Lopez-Ridaura R, et al. Factor structure of the Spanish version of the patient health questionnaire-9 in Mexican women. Int J Methods Psychiatr Res. 2015. https://doi.org/10.1002/mpr.1461.
17. Galenkamp H, Stronks K, Snijder MB, Derks EM. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. BMC Psychiatry. 2017. https://doi.org/10.1186/s12888-017-1506-9.
18. González-Blanch C, Medrano LA, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana JA, Limonero JT, et al. Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS ONE. 2018. https://doi.org/10.1371/journal.pone.0193356.
19. Harry ML, Waring SC. The measurement invariance of the patient health questionnaire-9 for American Indian adults. J Affect Disord. 2019. https://doi.org/10.1016/j.jad.2019.05.017.
20. Kalpakjian CZ, Toussaint LL, Albright KJ, Bombardier CH, Krause JK, Tate DG. Patient health questionnaire-9 in spinal cord injury: an examination of factor structure as related to gender. J Spinal Cord Med. 2009. https://doi.org/10.1080/10790268.2009.11760766.
21. Keding A, Böhnke JR, Croudace TJ, Richmond SJ, MacPherson H. Validity of single item responses to short message service texts to monitor depression: an mHealth sub-study of the UK ACUDep trial. BMC Med Res Methodol. 2015. https://doi.org/10.1186/s12874-015-0054-6.
22. Keum BT, Miller MJ, Inkelas KK. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol Assess. 2018. https://doi.org/10.1037/pas0000550.
23. Kim YE, Lee B. The psychometric properties of the patient health questionnaire-9 in a sample of Korean university students. Psychiatry Investig. 2019. https://doi.org/10.30773/pi.2019.0226.
24. Lopez MB, Goldstein LH, Robinson EJ, Vitoratou S, Chalder T, Carson A, et al. Validation of the PHQ-9 in adults with dissociative seizures. J Psychosom Res. 2021. https://doi.org/10.1016/j.jpsychores.2021.110487.
25. Moreno-Agostino D, Chua KC, Peters TJ, Scazufca M, Araya R. Psychometric properties of the PHQ-9 measure of depression among Brazilian older adults. Aging Ment Health. 2022. https://doi.org/10.1080/13607863.2021.1963951.
26. Schuler M, Strohmayer M, Mühlig S, Schwaighofer B, Wittmann M, Faller H, et al. Assessment of depression before and after inpatient rehabilitation in COPD patients: psychometric properties of the German version of the Patient Health Questionnaire (PHQ-9/PHQ-2). J Affect Disord. 2018. https://doi.org/10.1016/j.jad.2018.02.037.
27. Shaff J, Kahn G, Wilcox HC. An examination of the psychometric properties of the Patient Health Questionnaire-9 (PHQ-9) in a multiracial/ethnic population in the United States. Front Psychiatry. 2024. https://doi.org/10.3389/fpsyt.2023.1290736.
28. Villarreal-Zegarra D, Copez-Lonzoy A, Bernabé-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. https://doi.org/10.1371/journal.pone.0221717.
29. Yu X, Tam WWS, Wong PTK, Lam TH, Stewart SM. The patient health questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatry. 2012. https://doi.org/10.1016/j.comppsych.2010.11.002.
30. Beard C, Hsu K, Rifkin LS, Busch AB, Björgvinsson T. Validation of the PHQ-9 in a psychiatric sample. J Affect Disord. 2016. https://doi.org/10.1016/j.jad.2015.12.075.
31. Chilcot J, Rayner L, Lee W, Price A, Goodwin L, Monroe B, et al. The factor structure of the PHQ-9 in palliative care. J Psychosom Res. 2013. https://doi.org/10.1016/j.jpsychores.2012.12.012.
32. Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, et al. The factor structure of major depression symptoms: a test of four competing models using the patient health questionnaire-9. Psychiatry Res. 2012. https://doi.org/10.1016/j.psychres.2012.05.018.
33. Granillo MT. Structure and function of the patient health questionnaire-9 among Latina and non-Latina white female college students. J Soc Soc Work Res. 2012. https://doi.org/10.5243/jsswr.2012.6.
34. Hall BJ, Patel A, Lao L, Liem A, Mayawati EH, Tjipto S. Structural validation of the Patient Health Questionnaire-9 (PHQ-9) among Filipina and Indonesian female migrant domestic workers in Macao: STRUCTURAL VALIDATION OF PHQ-9. Psychiatry Res. 2021. https://doi.org/10.1016/j.psychres.2020.113575.
35. Janssen EPCJ, Köhler S, Stehouwer CDA, Schaper NC, Dagnelie PC, Sep SJS, et al. The Patient Health Questionnaire-9 as a screening tool for depression in individuals with type 2 diabetes mellitus: the Maastricht study. J Am Geriatr Soc. 2016. https://doi.org/10.1111/jgs.14388.
36. Krause JS, Reed KS, McArdle JJ. Factor structure and predictive validity of somatic and nonsomatic symptoms from the patient health questionnaire-9: a longitudinal study after spinal cord injury. Arch Phys Med Rehabil. 2010. https://doi.org/10.1016/j.apmr.2010.04.015.
37. Krause JS, Saunders LL, Bombardier C, Kalpakjian C. Confirmatory factor analysis of the patient health questionnaire-9: a study of the participants from the spinal cord injury model systems. PM&R. 2011. https://doi.org/10.1016/j.pmrj.2011.03.003.
38. 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. Depress Anxiety. 2019. https://doi.org/10.1002/da.22940.
39. Petersen JJ, Paulitsch MA, Hartig J, Mergenthal K, Gerlach FM, Gensichen J. Factor structure and measurement invariance of the patient health questionnaire-9 for female and male primary care patients with major depression in Germany. J Affect Disord. 2015. https://doi.org/10.1016/j.jad.2014.08.053.
40. Richardson E, Richards J. Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. Rehabil Psychol. 2008. https://doi.org/10.1037/0090-5550.53.2.243.
41. Marcos-Nájera R, Le HN, Rodríguez-Muñoz MF, Olivares Crespo ME, Izquierdo Mendez N. The structure of the patient health questionnaire-9 in pregnant women in Spain. Midwifery. 2018. https://doi.org/10.1016/j.midw.2018.03.011.
42. 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. https://doi.org/10.1016/j.jad.2020.06.066.
43. Chae D, Lee J, Lee EH. Internal structure of the patient health questionnaire-9: a systematic review and meta-analysis. Asian Nurs Res. 2025. https://doi.org/10.1016/j.anr.2024.12.005.
44. De Man J, Absetz P, Sathish T, Desloge A, Haregu T, Oldenburg B, et al. Are the PHQ-9 and GAD-7 suitable for use in India? A psychometric analysis. Front Psychol. 2021. https://doi.org/10.3389/fpsyg.2021.676398.
45. Hart C, Draper CE, Soepnel LM, Godongwana M, Mabetha K, Nyati LH, et al. Examining the psychometric properties of the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 among young urban South African women. J Affect Disord. 2025. https://doi.org/10.1016/j.jad.2024.09.145.
46. Ahmadi A, Galusha JM, Ponder WN, Carbajal J, Schuman DL, Whitworth J, et al. Validation of the PCL-5, PHQ-9, and GAD-7 in a sample of first responders. J Occup Environ Med. 2023. https://doi.org/10.1097/JOM.0000000000002823.
47. Stanyte A, Fineberg NA, Podlipskyte A, Gecaite-Stonciene J, Macijauskiene J, Steibliene V, et al. Validation of the patient health questionnaire-9 and the generalized anxiety disorder-7 in Lithuanian individuals with anxiety and mood disorders. J Psychiatr Res. 2023. https://doi.org/10.1016/j.jpsychires.2023.06.027.
48. Rahman MA, Dhira TA, Sarker AR, Mehareen J. Validity and reliability of the patient health questionnaire scale (PHQ-9) among university students of Bangladesh. PLoS ONE. 2022. https://doi.org/10.1371/journal.pone.0269634.
49. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J Affect Disord. 2015. https://doi.org/10.1016/j.jad.2014.10.010.
50. DiStefano C, Motl RW. Further investigating method effects associated with negatively worded items on self-report surveys. Struct Equ Modeling. 2006. https://doi.org/10.1207/s15328007sem1303_6.
51. Schmalbach B, Zenger M, Michaelides MP, Schermelleh-Engel K, Hinz A, Körner A, et al. From bi-dimensionality to uni-dimensionality in self-report questionnaires. Eur J Psychol Assess. 2021. https://doi.org/10.1027/1015-5759/a000583.
52. Kam C, Meyer JP. Do optimism and pessimism have different relationships with personality dimensions? A re-examination. Pers Individ Dif. 2012. https://doi.org/10.1016/j.paid.2011.09.011.
53. Marsh HW. Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? J Pers Soc Psychol. 1996. https://doi.org/10.1037/0022-3514.70.4.810.
54. Peterson NA, Speer PW, Hughey J. Measuring sense of community: a methodological interpretation of the factor structure debate. J Community Psychol. 2006. https://doi.org/10.1002/jcop.20109.
55. Vautier S, Pohl S. Do balanced scales assess bipolar constructs? The case of the STAI scales. Psychol Assess. 2009. https://doi.org/10.1037/a0015312.

56. Tourangeau R, Rips LJ, Rasinski K. The psychology of survey responses. Cambridge: Cambridge University Press; 2000.
57. Aromaa E, Tolvanen A, Tuulari J, Wahlbeck K. Predictors of stigmatizing attitudes towards people with mental disorders in a general population in Finland. Nord J Psychiatry. 2011. https://doi.org/10.3109/08039488.2010.510206.
58. Wood L, Birtel M, Alsawy S, Pyle M, Morrison A. Public perceptions of stigma towards people with schizophrenia, depression, and anxiety. Psychiatry Res. 2014. https://doi.org/10.1016/j.psychres.2014.07.012.
59. Yokoya S, Maeno T, Sakamoto N, Goto R, Maeno T. A brief survey of public knowledge and stigma towards depression. J Clin Med Res. 2018. https://doi.org/10.14740/jocmr3282w.
60. Gaudiano BA, Miller IW. Self-stigma and attitudes about treatment in depressed patients in a hospital setting. Int J Soc Psychiatry. 2013. https://doi.org/10.1177/0020764012446404.
61. Yen CF, Chen CC, Lee Y, Tang TC, Yen JY, Ko CH. Self-stigma and its correlates among outpatients with depressive disorders. Psychiatr Serv. 2005. https://doi.org/10.1176/appi.ps.56.5.599.
62. Pyne JM, Kuc EJ, Schroeder PJ, Fortney JC, Edlund M, Sullivan G. Relationship between perceived stigma and depression severity. J Nerv Ment Dis. 2004. https://doi.org/10.1097/01.nmd.0000120886.39886.a3.
63. Sirey JA, Bruce ML, Alexopoulos GS, Perlick DA, Raue P, Friedman SJ, et al. Perceived stigma as a predictor of treatment discontinuation in young and older outpatients with depression. Am J Psychiatry. 2001. https://doi.org/10.1176/appi.ajp.158.3.479.
64. Hunt M, Auriemma J, Cashaw ACA. Self-report bias and underreporting of depression on the BDI-II. J Pers Assess. 2003. https://doi.org/10.1207/S15327752JPA8001_10.
65. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull. 2007. https://doi.org/10.1037/0033-2909.133.5.859.
66. Robinson J, Khan N, Fusco L, Malpass A, Lewis G, Dowrick C. Why are there discrepancies between depressed patients' global rating of change and scores on the patient health questionnaire depression module? A qualitative study of primary care in England. BMJ Open. 2017. https://doi.org/10.1136/bmjopen-2016-014519.
67. Cross DS, Wallace R, Cross J, Coimbra MF. Understanding pilots' perceptions of mental health issues: a qualitative phenomenological investigation among airline pilots in the United States. Cureus. 2024. https://doi.org/10.7759/cureus.66277.
68. Marshall RE, Milligan-Saville J, Petrie K, Bryant RA, Mitchell PB, Harvey SB. Mental health screening amongst police officers: factors associated with under-reporting of symptoms. BMC Psychiatry. 2021. https://doi.org/10.1186/s12888-021-03125-1.
69. Hudson E, Arnaert A, Lavoie-Tremblay M. Healthcare professional disclosure of mental illness in the workplace: a rapid scoping review. J Ment Health. 2021. https://doi.org/10.1080/09638237.2021.1979485.
70. Kristófersdóttir KH, Kristjánsdóttir H, Asgeirsdottir RL, Karlsson T, Vésteinsdóttir V, Thorsdottir F. Investigating the PHQ-9 with Mokken scale analysis and cognitive interviews. Assessment. 2023. https://doi.org/10.1177/10731911231216961.
71. Malpass A, Shaw A, Kessler D, Sharp D. Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study. Br J Gen Pract. 2010. https://doi.org/10.3399/bjgp10X502119.
72. Velloza J, Njoroge J, Ngure K, Thuo N, Kiptinness C, Momanyi R, et al. Cognitive testing of the PHQ-9 for depression screening among pregnant and postpartum women in Kenya. BMC Psychiatry. 2020. https://doi.org/10.1186/s12888-020-2435-6.
73. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh postnatal depression scale. Br J Psychiatry. 1987. https://doi.org/10.1192/bjp.150.6.782.
74. Davison ML, Zieffler A, Cabrera J, Karl SR, Cohen HS. Automated assembly of optimally spaced and balanced paired comparisons: controlling order effects. Behav Res Methods. 2012. https://doi.org/10.3758/s13428-011-0170-0.
75. Ross RT. Optimum orders for the presentation of pairs in the method of paired comparisons. J Educ Psychol. 1934. https://doi.org/10.1037/h0070754.
76. R Core Team. R: a language and environment for statistical computing, v4.4.0. Vienna, Austria: R Foundation for Statistical Computing. 2024. Available from: https://www.R-project.org/
77. Wickelmaier F, Schmid C. A matlab function to estimate choice model parameters from paired-comparison data. Behav Res Methods Instrum Comput. 2004. https://doi.org/10.3758/BF03195547.
78. Kendall MG. Rank correlation methods. 4th ed. London: Charles Griffin & Company LTD; 1975.
79. Kendall MG, Smith BB. On the method of paired comparisons. Biometrika. 1940. https://doi.org/10.2307/2332613.
80. Edwards AL. Techniques of attitude scale construction. New York: Appleton-Century-Crofts; 1957.
81. Guilford JP. Psychometric methods. New York: McGraw-Hill; 1954.
82. Thurstone LL. The method of paired comparisons for social values. J Abnorm Soc Psychol. 1927. https://doi.org/10.1037/h0065439.
83. Torgerson WS. Theory and methods of scaling. New York: Wiley; 1958.
84. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
85. Wild BS, Cabral RM. The scaling of paired comparison data with one or more extreme stimuli: a comparison of three models. Educ Psychol Meas. 1976. https://doi.org/10.1177/001316447603600311.
86. Paulhus DL. Measurement and control of response bias. In: Robinson JP, Shaver PR, Wrightsman LS, editors. Measures of personality and social psychological attitudes. San Diego, CA: Academic Press; 1991. p. 17–59.
87. ZagerKocjan G, Lavtar D, Sočan G. The effects of survey mode on self-reported psychological functioning: measurement invariance and latent mean comparison across face-to-face and web modes. Behav Res Methods. 2023. https://doi.org/10.3758/s13428-022-01867-8.
88. Krendl AC, Pescosolido BA. Countries and cultural differences in the stigma of mental illness: the East-West divide. J Cross-Cult Psychol. 2020. https://doi.org/10.1177/0022022119901297.
89. Abdelnour E, Jansen MO, Gold JA. ADHD diagnostic trends: increased recognition or overdiagnosis? Mo Med. 2022;119:467–73.
90. Adler M, Hetta J, Isacsson G, Brodin U. An item response theory evaluation of three depression assessment instruments in a clinical sample. BMC Med Res Methodol. 2012. https://doi.org/10.1186/1471-2288-12-84.
91. Bianchi R, Verkuilen J, Toker S, Schonfeld IS, Gerber M, Brähler E, et al. Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. Psychol Assess. 2022. https://doi.org/10.1037/pas0001124.
92. Boothroyd L, Dagnan D, Muncer S. Psychometric analysis of the Generalized Anxiety Disorder scale and the Patient Health Questionnaire using Mokken scaling and confirmatory factor analysis. Health Prim Car. 2018. https://doi.org/10.15761/HPC.1000145.
93. Flores-Kanter PE, Mosquera M. How do you behave as a psychometrician? Research conduct in the context of psychometric research. Span J Psychol. 2023. https://doi.org/10.1017/SJP.2023.14.
94. Perinelli E, Gremigni P. Use of social desirability scales in clinical psychology: a systematic review. J Clin Psychol. 2016. https://doi.org/10.1002/jclp.22284.
95. Vésteinsdóttir V, Reips UD, Joinson A, Thorsdottir F. An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic internet panel data and cognitive interviews. Pers Individ Dif. 2017. https://doi.org/10.1016/j.paid.2016.11.023.
96. Fowler FJ. Improving survey questions: design and evaluation. Thousand Oaks, CA: SAGE Publications; 1995.
97. Charles JLK, V. Dattalo P. Minimizing social desirability bias in measuring sensitive topics: the use of forgiving language in item development. J Soc Serv Res. 2018; https://doi.org/10.1080/01488376.2018.1479335
98. Näher AF, Krumpal I. Asking sensitive questions: the impact of forgiving wording and question context on social desirability bias. Qual Quant. 2012. https://doi.org/10.1007/s11135-011-9469-2.
99. Yan T. Consequences of asking sensitive questions in surveys. Annu Rev Stat Appl. 2021. https://doi.org/10.1146/annurev-statistics-040720-033353.
100. Abdullah T, Brown TL. Mental illness stigma and ethnocultural beliefs, values, and norms: an integrative review. Clin Psychol Rev. 2011. https://doi.org/10.1016/j.cpr.2011.05.003.
101. Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Qual Quant. 2013. https://doi.org/10.1007/s11135-011-9640-9.
102. Holbrook AL, Green MC, Krosnick JA. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. Public Opin Q. 2003. https://doi.org/10.1086/346010.
103. Kreuter F, Presser S, Tourangeau R. Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. Public Opin Q. 2008. https://doi.org/10.1093/poq/nfn063.

Publisher's Note