

Lost in translation? Not anymore: a Mokken scale analysis of the revised Icelandic translation of the PHQ-9

Laufey Ósk Jóns, Kristín Hulda Kristófersdóttir, Vaka Vésteinsdóttir, Hafrún Kristjánsdóttir, Thorlakur Karlsson & Fanney Thorsdóttir

To cite this article: Laufey Ósk Jóns, Kristín Hulda Kristófersdóttir, Vaka Vésteinsdóttir, Hafrún Kristjánsdóttir, Thorlakur Karlsson & Fanney Thorsdóttir (20 Jan 2026): Lost in translation? Not anymore: a Mokken scale analysis of the revised Icelandic translation of the PHQ-9, *Nordic Psychology*, DOI: [10.1080/19012276.2026.2615958](https://doi.org/10.1080/19012276.2026.2615958)

To link to this article: <https://doi.org/10.1080/19012276.2026.2615958>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 20 Jan 2026.



[Submit your article to this journal](#)



Article views: 551



[View related articles](#)



[View Crossmark data](#)

Lost in translation? Not anymore: a Mokken scale analysis of the revised Icelandic translation of the PHQ-9

LAUFEY ÓSK JÓNS¹, KRISTÍN HULDA KRISTÓFERSDÓTTIR¹, VAKA VÉSTEINSDÓTTIR¹,
HAFRÚN KRISTJÁNSDÓTTIR², THORLAKUR KARLSSON³ & FANNEY THORSÐÓTTIR¹

Correspondence address: Kristín Hulda Kristófersdóttir Faculty of Psychology, School of Health Sciences, University of Iceland, Sæmundargata 12, Reykjavík 102, Iceland. Email: kristinhk@hi.is

Abstract

The Patient Health Questionnaire-9 (PHQ-9) is among the most widely used self-report instruments for assessing depressive symptoms. Although previously translated into Icelandic and widely adopted in clinical and research contexts, recent studies identified notable translation issues that may have affected its psychometric quality. This study examined a revised Icelandic version of the PHQ-9, designed to improve clarity and fidelity to the original English version. A general population sample ($n = 139$) completed the revised version in an online survey. Mokken scale analysis was conducted to evaluate scalability, monotonicity, local independence, and invariant item ordering. Results indicated that the revised version demonstrated stronger psychometric properties compared to the original translation. The overall scalability coefficient suggested a strong scale ($H = .548$) according to Mokken's criteria; however, the confidence interval (95% CI [.466, .630]) crossed .50, indicating that strong scalability cannot be guaranteed at the population level. All items surpassed the conventional .30 benchmark, with most achieving $H_i \geq .50$. No significant violations of monotonicity were observed, and only minor indications of negative local dependence emerged. While partial evidence for invariant item ordering was found, it remained limited. These findings suggest that the revised Icelandic PHQ-9 is a more internally consistent and interpretable assessment of depression severity, broadly comparable to findings reported in other language versions. The scale appears suitable for use in research settings and for screening purposes in Iceland, although further validation in clinical populations is needed.

Keywords: PHQ-9, depression assessment, Mokken scale analysis, translation validation, psychometric evaluation

Depression is a major public health concern with profound individual and societal consequences. According to the World Health Organization (2017, 2023), depression affects an estimated 280 million people worldwide and is a leading cause of disability. A meta-analysis

¹Faculty of Psychology, University of Iceland, Reykjavík, Iceland;

²Department of Sport Science, Reykjavík University, Reykjavík, Iceland;

³Brandr, Reykjavík, Iceland

Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19012276.2026.2615958>.

of community-based studies across 30 countries reported an aggregate point prevalence of 12.9%, a one-year prevalence of 7.2%, and a lifetime prevalence of 10.8% (Lim et al., 2018), highlighting the substantial proportion of the global population affected at any given time. Depression contributes significantly to the global burden of disease and imposes considerable economic costs due to reduced productivity and increased healthcare demands (OECD, 2021). These reports underscore the necessity of ongoing research into the causes, treatment, and prevention of depression.

Iceland offers favourable conditions for conducting depression research, thanks to its nationwide health and population registries, as well as the assignment of personal identity numbers to all residents, which enables linkage across datasets (Laugesen et al., 2021). This infrastructure presents a unique opportunity to generate evidence that informs prevention strategies, clinical interventions, and public health policy. Additionally, Iceland's small, well-defined, and relatively homogeneous population facilitates the identification of genetic and environmental risk factors for depression, which can later be validated in more diverse populations. These strengths make Iceland an ideal setting for innovative, high-quality research on depression. However, the accuracy of such research ultimately depends on how well depression is identified and assessed. If assessment tools (e.g. diagnostic interviews, questionnaires, or clinical scales) lack validity or reliability, their scores may lead to misclassifications, biased research findings, or misleading conclusions.

The Patient Health Questionnaire-9

The Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001) is a widely used self-report instrument for assessing depression globally and within the Icelandic healthcare system. Developed as part of the broader Patient Health Questionnaire (PHQ; Spitzer et al., 1999), it was designed as a brief but effective measure that closely aligns with the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV; American Psychiatric Association, 1994) criteria for major depressive disorder (MDD). The PHQ-9 has demonstrated good psychometric properties, clinical utility, and ease of use across diverse settings (Chae et al., 2025; Inoue et al., 2012; Kroenke, 2021). Its simplicity allows for rapid administration and scoring, making it practical for clinical practice, while also supporting large-scale epidemiological research on depression prevalence and public health impact. Reflecting its utility, it has been recommended by leading organizations, including the American Psychiatric Association (Silverman et al., 2015), the National Institute for Health and Care Excellence (2011), and the U.S. Department of Veterans Affairs and Department of Defense (Management of Major Depressive Disorder Working Group, 2016).

Kroenke et al. (2001) conducted the initial validation study using primary care and obstetrics-gynaecology clinic patients, demonstrating strong reliability and validity. The PHQ-9 exhibited high internal consistency (Cronbach's $\alpha = .86$ and $.89$) and test-retest reliability ($r = .84$).¹ Criterion validity was assessed by comparing PHQ-9 scores with diagnoses made by mental health professionals, with a sensitivity and specificity of 88% for detecting major depression at a cutoff score of 10 or higher.

The unidimensionality of the PHQ-9 has generally been supported in the literature (Chae et al., 2025). Although some studies suggest the presence of two distinct factors, a Cognitive-Affective factor and a Somatic factor (e.g. Beard et al., 2016; Patel et al., 2019),

these factors are typically highly correlated, raising doubts about their conceptual distinctiveness. A systematic review by Lamela et al. (2020) found evidence supporting both one- and two-factor models, although the single-factor structure was more frequently endorsed. Furthermore, Bianchi et al. (2022), in a large-scale psychometric study, concluded with “an unprecedented level of confidence” (p. 595) that the PHQ-9 functions as a unidimensional measure.

Across multiple studies, Mokken scale analyses (MSAs) have generally found that the PHQ-9 demonstrates acceptable to strong scalability at both the item and scale levels (Adler et al., 2012; Bianchi et al., 2022; Boothroyd et al., 2018; Cosci et al., 2024; Pretorius & Padmanabhanunni, 2024; Stochl et al., 2022). According to Mokken’s (1971) criteria, reported item scalability coefficients ($H_i = .39-.62$) and overall scale scalability ($H = .47-.55$) suggest the PHQ-9 total score can be used to accurately order people according to their level of depression severity.

Studies have also shown strong positive correlations between PHQ-9 scores and other established depression measures, such as the Beck Depression Inventory–II (Titov et al., 2011; Urtasun et al., 2019) and the Hamilton Depression Rating Scale (Ma et al., 2021), providing convergent evidence supporting the validity of PHQ-9 scores as indicators of depressive symptom severity.

The PHQ-9 has also proven to be an effective screening tool for depression. In a recent meta-analysis involving more than 44,000 participants across 100 studies, researchers found that the PHQ-9 demonstrated both sensitivity and specificity of 85% when using the standard cutoff score of ≥ 10 to detect major depression (Negeri et al., 2021).

These findings suggest that the PHQ-9 is a psychometrically sound instrument for assessing depressive symptoms across different populations and settings. However, its effectiveness relies on accurate translation and cultural adaptation to ensure that items retain their intended meaning. As such, evaluating the quality and measurement properties of local language versions is critical to maintaining the scale’s validity and utility.

The PHQ-9 in Iceland

In Iceland, the use of PHQ-9 is recommended in clinical guidelines by Landspítali—The National University Hospital (Agnarsdóttir et al., 2011) and is also supported in primary care, where it serves as a minimum standard for evaluating depressive symptoms (Agnarsdóttir et al., 2021). Beyond clinical applications, the PHQ-9 is widely utilised in research, both in Iceland and internationally (Kroenke, 2021). The original Icelandic translation has been incorporated into diverse studies across disciplines, including clinical psychology (e.g. Wang et al., 2024), medicine (e.g. Saevarsdóttir et al., 2021), sport psychology (e.g. Jónsdóttir et al., 2022; Tahtinen et al., 2021), and socio-economic research (e.g. Einarsdóttir et al., 2022).

Because the PHQ-9 is widely used in both clinical practice and research in Iceland, the Icelandic version must be accurate and reliable. Poor translation can undermine item performance and lead to biased or misleading results. Until recently, however, little attention had been given to the quality of the Icelandic translation.

A comprehensive evaluation by Kristófersdóttir et al. (2024) addressed this gap by applying MSA to a primary care sample ($N=618$) and conducting cognitive interviews with

individuals who both had and did not have a history of depression ($N=53$). The findings revealed several issues, including translation errors, that may have contributed to the poor performance of specific items and reduced the overall measurement quality of the scale.

In response to these concerns, a revised Icelandic translation of the PHQ-9 was developed by our research group, based on the issues identified in Kristófersdóttir et al. (2024), to improve linguistic clarity and conceptual equivalence with the original English version. The objective of the current study is to present this revised version and conduct its first psychometric evaluation in a general population sample using MSA.

Methods

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. Of the 295 individuals who opened the survey link, 17 did not respond to any items and were excluded, resulting in 278 valid responses.

For the purpose of a separate study, these 278 participants were randomly assigned to two groups: one received an updated translation of Item 7 as described in the Instrument section below, and the other received a version *without* the examples (*“reading the newspaper or watching television”*). In the present study, analyses are reported only for the subgroup who received the version *with* examples in Item 7 ($n=139$), as this format is closest to the English version most commonly used in practice. The potential impact of including or omitting the examples will be examined in a separate study that may inform future recommendations regarding item wording.

Of these 139 participants, 62 (44.6%) were men and 77 (55.4%) were women. No non-binary individuals participated in the survey. Participants' age ranged from 27 to 78 years, with a mean age of 56.3 ($SD=13.1$). Two participants did not report their age.

Instrument

The PHQ-9 (Kroenke et al., 2001) is a nine-item self-report instrument developed to screen for depression in primary care and to assess symptom severity. Respondents are asked to rate how often they have been bothered by the listed problems over the last two weeks. The items are rated on a 4-point scale (0 = *Not at all*, 1 = *Several days*, 2 = *More than half the days*, and 3 = *Nearly every day*). Total scores of 5, 10, 15, and 20 have been recommended as thresholds for mild, moderate, moderately severe, and severe depression, respectively (Kroenke et al., 2001). A score of 10 or above is commonly used as a cutoff for identifying probable cases of depression.

The scale was first translated from English to Icelandic in 2005 and has been used in both clinical practice and research since then. However, a recent study by Kristófersdóttir et al. (2024) identified several mistranslations and ambiguities in this version and recommended improvements to increase its accuracy. Building on these findings, the Icelandic PHQ-9 was revised to enhance linguistic clarity and fidelity to the original English version.

Table 1. An overview of the original PHQ-9 item wording in English, the original Icelandic translation, and the revised Icelandic translation.

Item	English by Kroenke et al. (2001)	Icelandic Translation: Original	Icelandic Translation: Revised
1*	Little interest or pleasure in doing things	Little interest or joy in doing things	Little interest or little pleasure in doing things
2	Feeling down, depressed, or hopeless	Felt down, depressed or hopeless	Felt down, depressed or hopeless
3*	Trouble falling or staying asleep, or sleeping too much	Had difficulty going to sleep or sleeping the whole night	Had difficulty going to sleep or sleeping the whole night—or slept too much
4*	Feeling tired or having little energy	Fatigue and lack of energy	Fatigue or lack of energy
5	Poor appetite or overeating	Loss of appetite or overeating	Loss of appetite or overeating
6*	Feeling bad about yourself—or that you are a failure or have let yourself or your family down	Feeling bad about yourself or felt that you have failed or not stood yourself in the piece toward yourself or your family	Been displeased with yourself—or felt like you are a failure or that you have let yourself or your family down
7*	Trouble concentrating on things, such as reading the newspaper or watching television	Trouble concentrating e.g. when reading the newspapers or watching television	Had difficulty concentrating on things, e.g. when reading the newspapers or watching television
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	Moving or speaking so slowly that others have noticed? Or the opposite – been so fidgety or restless that you moved a lot more than usual	Moving or speaking so slowly that other people could have noticed? Or the opposite—been so fidgety or restless that you moved a lot more than usual
9*	Thoughts that you would be better off dead or of hurting yourself in some way	Thought that it would be better that you were dead or thought about hurting yourself in some way	Thought that it would be better if you were dead or thought about hurting yourself in some way

Note. Icelandic translations are back-translations into English.

*Item was changed from the original Icelandic translation.

Table 1 presents the original and revised Icelandic items alongside their English counterparts. The Icelandic items shown are back-translations into English, initially produced by LÓJ and subsequently reviewed and revised by KHK and FT. As a result, the English wording may appear somewhat unnatural, though it accurately reflects appropriate and idiomatic Icelandic usage. The most notable changes are described below.

For Item 1, a minor revision was made: the original phrase *“Little interest or joy in doing things”* was updated to *“Little interest or little pleasure in doing things”*. This subtle change, informed by interview feedback from Kristófersdóttir et al. (2024), was intended to enhance clarity and better reflect the distinction between interest and pleasure.²

A more substantial issue was identified in Item 3, where the original translation failed to capture the full scope of sleep disturbances. Specifically, it omitted reference to

hypersomnia, only mentioning difficulty falling or staying asleep. This omission may have led to some cases of excessive sleep going unreported. The revised version corrects this by including both insomnia and hypersomnia, ensuring that the item more accurately reflects the relevant symptom criterion for major depression.

Item 4 contained a small but significant mistranslation in the original Icelandic version, where “and” was used instead of “or”, making the item more restrictive and potentially reducing endorsement of the symptom. The updated translation corrects this to better match the original English phrasing.

Item 6 was found to be weaker in severity compared to the English version due to a translation inconsistency (Kristófersdóttir et al., 2024). This discrepancy could have led to underreporting of depressive symptoms, reducing the scale’s ability to differentiate between varying levels of severity. Three key changes were made in the revised translation. First, “*Feeling bad about yourself*” was adjusted to use stronger wording in Icelandic. Second, “*felt that you had failed*” was revised to “*felt like you are a failure*”, making the phrasing both stronger and closer in meaning to the original English version. Lastly, “*not stood yourself in the piece*,” an Icelandic idiom rarely used in everyday speech, was replaced with a clearer expression of the intended meaning to improve comprehension.

Item 7 underwent minor adjustments to improve clarity and better capture the intended breadth of concentration difficulties. In the original Icelandic translation, the phrase corresponding to “*on things*” was omitted, which might have led some respondents to interpret the examples (reading the newspaper or watching television) too literally (Kristófersdóttir et al., 2024). The revised version adds this element to indicate that difficulties can occur across a wider range of activities, while retaining the original examples for context and equivalence to the English version.

The wording of Item 8 was also problematic in the original translation, as it implied that others must have actually noticed changes in movement or speech. Some interviewees in Kristófersdóttir et al. (2024) even interpreted this as requiring that someone comment on the changes, which could lead to underreporting. To address this, the revised version now refers to whether others *could* have noticed such changes in behaviour, a more inclusive phrasing that still captures the symptom’s observable nature without requiring external confirmation.

Only a small grammatical adjustment was made to Item 9, replacing “*that you were dead*” with “*if you were dead*” to improve clarity and the naturalness of expression in Icelandic.

Items 2 and 5 were left unchanged in the revised version of the scale as they did not pose any major translational difficulty.

Kristófersdóttir et al. (2024) also highlighted a translation error in the third response option, which had been rendered as “*More than half the time*” in Icelandic but should have been “*More than half the days*” to be consistent with the original English version. This has been corrected in the revised scale.

Finally, because Icelandic is a gendered language with three grammatical genders, three distinct versions of the scale were created: one for men, one for women, and one for non-binary individuals. The original translation used a single version that included both masculine and feminine grammatical endings within each item, which made them overly complex and challenging to read while excluding non-binary individuals. By creating separate gendered versions, the revised translation improves clarity and inclusivity for all respondents.

Procedure

The study protocol received a positive review from the Science Ethics Committee of Icelandic Universities (ID: SHV2024-073). Additional review was sought from the National Bioethics Committee in Iceland, which determined that formal approval was not required.

The PHQ-9 was administered online using the survey platform QuestionPro in November 2024. The survey remained open for two weeks, and two reminder emails were sent to individuals who had not responded. Upon accessing the survey, participants were presented with an information and consent form. At the bottom of the page, they were required to indicate whether they had read the text and consented to participate. Only those who provided consent could proceed to the next page; the survey was terminated for those who declined.

On the following page, participants were asked to report their birth year and gender. The gender question offered five response options: "Man," "Woman," "Non-binary," "I define my gender differently," and "Prefer not to say." If a participant selected either of the last two options, an additional question appeared asking how they would prefer to be addressed grammatically in Icelandic.

After the demographic questions, participants completed the revised Icelandic version of the PHQ-9. This version was tailored to each respondent's grammatical gender based on their previous answers. The different grammatical versions can be seen in the [supplementary material](#). For the purpose of another study, participants also completed the Marlowe-Crowne Social Desirability-Short Form (MCSD-SF; Vésteinsdóttir et al., 2017), but these data are not used in the current analysis. At the end of the survey, participants were provided with contact information in case they experienced any emotional discomfort or had questions about the study. Most participants completed the survey within five minutes.

Data analysis

All data analyses were conducted in R (version 4.4.1; R Core Team, 2024). The R package *mokken* (Van der Ark, 2007, 2012) was used to perform the MSA and the *robustbase* package (Maechler et al., 2024) was used to compute the statistics needed for the adjusted boxplot. Participants who did not respond to any of the PHQ-9 items ($n = 17$) were removed from the sample. Among the remaining participants, there was no item-level missing data.

MSA (Mokken, 1971) is a nonparametric item response theory (IRT) method that provides an appropriate first step in evaluating the psychometric quality of a scale. Unlike parametric IRT models, MSA does not assume a specific functional form for the relation between the latent trait and item responses. This property makes MSA especially valuable when parametric assumptions are difficult to justify or when sample sizes are modest.

Within MSA, two nested models are commonly evaluated: the monotone homogeneity model (MHM) and the double monotonicity model (DMM) (Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017). The MHM is defined by three assumptions: unidimensionality, local independence, and monotonicity. When these are met, the total score can be interpreted as an ordinal indicator of a respondent's level on the latent trait. In clinical settings, this supports interpreting higher total scores on the PHQ-9 as reflecting higher levels of depression severity. The DMM builds on the MHM by also requiring that item response functions do not intersect. This additional condition, known as invariant item ordering (IIO),

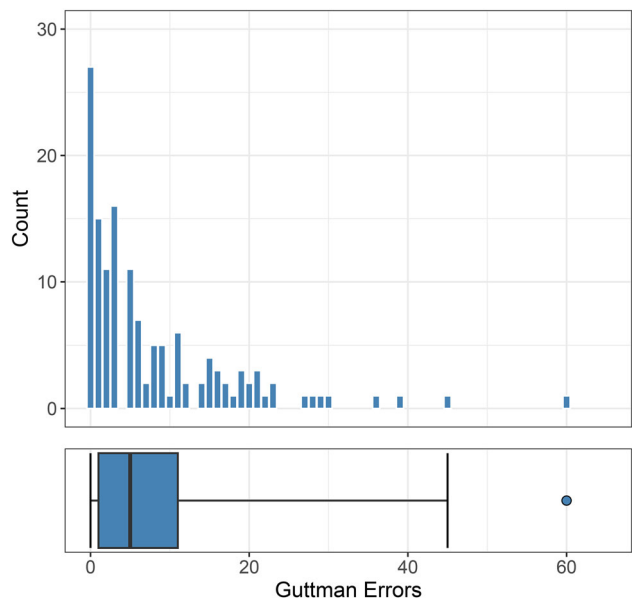


Figure 1. The distribution of the number of Guttman errors in the sample ($n = 139$).

implies that items’ relative difficulty or severity remains consistent across all levels of the latent trait, allowing for stronger interpretive claims about the hierarchy of symptoms.

If a clinical scale like the PHQ-9 fits the MHM, it supports using the total score as a meaningful indicator of depression severity, reinforcing current clinical practices that rely on cut-off scores and symptom counts. If the scale also fits the DMM, it further suggests that symptoms emerge in a predictable order as depression severity increases. For example, an item assessing a symptom such as “fatigue” would generally be endorsed before more severe symptoms like “suicidal ideation.” This hierarchy can offer clinicians additional insights into symptom progression and improve the interpretability of total scores across patients. However, even if the DMM is not satisfied, the validity of the total score to order respondents along the depression severity dimension remains intact if the MHM holds.

Before assessing the MSA assumptions, item-score patterns were checked for outliers, following the procedures recommended by [Sijtsma and Van der Ark \(2017\)](#). Given the highly skewed distribution of Guttman errors (GEs), an adjusted boxplot was used to identify extreme cases (see [Figure 1](#)). GEs occur when respondents exhibit inconsistent response patterns, that is, endorsing an item reflecting a higher level of the trait while not endorsing an item reflecting a lower level. One participant was identified as an outlier, as it was the only case exceeding the upper fence of 57 GEs in the adjusted boxplot. To evaluate its influence, analyses were conducted both with and without this participant. Given that the outlier notably affected the scalability coefficient for Item 8 and slightly increased the scalability of the whole scale, the decision was made to exclude this participant, resulting in a final sample size of $n = 138$.

Scalability

The scalability of the PHQ-9 was evaluated by calculating scalability coefficients for each item pair (H_{ij}), each item (H_i), and the overall scale (H) ([Sijtsma & Molenaar, 2002](#)). Scalability

coefficients indicate how closely the set of items approximates a Guttman scale, meaning a perfectly ordered scale without GEs. If no GEs are present, all coefficients (H_{ij} , H_i , and H) are equal to 1, indicating perfect unidimensionality. As the number of GEs increases at the item-pair, item, or scale level, the corresponding coefficients decrease, reflecting weaker ordering of persons and reduced scale coherence.

The H_{ij} coefficient is computed by comparing the observed number of GEs between two items to the number expected under the assumption of item independence (Molenaar, 1991; Sijtsma & Molenaar, 2002). Specifically, H_{ij} is calculated as one minus the ratio of observed to expected GEs. This value reflects how well two items cohere in their ordering and can be considered the proportion of their maximum possible association given their marginal distributions.

The H_i coefficient is calculated similarly, but it summarises the performance of a single item in relation to all others in the scale (Sijtsma & Molenaar, 2002). It is computed by aggregating the GEs involving that item across all item pairs, dividing this total by the expected number of GEs under the assumption of independence, and subtracting the resulting ratio from one. H_i serves as a nonparametric indicator of item discrimination. A higher H_i value suggests that the item strongly relates to the latent trait being measured and effectively distinguishes between individuals with lower and higher levels of it. Conversely, a H_i coefficient near zero indicates that the item contributes very little to a reliable person ordering, as individuals across all levels of the latent trait are equally likely to endorse it. Conventionally, a H_i value of at least .30 is considered the minimum acceptable threshold for item scalability (Mokken, 1971; Sijtsma & Van der Ark, 2017).

The H coefficient summarises the scale's ability to explain a single latent trait, such that ordering respondents by their total score is meaningful (Sijtsma & Molenaar, 2002). It is computed by subtracting from one the ratio of the total number of observed GEs across all items to the number expected under the assumption of independence. In simpler terms, H can be interpreted as the average of the H_i values.

Interpretation of scalability followed Mokken's (1971) guidelines: values of $.30 \leq H < .40$ indicate a weak scale, $.40 \leq H < .50$ a medium-strength scale, and $H \geq .50$ a strong scale. This classification provides a standardized method for assessing scale quality and comparing different instruments or translations. However, because the PHQ-9 is used in clinical settings to inform important decisions about individuals, a minimum acceptable H value of .50 was adopted. Accordingly, a threshold of .50 was also applied to each H_i coefficient instead of the traditional .30 benchmark, reflecting the importance of robust item performance for clinical applications. H_{ij} values were reviewed to confirm they were non-negative.

The validity of the scalability coefficients depends on three core assumptions of the Mokken model: unidimensionality, monotonicity, and local independence (Sijtsma & Molenaar, 2002). These assumptions were evaluated using the diagnostic procedures described in the following sections.

Automated Item Selection Procedure (AISP)

To examine the dimensionality of the PHQ-9, the Automated Item Selection Procedure (AISP; Sijtsma & Molenaar, 2002) was applied. AISP is a bottom-up clustering algorithm that assigns items to one or more scales based on two criteria: a) $H_{ij} > 0$ and b) $H_i \geq c > 0$,

where c is a user-defined lower-bound value. Following recommendations by Hemker et al. (1995), the procedure was run repeatedly across a range of lower-bound values ($c = 0$ to $.55$ in increments of $.05$) to examine the stability of scale membership and to identify the dimensionality pattern typical of unidimensional versus multidimensional item pools. Because the PHQ-9 constitutes an a priori defined scale, AISP was used here as a descriptive check rather than as a data-driven scale construction method.

Monotonicity

Monotonicity is the assumption that the probability of endorsing an item increases or remains constant as the level of the latent trait increases. It was evaluated using the *check-monotonicity* function in the *mokken* package (Van der Ark, 2007). This function assesses whether the cumulative probability of an item score exceeding a given threshold is a non-decreasing function of the respondent's rest score (i.e. the total score excluding the item in question). A violation occurs when this cumulative probability decreases as the rest score increases.

Rest-score groups were formed by aggregating respondents with the same or similar rest scores. In samples smaller than 250, Molenaar and Sijtsma (2000) recommend a minimum group size of $n/3$, with at least 50 respondents per group. However, it is also advised to evaluate monotonicity several times using different minimum group sizes (Van der Ark, 2007). Smaller rest-score groups can improve sensitivity to detect violations but may increase sampling variability, while larger groups reduce variability but may miss smaller violations. After considering these trade-offs, a minimum group size of 40 was selected to balance granularity and stability in the estimated response functions. This choice resulted in three rest-score groups being formed, compared to only two groups when using the default minimum size of 50.

Potential violations were examined both statistically and visually. The mean item response functions (IRFs) and item step response functions (ISRFs) were plotted for each item. For polytomous items, the mean IRF shows the average item score across rest-score groups and can reveal non-monotonic trends.

Statistically significant deviations were evaluated using a z-test comparing the proportion endorsing each item step across adjacent rest-score groups. The null hypothesis assumed equal proportions, while the alternative posited a decrease in endorsement probability in the higher group. Violations were flagged if they exceeded the default minimum threshold of 0.09 (i.e. the number of item steps multiplied by 0.03). The output included a *Crit* value for each item, which is a weighted summary of the item's H_i coefficient, the number and size of violations, and their statistical significance. Following Molenaar and Sijtsma (2000), *Crit* values below 40 were considered likely to reflect sampling variation, whereas values above 80 indicated substantial violations and potential misfit with the model.

Local independence

To assess whether item scores were independent after accounting for respondents' levels of depression (i.e. whether they satisfied the assumption of local independence), the conditional association procedure was applied (Straat et al., 2016). Violations of local

independence can take two forms: positive dependence, where items are more strongly related than expected given the latent trait, and negative dependence, where they are less strongly related. Consistent with the rest-score grouping applied in the monotonicity checks, a minimum rest-score group size of 40 was used for the local independence analyses.

The procedure evaluates these forms of dependence using two indices: W_1 and W_3 . The W_1 index measures positive local dependence, with higher values indicating that a pair of items is positively dependent beyond the latent trait. Conversely, the W_3 index assesses negative local dependence, where higher values suggest that a pair of items is negatively dependent. Together, these indices help identify item pairs that potentially violate the assumption of local independence (Straat et al., 2016).

Ordering structure

The ordering structure of the PHQ-9 items, reflecting the additional assumption of invariant item ordering (IIO) required by the DMM, was evaluated using the manifest invariant item ordering (MIIO) method as implemented in the *check.iio* function of the *mokken* R package (Van der Ark, 2007, 2012).

While the items are not necessarily expected to follow a strict ordering by difficulty or severity, the interpretability of sum scores is enhanced if items can be meaningfully arranged from less to more severe manifestations of depression. Recent methodological work by Koopman and Braeken (2025) has emphasized that even when invariant ordering of individual items is not supported, it can still be informative to assess whether *clusters* of items are consistently ordered on average. To this end, items were grouped into three clusters based on their observed mean scores, reflecting the assumption that more frequently endorsed items represent less severe symptoms. This resulted in the following clusters, ordered from higher to lower mean scores: Cluster A (Items 3 and 4), Cluster B (Items 1, 2, 5, 6, and 7), and Cluster C (Items 8 and 9). For consistency with the other assumption checks, the minimum rest-score group size was set to 40 respondents.

The analysis tested three levels of ordering: manifest weak invariant cluster ordering (ICO), manifest strong ICO, and manifest invariant item ordering (IIO). Weak ICO evaluates whether clusters maintain the expected rank order in terms of average difficulty across levels of the latent trait. Strong ICO further requires that all items in a lower cluster are easier than all items in a higher cluster, and IIO additionally requires invariant ordering of items within clusters. Violations were identified based on significant differences in expected item or cluster scores across rest-score groups.

Global ordering consistency was quantified using H^T (total ordering consistency), H^{TB} (consistency attributable to clusters), and the H^{TB}/H^T ratio, which reflects the proportion of ordering explained by cluster-level structure rather than individual item ordering. Interpretation of H^T and H^{TB} followed the guidelines proposed by Koopman and Braeken (2025), wherein H^T values below .30 indicate too little consistency to justify interpretation, values between .30 and .40 are considered low consistency, .40–.50 moderate consistency, and values above .50 high consistency. For the H^{TB}/H^T ratio, values below .30 indicate that the cluster ordering is too inconsistent to be useful, .30–.60 suggest a reasonable contribution of cluster-level ordering, and values above .60 indicate high consistency (i.e. the

Table 2. Descriptive statistics for the PHQ-9 total scores and items.

	<i>M</i>	<i>SD</i>	<i>Skew</i>	<i>Kurtosis</i>	<i>n</i>
Total score					
Full sample	5.46	4.85	1.05	0.49	138
Men	4.84	5.16	1.07	0.24	62
Women	5.96	4.55	1.12	0.80	76
Item					
1	0.68	0.81	1.29	1.44	
2	0.54	0.71	1.17	0.87	
3	1.03	0.99	0.71	−0.52	
4	1.09	0.92	0.54	−0.51	
5	0.54	0.74	1.29	1.18	
6	0.61	0.86	1.25	0.59	
7	0.62	0.81	1.27	1.05	
8	0.22	0.48	2.43	7.46	
9	0.12	0.35	2.76	7.14	

ordering is largely attributable to the cluster-level structure). Estimated mean IRFs were also visually inspected to illustrate observed patterns and any deviations from the hypothesized ordering (Meijer & Egberink, 2012).

Although the MSA in this study was based on a relatively small sample, the method remains valuable for detecting problematic item characteristics under limited data conditions. Simulation studies by Wind (2022) have demonstrated that item-level scalability coefficients can identify issues such as monotonicity violations, low discrimination, and problems with IIO, even when these occur within specific regions of the latent trait. While larger samples increase detection power, meaningful insights were still obtained in samples as small as 100, particularly when multiple item quality indicators were considered. This supports the method’s appropriateness for the present study.

Results and discussion

Descriptive statistics

Table 2 presents descriptive statistics for the PHQ-9 total scores and individual items. The distribution of total scores in the sample was positively skewed, with most participants reporting relatively low levels of depressive symptoms. According to the severity categories defined by Kroenke et al. (2001), 55.1% of participants reported no depressive symptoms (scores 0–4), 26.1% reported mild symptoms (scores 5–9), 11.6% reported moderate symptoms (scores 10–14), 5.8% reported moderately severe symptoms (scores 15–19), and 1.4% reported severe depressive symptoms (scores ≥ 20). Mean item scores ranged from 0.12 (Item 9) to 1.09 (Item 4). As expected in a general population sample, items assessing more severe symptoms, such as thoughts of death or self-harm (Item 9) and psychomotor changes (Item 8), demonstrated marked positive skewness and kurtosis.

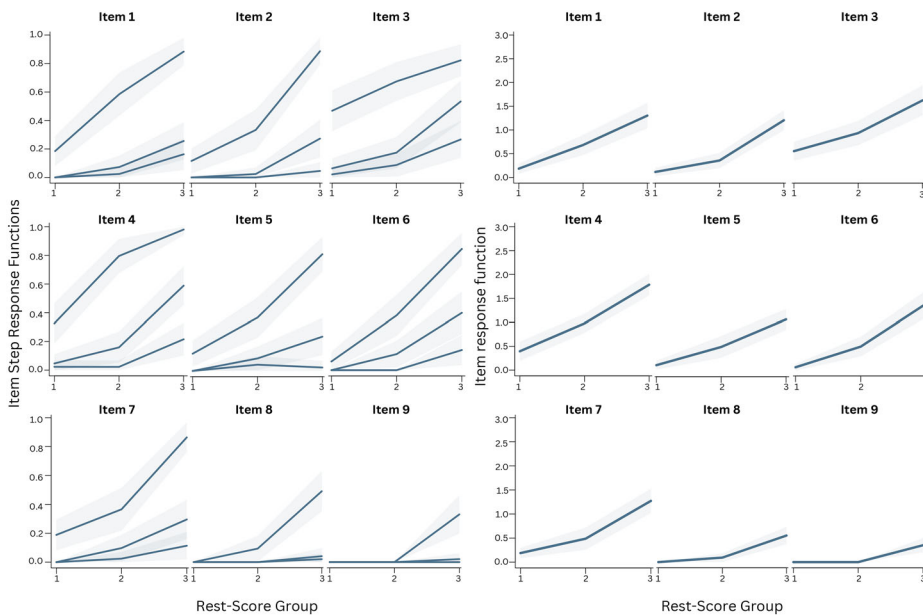


Figure 2. Item step response functions (ISRFs) and mean item response functions (IRFs) for the PHQ-9 items.

Note. Rest-score groups are numbered so that lower numbers indicate lower rest scores.

Monotonicity

No statistically significant violations of monotonicity were detected across the nine items. Visual inspection confirmed that the ISRFs and mean IRFs were monotonically non-decreasing (see Figure 2). Although a slight decline was observed in the ISRF for Item 5, it was not considered a meaningful violation. Overall, the monotonicity assumption was supported for the scale.

Local independence

No PHQ-9 item pairs were identified as positively locally dependent. However, several pairs were flagged as negatively locally dependent. Item 3 was involved in most of these associations, showing negative dependence with Items 1, 2, 5, and 6. Additionally, negative dependence was observed between Items 4 and 6. These results suggest that certain items may capture partially non-overlapping aspects of depressive symptomatology, leading some respondents to endorse one item while being less likely to endorse another at similar levels of overall depression severity.

To assess the stability of these findings, the analyses were repeated using the default grouping criterion of at least 50 respondents per rest-score group, rather than the minimum of 40 used for consistency with the monotonicity checks. Under this specification, no violations were detected. This sensitivity to rest-score group size indicates that the evidence for negative local dependence in this sample should be interpreted cautiously, as the flagged item-pairs may reflect sampling variability rather than systematic response patterns.

Table 3. Item and total scalability coefficients (H_i and H) and standard errors for the PHQ-9.

No.	Item	H_i	SE
1	Little interest or pleasure in doing things	.589	0.05
2	Feeling down, depressed or hopeless	.622	0.04
3	Trouble falling or staying asleep, or sleeping too much	.450	0.06
4	Feeling tired or having little energy	.586	0.05
5	Poor appetite or overeating	.492	0.06
6	Feeling bad about yourself – or that you are a failure or have let yourself or your family down	.590	0.05
7	Trouble concentrating on things, such as reading the newspaper or watching television	.542	0.05
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	.535	0.07
9	Thoughts that you would be better off dead, or of hurting yourself in some way	.516	0.07
Total		.548	0.04

Scalability

Item pair scalability coefficients (H_{ij}) were all positive and ranged from .308 (Items 3 and 5) to .759 (Items 1 and 2). The item scalability coefficients (H_i), summarised in Table 3, ranged between .450 and .622, with no item below the traditional .30 benchmark. Only Items 3 and 5 were slightly below the .50 criterion that we set for acceptable item scalability. However, the lower bounds of their confidence intervals still exceeded .30 (Item 3: 95% CI [.334, .566]; Item 5: 95% CI [.380, .604]).

The total scalability coefficient for all PHQ-9 items was $H = .548$ (95% CI [.466, .630]), exceeding the .50 criterion set for this study. This indicates that the items form a strong scale according to Mokken’s (1971) guidelines for interpreting overall scalability. However, the confidence interval includes values below .50, suggesting that the true scalability could fall below this threshold in the population.

Although several item pairs demonstrated negative local dependence, scalability coefficients remained moderate to strong overall. This suggests that the PHQ-9 items, despite some inconsistent co-endorsement patterns, are largely cohesive and informative for ordering respondents by depression severity. The negative dependence may have slightly attenuated the scalability of specific items, particularly Item 3 (as reflected in its lower H_i coefficient), but did not substantially compromise the scale’s overall performance.

Automated Item Selection Procedure (AISP)

To further examine whether the items form a single coherent scale, the AISP was applied across a range of lower-bound values ($c = 0, .05, .10, \dots, .55$). Across all lower-bound levels up to and including .45, all nine PHQ-9 items were assigned to a single scale. Item 3 was unscalable at $c \geq .50$, and Item 5 also became unscalable at $c = .55$. Because increasing the lower-bound value imposes progressively stricter scalability requirements, items with

comparatively lower H_i coefficients may fail to meet these criteria at higher c thresholds, even when the item set is unidimensional. Importantly, no items were reassigned to a separate scale at any lower-bound value, indicating that the AISP never suggested a multi-dimensional solution. Overall, the AISP results support interpreting the revised Icelandic PHQ-9 as a unidimensional scale.

Ordering structure

As the assumptions of monotonicity and local independence were largely satisfied, the ordering structure of the PHQ-9 was examined next. There were no violations of manifest weak invariant cluster ordering (weak ICO), indicating that the clusters were consistently ordered on average across the latent depression continuum. Similarly, no violations of manifest strong ICO were detected, suggesting that between-cluster ordering also held for the items. For manifest invariant item ordering (IIO), one potential within-cluster violation was detected involving Items 3 and 4. However, this violation did not reach statistical significance.

To further evaluate the overall consistency of the ordering structure, global fit was examined. The H^T coefficient was .407, indicating moderate consistency of ordering in the data. The H^{TB} coefficient, which reflects the consistency of ordering at the cluster level, was .237. The ratio of H^{TB} to H^T was .573, suggesting that about half of the observed ordering consistency can be attributed to stable ordering among clusters, with the remainder reflecting item-level ordering within clusters.

Repeating the analyses with the default minimum rest-score group size of 50 (instead of 40) yielded similar patterns, with no statistically significant violations and unchanged global fit coefficients.

Figure 3 shows the estimated mean IRFs across rest-score groups. For consistency, the most common rest-score grouping from the *check.monotonicity* function was used. Because this grouping differs somewhat from that used in the MIIO procedure, the observed intersections occur in slightly different rest score intervals than those flagged by the MIIO method. The plot shows that Items 8 and 9 were the only items whose IRFs did not overlap with any others, while Items 3 and 4 intersected once, consistent with the MIIO results. The remaining items showed multiple intersections, suggesting that their ordering is less stable across the latent trait continuum.

If an attempt were made to order the items by difficulty, it would only be evident that Items 3 and 4 have the highest expected item score across all rest-score groups, while Items 8 and 9 have the lowest. This pattern aligns with general expectations about how different depressive symptoms would typically emerge and intensify: sleep disturbances (Item 3) and fatigue (Item 4) are often reported even at lower levels of overall symptom severity, whereas psychomotor changes (Item 8) and thoughts of death or self-harm (Item 9) tend to occur less frequently and more at higher severity levels. The remaining items form a middle group with overlapping expected scores, indicating that their relative difficulty is less distinct and may vary across individuals.

While these results suggest an overall coherent ordering structure in this sample, they should be interpreted cautiously. The modest sample size ($n = 138$) limits power to detect violations and increases uncertainty around the estimates. Simulation studies (Wind, 2022)

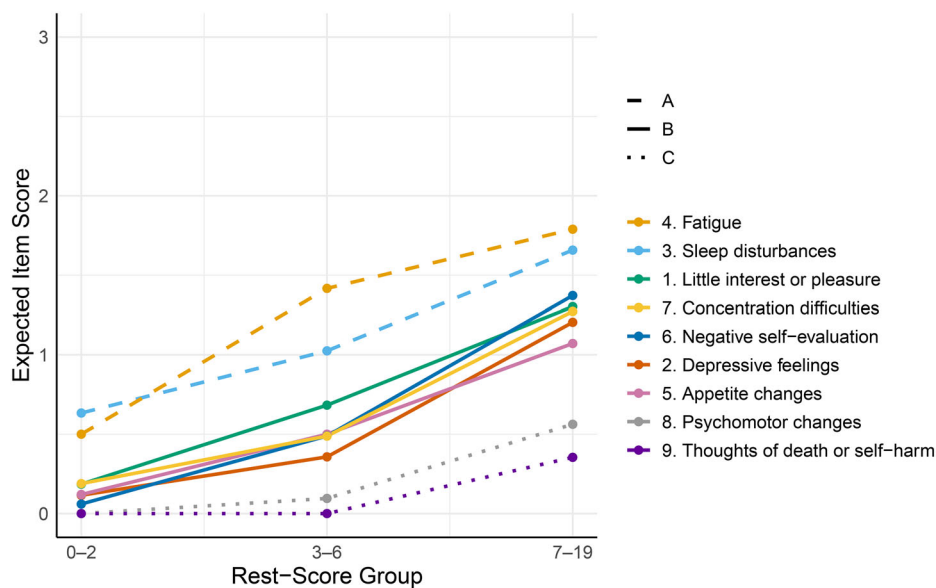


Figure 3. Mean item response functions (IRFs) across rest-score groups.
Note. Item clusters are indicated by different line types: Cluster A (dashed), Cluster B (solid), Cluster C (dotted).

have shown that even though items were not flagged for statistically significant IIO violations, they still exhibited problematic item characteristics. Taken together, these findings indicate that although the PHQ-9 shows some degree of ordering consistency at both the cluster and item levels, evidence for a strict invariant ordering structure is limited in this sample. Further research using larger samples could clarify whether the observed patterns reflect genuine properties of the scale or sample-related variability.

Comparison with the original translation

These findings suggest improvements compared to the original Icelandic translation evaluated by Kristófersdóttir et al. (2024). It is, however, important to note that given differences in sample composition (primary care vs. general population), the changes in scalability cannot be attributed solely to the revised wording. As summarised in Table 4, all H_i coefficients were higher in the present study. The total H coefficient also increased from .431 for the original translation to .548 in the revised version, an increase of .117. This comparison helps illustrate how specific modifications might have influenced the psychometric properties of individual items.

The items that changed the most were Items 3, 6, 7, and 8, all of which were revised from the original translation. Item 8 demonstrated the largest increase in scalability, rising from .367 to .535, an improvement of .168. This means it moved from an item that was just above the conventional .30 threshold to exceeding the .50 mark set for acceptable scalability in the present study. This notable change may reflect the modification from “that other people have noticed” to “that other people could have noticed.” The revision may have made the item less demanding to endorse, as the original wording implied that a behaviour must be severe enough for others to observe and comment on it. In cognitive interviews,

Table 4. Comparison of scalability coefficients: original vs. revised Icelandic translation of the PHQ-9.

Item	H_i (Original) ^a	H_i (Revised)	Difference
1*	.495	.589	.094
2	.510	.622	.112
3*	.296	.450	.154
4*	.522	.586	.064
5	.397	.492	.095
6*	.469	.590	.121
7*	.412	.542	.130
8*	.367	.535	.168
9*	.410	.516	.106
Min	.296	.450	
Max	.522	.622	
Range	.226	.172	
H 95% CI	.431 [.396, .466]	.548 [.466, .630]	.117
n	618	138	

Note. Scalability coefficients above .50 are in bold. PHQ-9 = Patient Health Questionnaire-9; CI = Confidence interval.

^aItem scalability coefficients from Kristófersdóttir et al. (2024) using the original translation of the PHQ-9.

*Item was changed from the original Icelandic translation.

participants often interpreted this requirement literally, which they reported made it difficult to assess whether it applied to them (Kristófersdóttir et al., 2024). Therefore, this revision has the potential to shift the focus to respondents' perceptions of their behaviour rather than requiring evidence of others' observations or comments.

Although Item 3 did not meet the .50 criterion for acceptable scalability in this study, its H_i coefficient increased from .296 to .450, representing a notable improvement of .154. This change places it above the conventional .30 threshold, elevating it from an item previously considered unscalable to one approaching acceptable performance. This improvement may reflect the significant modification of including the phrase "*or sleeping too much*," consistent with the original English version of the scale. This revision enables individuals experiencing hypersomnia, a common symptom of depression (Lee et al., 2024), to respond accurately to this item, thereby increasing the likelihood that their symptoms are reflected in the total score. However, it is also worth noting that Item 3 demonstrated negative local dependence with several other items in the scale, which may have contributed to its lower scalability compared to the other items despite the improved wording.

Item 7 showed a meaningful increase in scalability, rising from .412 to .542 (+.130), suggesting that the revised translation more effectively captures concentration difficulties associated with depression. The changes primarily involved clarifying that concentration problems can occur across a wider range of activities by adding the phrase "*on things*," which was omitted in the original translation. While the revision might have improved the item's performance, some limitations remain regarding content validity and relevance, as the examples "*reading the newspaper or watching television*" are still included (Kristófersdóttir et al., 2024).

The scalability coefficient of Item 6 increased by .121, from .469 to .590. The revisions involved making the wording more severe and removing an idiom that some respondents might find confusing or unclear. However, the item remains somewhat complex, as it asks about three different things simultaneously: feeling bad about oneself, feeling like a failure, and feeling that one has let oneself or one's family down. This triple-barrelled nature of the item reflects the original English version rather than a translation issue and, therefore, could not be fully resolved in the Icelandic translation.

Items 1, 4, and 9 were also revised, though the magnitude of improvement in scalability was less pronounced than for the previously discussed items. Each of these revisions was relatively minor and mainly intended to improve clarity and alignment with the original English version. While their H_i coefficients increased modestly, all three items surpassed the .50 threshold for acceptable scalability in the present study.

Interestingly, the scalability of Items 2 and 5 improved despite remaining unchanged. Several factors may help explain this. First, the current sample was drawn from the general population, whereas the earlier study used a primary care sample, which could have affected how these and other items performed. Second, the differences might reflect normal fluctuations in measurement. Third, Icelandic has three grammatical genders, and the revised version presented separate forms for each of them which made the items much more readable than the original translation that combined all genders into a single form. Fourth, the revision also corrected a translation error in the third response option, which was changed from *"More than half the time"* to *"More than half the days,"* potentially improving the interpretability of all items, including those not otherwise altered. Finally, the most likely explanation lies in how the H_i coefficients are computed: each item's scalability depends on its relationship with all other items. When the other items improve in clarity and better capture the underlying construct, the scalability of unchanged items can also increase because they are paired with better-performing "partners." These same considerations may also apply to items whose wording was revised, meaning that improvements in scalability cannot be attributed solely to changes in item phrasing.

Beyond improvements in individual items, the revised translation also demonstrates higher overall scalability. The H coefficient for the total scale improved from .431 to .548, surpassing the .50 threshold set for this study, though its confidence interval included values slightly below .50. Additionally, the range of H_i coefficients decreased from .226 in the original translation to .172 in the revised version, indicating that item performance has become more consistent across the scale. This greater uniformity suggests that disparities between the strongest and weakest items have been reduced, making the PHQ-9 a more cohesive instrument. Enhanced consistency in item scalability strengthens the scale's utility by ensuring that each item contributes meaningfully to assessing depressive symptom severity.

International studies

In addition to improvements observed between the Icelandic translations, the revised PHQ-9 also compares favourably with results reported in several international MSA studies. As detailed in Table 5, the overall scalability coefficient in the present study ($H = .548$) exceeds

Table 5. Comparison of scalability coefficients reported in MSA studies on the PHQ-9.

H_i	Adler et al. (2012) ^a	Boothroyd et al. (2018) ^a	Stochl et al. (2022) ^b	Bianchi et al. (2022)	Cosci et al. (2024)	Pretorius & Padmanabhanunni (2024)	The present study
1	.584	.524		.521	.44	.52	.589
2	.603	.538		.594	.56	.55	.622
3	.481	.452		.550	.48	.53	.450
4	.478	.481		.616	.54	.47	.586
5	.385	.442		.537	.45	.52	.492
6	.588	.467		.564	.53	.53	.590
7	.567	.477		.539	.46	.54	.542
8	.400	.415		.507	.44	.53	.535
9	.468	.439		.533	.52	.52	.516
Min	.385	.415	.420	.507	.44	.47	.450
Max	.603	.538	.547	.616	.56	.55	.622
Range	.218	.130	.127	.109	.12	.08	.172
H	.510	.471	.482	.554	.49	.52	.548
n	59	7,763	22,362	58,272	3,398	429	138

Note. Scalability coefficients above .50 are in bold. Language versions used in each study: Adler et al. = Swedish; Boothroyd et al. = English; Stochl et al. = English; Bianchi et al. = combined sample across French, German, Hebrew, English, and Spanish versions; Cosci et al. = not reported (administered in Italy); Pretorius & Padmanabhanunni = not reported (administered in South Africa). MSA = Mokken scale analysis; PHQ-9 = Patient Health Questionnaire-9.

^aStudies using a clinical sample.

^bItem-level scalability coefficients were not reported in the study.

those reported in most large-scale investigations (Boothroyd et al., 2018; Cosci et al., 2024; Stochl et al., 2022) and is similar to that observed by Bianchi et al. (2022).

At the item level, several items demonstrated scalability comparable to or exceeding that reported elsewhere. For example, Item 2 reached a H_i of .622, the highest value among all studies compared. Item 6 also performed well, with a H_i of .590. Most items in the present study surpassed the .50 threshold adopted here as an indicator of adequate scalability.

Item 3 showed weaker scalability (H_i = .450), a trend observed across most international studies. Item 5 similarly showed lower scalability, consistent with prior findings. These results suggest that challenges associated with these items may reflect broader conceptual characteristics of the PHQ-9 rather than issues unique to translation.

The range of H_i coefficients in the present study was somewhat wider than those reported in certain larger studies, indicating that while some variability in item performance exists, the items function with relative consistency overall. Nonetheless, the pattern of results suggests that the revised Icelandic PHQ-9 demonstrates scalability broadly in line with findings from other countries, populations, and language versions. While the sample size in the present study was smaller than in most other investigations, limiting the precision of the estimates, simulation studies (Wind, 2022) indicate that MSA can yield informative results in such contexts, particularly when multiple indicators of item quality are considered. Taken

together, these observations enhance confidence in the scale's cross-cultural applicability and support its use in Iceland.

Conclusions

The PHQ-9 is among the most widely used self-report scales to assess depression worldwide and is firmly established in Icelandic clinical practice and research. Ensuring that the instrument demonstrates strong psychometric properties in local language versions is critical for producing reliable assessments and supporting effective care.

This study provides preliminary evidence that the revised Icelandic translation of the PHQ-9 exhibits improved psychometric properties compared to the original version, as evaluated by Kristófersdóttir et al. (2024). All items had higher scalability coefficients, and the total scale scalability exceeded the threshold for a strong scale, although the confidence interval included values below .50, cautioning that strong scalability cannot be fully assumed at the population level. In addition, the revised version's scalability was broadly comparable to results reported in international studies, supporting its validity in cross-cultural comparisons. Importantly, the findings indicate that several of the translation issues identified in the earlier version have been addressed. Nevertheless, the observed improvements cannot be interpreted as resulting exclusively from the revised wording, as factors such as sample differences may also have contributed to the higher scalability estimates in this study.

Although the results indicate some degree of ordering consistency among the items, evidence for a strict invariant ordering structure was limited. This suggests that, although the revised translation of the PHQ-9 can be used to rank individuals by total summed scores, caution is warranted in interpreting the relative difficulty or expected severity ordering of specific items across respondents. While fatigue (Item 4) and sleep disturbances (Item 3) appeared relatively frequent even at lower severity levels, and psychomotor changes (Item 8) and thoughts of death or self-harm (Item 9) were rare unless overall severity was higher, the remaining symptoms showed considerable overlap. As a result, the scale's total score remains informative, but the ordering of individual symptoms should not be interpreted as reflecting a fixed progression of depressive severity.

Several limitations must be considered. First, the sample size was smaller than typically recommended for MSA (see Straat et al., 2016; Watson et al., 2018), which can reduce statistical power and limit the ability to detect violations of model assumptions. While simulation studies (e.g. Wind, 2022) suggest that informative results can still be obtained in smaller samples, the estimates should be interpreted with care. The analytic sample was reduced because participants were randomly assigned to two experimental conditions as part of a separate project, and only the subgroup aligned with the standard English PHQ-9 was relevant for this analysis. Recruitment was also limited by the fixed size of the available panel and the time available for data collection.

A second factor limiting the direct comparability between the old and new translations of the scale is the difference in the populations on which they were tested. The original translation was evaluated in a clinical population, whereas the revised version was tested in a non-clinical (general) population. While this is not a flaw, as the PHQ-9 is widely used in non-clinical settings and must be validated for that context, it does mean that conclusions about performance in clinical populations should be made cautiously. Further validation in

clinical samples is necessary to fully establish the revised version's utility across all relevant settings. Nevertheless, the current study represents an important and promising step in that broader validation process, offering preliminary support for the revised translation's psychometric strength and potential to outperform the original.

Finally, it is worth noting that, while the revised version improved clarity and alignment with the original English items, some substantive content issues noted by Kristófersdóttir et al. (2024) likely still persist. These reflect limitations inherent in the PHQ-9 items themselves and are, therefore, not specific to the Icelandic translation.

Overall, this study marks an important initial step toward improving the assessment of depressive symptoms in Iceland. Although further validation, particularly in clinical settings and larger samples, is needed, the findings suggest that the revised PHQ-9 may be a more consistent and interpretable instrument than the original translation. Future research should build on this foundation to more comprehensively establish the revised version's reliability, validity, and clinical utility.

Notes

1. The correlation between self-reported PHQ-9 total scores and scores from a telephone interview conducted within 48 hours.
2. In the original Icelandic translation, the word *gleði* ("joy") was used, whereas we considered *ánægja* a closer equivalent to "pleasure" in the original English version.

Acknowledgements

The authors declare that generative AI was used in the preparation of this manuscript. ChatGPT (GPT-5, OpenAI) was used exclusively for language editing, including improving clarity, grammar, and readability of the manuscript text. All analytic decisions, results, and interpretations were made and verified by the authors.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

No funding was received.

REFERENCES

- Adler, M., Hetta, J., Isacson, G., & Brodin, U. (2012). An item response theory evaluation of three depression assessment instruments in a clinical sample. *BMC Medical Research Methodology*, 12(1), 84. <https://doi.org/10.1186/1471-2288-12-84>
- Agnarsdóttir, A., Eyþórsdóttir, Á., Einarsson, E., Daniélsdóttir, G., Kristjánsdóttir, H., Sverrisdóttir, K., Tyrfinngsson, P., Jónsdóttir, S. E., Heiðarsdóttir, Þ. E., Sigurðsson, J. F., Ársælsdóttir, Þ., Ólafsdóttir, H. (2011). *Klínískar leiðbeiningar um þunglyndi og kvíða*. Landspítaliinn. https://www.landspitali.is/library/Sameiginlegar-skrar/Gagnasafn/BRUNNURINN/Kliniskar-leidbeiningar/thunglyndi-og-kvidi/klin_leid_thunglyndi_kvdi_agust_2011.pdf
- Agnarsdóttir, A., Gunnell, L. A., Þórisdóttir, K. (2021). *Gæðahandbók. Sálfræðiþjónusta fyrir fullorðna í heilsugæslu*. 1.línu þjónusta: Kvíðaraskanir, þunglyndi og áfallastreituröskun, væg til miðlungs

- einkenni. Þróunarmiðstöð Íslenskrar Heilsugæslu. <https://throunarmidstod.is/library/Salfraedingar/HANDB%c3%93K%202.1.22.pdf>
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders*. (4th ed.). APA.
- Beard, C., Hsu, K. J., Rifkin, L. S., Busch, A. B., & Björgvinsson, T. (2016). Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, 193, 267–273. <https://doi.org/10.1016/j.jad.2015.12.075>
- Bianchi, R., Verkuilen, J., Toker, S., Schonfeld, I. S., Gerber, M., Brähler, E., & Kroenke, K. (2022). Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. *Psychological Assessment*, 34(6), 595–603. <https://doi.org/10.1037/pas0001124>
- Boothroyd, L., Dagnan, D., & Muncer, S. (2018). Psychometric analysis of the Generalized Anxiety Disorder scale and the Patient Health Questionnaire using Mokken scaling and confirmatory factor analysis. *Health and Primary Care*, 2(4), 1–4. <https://doi.org/10.15761/HPC.1000145>
- Chae, D., Lee, J., & Lee, E. (2025). Internal structure of the Patient Health Questionnaire-9: A systematic review and meta-analysis. *Asian Nursing Research*, 19(1), 1–12. <https://doi.org/10.1016/j.anr.2024.12.005>
- Cosci, F., Christensen, K. S., Ceccatelli, S., Patierno, C., & Carrozzino, D. (2024). Patient Health Questionnaire-9: A clinimetric analysis. *Revista Brasileira de Psiquiatria (Sao Paulo, Brazil: 1999)*, 46, e20233449. Article e20233449. <https://doi.org/10.47626/1516-4446-2023-3449>
- Einarsdóttir, M., Stefánsson, K. H., & Gísladóttir, K. H. (2022). Vægi fjárhagsþrenginga í áhættu á þunglyndiseinkennum hjá launafólki á Íslandi á tímum kórónuveirunnar. *Íslenska Þjóðfélagið*, 13(1), 33–45.
- Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, 19(4), 337–352. <https://doi.org/10.1177/014662169501900404>
- Inoue, T., Tanaka, T., Nakagawa, S., Nakato, Y., Kameyama, R., Boku, S., Toda, H., Kurita, T., & Koyama, T. (2012). Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. *BMC Psychiatry*, 12(1), 73. <https://doi.org/10.1186/1471-244X-12-73>
- Jónsdóttir, M. K., Kristófersdóttir, K. H., Runólfssdóttir, S., Kristensen, I. S. U., Sigurjónsdóttir, H. Á., Eggertsdóttir Claessen, L. Ó., & Kristjánsdóttir, H. (2022). Concussion among female athletes in Iceland: Stress, depression, anxiety, and quality of life. *Nordic Psychology*, 74(4), 262–278. <https://doi.org/10.1080/19012276.2021.2004916>
- Koopman, L., & Braeken, J. (2025). Investigating the ordering structure of clustered items using nonparametric item response theory. *Educational and Psychological Measurement*, 85(2), 336–356. <https://doi.org/10.1177/00131644241274122>
- Kristófersdóttir, K. H., Kristjánsdóttir, H., Asgeirsdóttir, R. L., Karlsson, T., Vésteinsdóttir, V., & Thorsdóttir, F. (2024). Investigating the PHQ-9 with Mokken scale analysis and cognitive interviews. *Assessment*, 31(6), 1332–1355. <https://doi.org/10.1177/10731911231216961>
- Kroenke, K. (2021). PHQ-9: Global uptake of a depression scale. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 20(1), 135–136. <https://doi.org/10.1002/wps.20821>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Lamela, D., Soreira, C., Matos, P., & Morais, A. (2020). 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. *Journal of Affective Disorders*, 276, 220–233. <https://doi.org/10.1016/j.jad.2020.06.066>
- Laugesen, K., Ludvigsson, J. F., Schmidt, M., Gissler, M., Valdimarsdóttir, U. A., Lunde, A., & Sørensen, H. T. (2021). Nordic health registry-based research: A review of health care systems and key registries. *Clinical Epidemiology*, 13, 533–554. <https://doi.org/10.2147/CLEP.S314959>
- Lee, S. A., Mukherjee, D., Rush, J., Lee, S., & Almeida, D. M. (2024). Too little or too much: Nonlinear relationship between sleep duration and daily affective well-being in depressed adults. *BMC Psychiatry*, 24(1), 323. Article 323. <https://doi.org/10.1186/s12888-024-05747-7>
- Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific Reports*, 8(1), 2861. Article 2861. <https://doi.org/10.1038/s41598-018-21243-x>

- Ma, S., Yang, J., Yang, B., Kang, L., Wang, P., Zhang, N., Wang, W., Zong, X., Wang, Y., Bai, H., Guo, Q., Yao, L., Fang, L., & Liu, Z. (2021). The Patient Health Questionnaire-9 vs. the Hamilton Rating Scale for Depression in assessing major depressive disorder. *Frontiers in Psychiatry*, 12, 747139. Article 747139. <https://doi.org/10.3389/fpsy.2021.747139>
- Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibián-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., di Palma, M. A. (2024). *robustbase* Basic robust statisticsVersion R package version 0.99-4-1. <http://CRAN.R-project.org/package=robustbase>
- Management of Major Depressive Disorder Working Group. (2016). *VA/DoD clinical practice guideline for the management of major depressive disorder*. <https://www.healthquality.va.gov/guidelines/MH/mdd/VADoDMDDCPFINAL82916.pdf>
- Meijer, R. R., & Egberink, I. J. L. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, 72(4), 589–607. <https://doi.org/10.1177/0013164411429344>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. De Gruyter Mouton.
- Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, 12(37), 97–117.
- Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. iecProGAMMA.
- National Institute for Health and Care Excellence (2011). Common mental health problems: Identification and pathways to care. <https://www.nice.org.uk/guidance/cg123>
- Negeri, Z. F., Levis, B., Sun, Y., He, C., Krishnan, A., Wu, Y., Bhandari, P. M., Neupane, D., Brehaut, E., Benedetti, A., & Thombs, B. D. (2021). Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: Updated systematic review and individual participant data meta-analysis. *BMJ (Clinical Research ed.)*, 375, n2183. <https://doi.org/10.1136/bmj.n2183>
- OECD. (2021). *A new benchmark for mental health systems: Tackling the social and economic costs of mental ill-health*. OECD Health Policy Studies. <https://doi.org/10.1787/4ed890f6-en>
- Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). 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 and Anxiety*, 36(9), 813–823. <https://doi.org/10.1002/da.22940>
- Pretorius, T. B., & Padmanabhanunni, A. (2024). Examining the unidimensionality of the PHQ-9 with first responders: Evidence from different psychometric paradigms. *African Journal of Psychological Assessment*, 6(0), 165. Article a165 <https://doi.org/10.4102/ajopa.v6i0.165>
- R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Saevarsdóttir, K. S., Hilmarsdóttir, H. Ý., Magnúsdóttir, I., Hauksdóttir, A., Thordardóttir, E. B., Guðjónsdóttir, Á. B., Tomasson, G., Rúnarsdóttir, H., Jónsdóttir, H. L., Guðmundsdóttir, B., Pétursdóttir, G., Petersen, P. H., Kristinsson, S. Y., Love, T. J., Hansdóttir, S., Hardardóttir, H., Guðmundsson, G., Eythorsson, E., Guðmundsdóttir, D. G., ... Valdimarsdóttir, U. (2021). Illness severity and risk of mental morbidities among patients recovering from COVID-19: A cross-sectional study in the Icelandic population. *BMJ Open*, 11(7), e049967. <https://doi.org/10.1136/bmjopen-2021-049967>
- Sijtsma, K., & Molenaar, I. (2002). *5 Introduction to nonparametric item response theory*. Sage. <https://doi.org/10.4135/9781412984676>
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *The British Journal of Mathematical and Statistical Psychology*, 70(1), 137–158. <https://doi.org/10.1111/bmsp.12078>
- Silverman, J. J., Galanter, M., Jackson-Triche, M., Jacobs, D. G., Lomax, J. W., Riba, M. B., Tong, L. D., Watkins, K. E., Fochtmann, L. J., Rhoads, R. S., & Yager, J. (2015). The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults. *The American Journal of Psychiatry*, 172(8), 798–802. <https://doi.org/10.1176/appi.ajp.2015.1720501>
- Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA*, 282(18), 1737–1744. <https://doi.org/10.1001/jama.282.18.1737>
- Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B., & Perez, J. (2022). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment*, 29(3), 355–366. <https://doi.org/10.1177/1073191120976863>

- Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology*, 12(4), 117–123. <https://doi.org/10.1027/1614-2241/a000115>
- Tahtinen, R., Kristjansdóttir, H., Olason, D. T., & Morris, R. (2021). What lies beneath: Exploring different depressive symptoms across selected risk factors in Icelandic team sport athletes. *Journal of Clinical Sport Psychology*, 15(1), 54–79. <https://doi.org/10.1123/jcsp.2020-0040>
- Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy*, 40(2), 126–136. <https://doi.org/10.1080/16506073.2010.550059>
- Urtasun, M., Daray, F. M., Teti, G. L., Coppolillo, F., Herlax, G., Saba, G., Rubinstein, A., Araya, R., & Irazola, V. (2019). Validation and calibration of the Patient Health Questionnaire (PHQ-9) in Argentina. *BMC Psychiatry*, 19(1), 291. <https://doi.org/10.1186/s12888-019-2262-9>
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(11), 1–19. <https://doi.org/10.18637/jss.v020.i11>
- Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5), 1–27. <https://doi.org/10.18637/jss.v048.i05>
- Vésteinsdóttir, V., Reips, U.-D., Joinson, A., & Thorsdóttir, F. (2017). An item level evaluation of the Marlowe–Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews. *Personality and Individual Differences*, 107, 164–173. <https://doi.org/10.1016/j.paid.2016.11.023>
- Wang, Y., Hauksdóttir, A., Thordardóttir, E. B., Ge, F., Gísladóttir, E. U., Jakobsdóttir, J., Ásbjörnsdóttir, K. H., Rúnarsdóttir, H., Unnarsdóttir, A. B., Magnúsdóttir, I., Love, T. J., Kristinsson, S. Y., Palsson, R., Zoega, H., Fang, F., Tómasson, G., Song, H., Aspelund, T., & Valdimarsdóttir, U. (2024). Depressive symptom trajectories among general population during the COVID-19 pandemic in Iceland: A prospective cohort study (2020–2023). *BMJ Public Health*, 2(2), e001294. <https://doi.org/10.1136/bmjph-2024-001294>
- Watson, R., Egberink, I. J. L., Kirke, L., Tendeiro, J. N., & Doyle, F. (2018). What are the minimal sample size requirements for Mokken scaling? An empirical example with the Warwick–Edinburgh Mental Well-Being Scale. *Health Psychology and Behavioral Medicine*, 6(1), 203–213. <https://doi.org/10.1080/21642850.2018.1505520>
- Wind, S. A. (2022). Identifying problematic item characteristics with small samples using Mokken scale analysis. *Educational and Psychological Measurement*, 82(4), 747–756. <https://doi.org/10.1177/00131644211045347>
- World Health Organization (2017). *Depression and other common mental disorders: Global health estimates*. <https://iris.who.int/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1>
- World Health Organization (2023, March 31). *Depressive disorder (depression)*. <https://www.who.int/news-room/fact-sheets/detail/depression>