Adaptation of the PHQ-4 Short Screening for Depression and Anxiety to increase its Sensitivity to Subclinical Variability

Dominique Makowski[1], An Shu Te[1], & S.H. Annabel Chen[1, 2, 3, 4]

[1] School of Social Sciences, Nanyang Technological University, Singapore

[2] LKC Medicine, Nanyang Technological University, Singapore

[3] National Institute of Education, Singapore

[4] Centre for Research and Development in Learning, Nanyang Technological University, Singapore

Author Note

The authors made the following contributions. Dominique Makowski: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft; An Shu Te: Project administration, Resources, Software, Investigation, Writing – original draft; S.H. Annabel Chen: Project administration, Supervision, Writing – review & editing.

Correspondence concerning this article should be addressed to Dominique Makowski, HSS 04-18, 48 Nanyang Avenue, Singapore. E-mail: dom.makowski@gmail.com

Abstract

The PHQ-4 is an ultra-brief (4 items) screening questionnaire for depression and anxiety. We propose to add one additional response option ("Once or twice", in between "Not at all" and "Several days") to improve its sensitivity to milder alterations, and thus increase its usefulness in subclinical populations. Using Item Response Theory (IRT), we provide evidence that this new option does indeed capture specific portions of the measured constructs in the general population.

Significance Statement. Developing reliable and sensitive instruments for mood disorders screening is critical in a global context marked by international crises (pandemics, wars), where increasingly more surveys are done online. In this study, we show that a small modification to the widely used PHQ-4 scale (adding the "Once or twice" response option) can increase its ability to capture the mild fluctuations prevalent in subclinical samples.

Keywords: PHQ-4; depression; anxiety; brief questionnaire; short scale

Word count: 1249

Adaptation of the PHQ-4 Short Screening for Depression and Anxiety to increase its Sensitivity to Subclinical Variability

Introduction

The Patient Health Questionnaire-4 (PHQ-4) is an ultra brief measurement of core signs of depression and anxiety (Kroenke et al., 2009). It consists of two items for depression (PHQ–2, Kroenke et al., 2003) and anxiety (GAD–2, Kroenke et al., 2007), each corresponding to DSM-5 diagnostic symptoms for major depressive disorder (MDD) and generalized anxiety disorder (GAD). It has been validated across many languages and populations (Christodoulaki et al., 2022; Materu et al., 2020; Mendoza et al., 2022), becoming one of the most popular screening instruments for depression and anxiety (Maurer et al., 2018).

While the scale has been validated and used in the general population and non-clinical samples (Hajek & König, 2020; Löwe et al., 2010), its initial purpose was to reliably discriminate and identify potential MDD/GAD patients. This discriminative goal materializes in the scale's design and the existence of categorical cut-offs, which does not necessary entail a strong sensitivity to milder mood alterations. In particular, the gap between the two lowest possible answers, "Not at all" and "Several days", is quite large and leaves out the possibility of more subtle occurrences. While this is not necessarily an issue in clinical and diagnostic contexts, it might lead to a sub-optimal discrimination of affective levels on the lower end of the spectrum, important for instance in the context of subclinical variability quantifications. The goal of this study is hence to enhance, with minimal changes to the original scale, the sensitivity to mild mood level inflections.

Methods

Original Scale and Revision

The instructions *"Over the last 2 weeks, how often have you been bothered by the following problems?"* are followed with 4 items (A1 - *Feeling nervous, anxious or on edge*; A2 - *Not being able to stop or control worrying*; D1 - *Little interest or pleasure in doing things*; D2 - *Feeling down, depressed, or hopeless*). The original answer options are "Not at all" (0), "Several days" (1), "More than half the days" (2), "Nearly every day" (3). The total score is computed by summing the responses of each facet.

In order to better capture potential mild mood inflections without altering the scale scoring or structure, we added a "Once or twice" option between "Not at all" and "Several days" (see Dobson & Mothersill, 1979 for the choice of the label).

Participants

The sample consists of 485 English-speaking participants (Mean age = 30.1, SD = 10.1, range: [18, 73]; Sex: 50.3% females, 49.7% males) from the general population recruited via *Prolific*, a crowd-sourcing platform recognized for providing high quality data (Peer et al., 2022). The only inclusion criterion was a fluent proficiency in English to ensure that the task instructions would be well-understood. Note that the participants were administered the refined PHQ-4 online as part of another study, which data is available in open-access at https://github.com/RealityBending/IllusionGameReliability. This study was approved by the NTU Institutional Review Board (NTU IRB-2022-187). All participants provided their informed consent prior to participation and were incentivized after completing the study.

Results

Descriptive Statistics

The reliability of the anxiety (*Cronbach's* $\alpha = 0.903$; RMSEA = 0.031) and depression (*Cronbach's* $\alpha = 0.841$; RMSEA = 0.044) subscales is excellent. The proportion of response types stratified by item (see **Figure 1A**) shows that the new "Once or twice" option was the most prevalent response for all items (on average selected in 29.12% of cases).

Item Response Theory

Item Response Theory (IRT) provides insights into how well items and responses capture an underlying latent trait θ. For each of the subscales, we fitted a unidimensional graded response model (GRM, Samejima, 1997). For anxiety, the two items captured 89.2% of the variance of the latent anxiety dimension ($\theta_{anxiety}$). The discrimination parameters suggested that the first item was less precise ($\alpha = 3.42$) than the second item ($\alpha = 12.55$) in its ability to discriminate between various levels of anxiety (i.e., each response on the second item covers a more exclusive range of $\theta_{anxiety}$, as can be seen in **Figure 1B**). The two depression items captured 82.8% of the variance of its latent trait ($\theta_{depression}$), and the opposite pattern was found: the first item had a higher precision ($\alpha = 16.46$) than the first ($\alpha = 2.41$). However, it is important to note that the "less precise" items were also the ones covering a larger portion of the latent space (being more sensitive especially on the lower end of the spectrum), offering an interesting trade-off between sensitivity and precision. Importantly for our objective, the added "Once or twice" option does cover a selective and unique portion of the latent space.

Scoring

We propose two types of scoring procedures. The first aims at minimally disrupting the original scale and making its scores comparable, enabling comparisons across studies

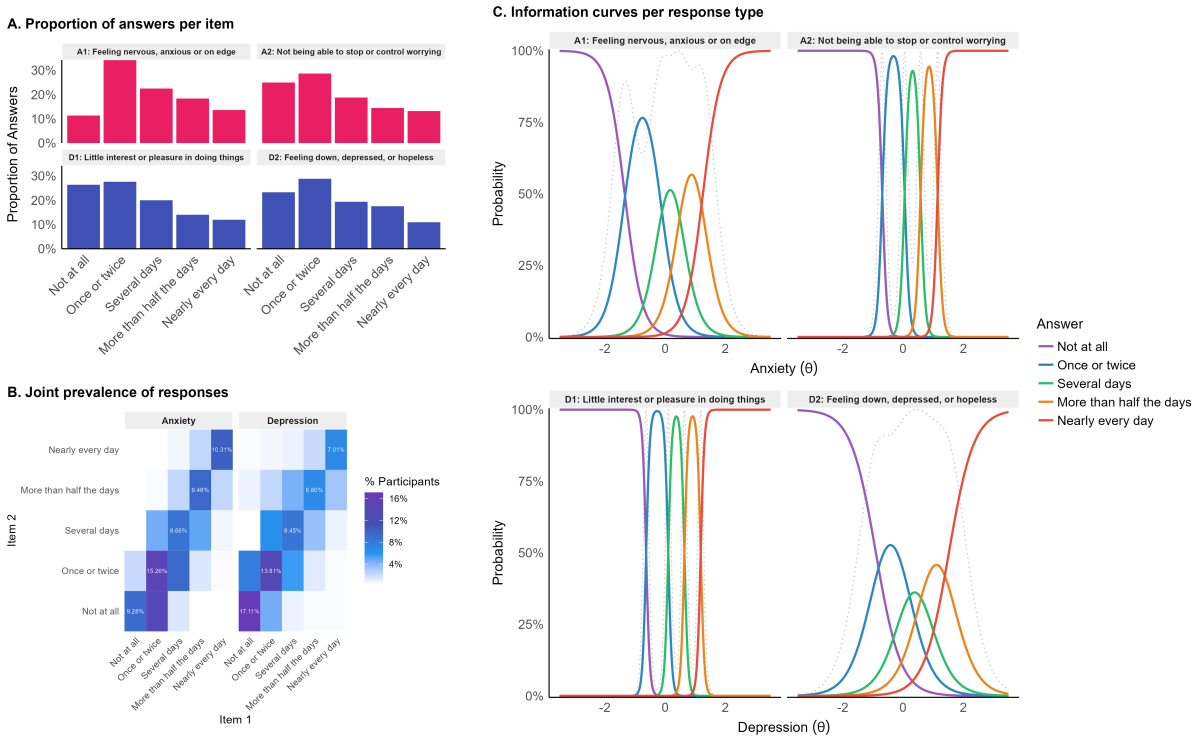

Figure 1. A) Proportion of answers of each type. B) Prevalence of answer pairs. C) Item Information Curves from IRT showing the coverage by each item and response of the latent dimension. Typically, an optimally informative item would display a large coverage over theta, with each response presenting a narrow coverage (high discrimination between different levels).

and the usage of developed cut-offs and norms. We suggest to score the new option, "Once or twice", as 0.5, occupying the middle space between 0 and 1 (preserving in this way the total score range of 0 - 8 per dimension).

The second scoring method we propose takes into account the non-linear spacing between responses, as indicated by the peak of their measurement information. Specifically, we found for each dimension the lower and higher values of the latent trait θ for which the probability of identification is 95%. We considered these points as the 0 and 1 ends, and normalized the location of the peaks of each response in between.

To illustrate, a person answering "Several days" and "Once or twice" to the first and second item of the anxiety subscale would have a score of 1.5 (1 + 0.5 following the basic scoring) and a normalized score of 0.49 ($(0.55 + 0.44)/2$).

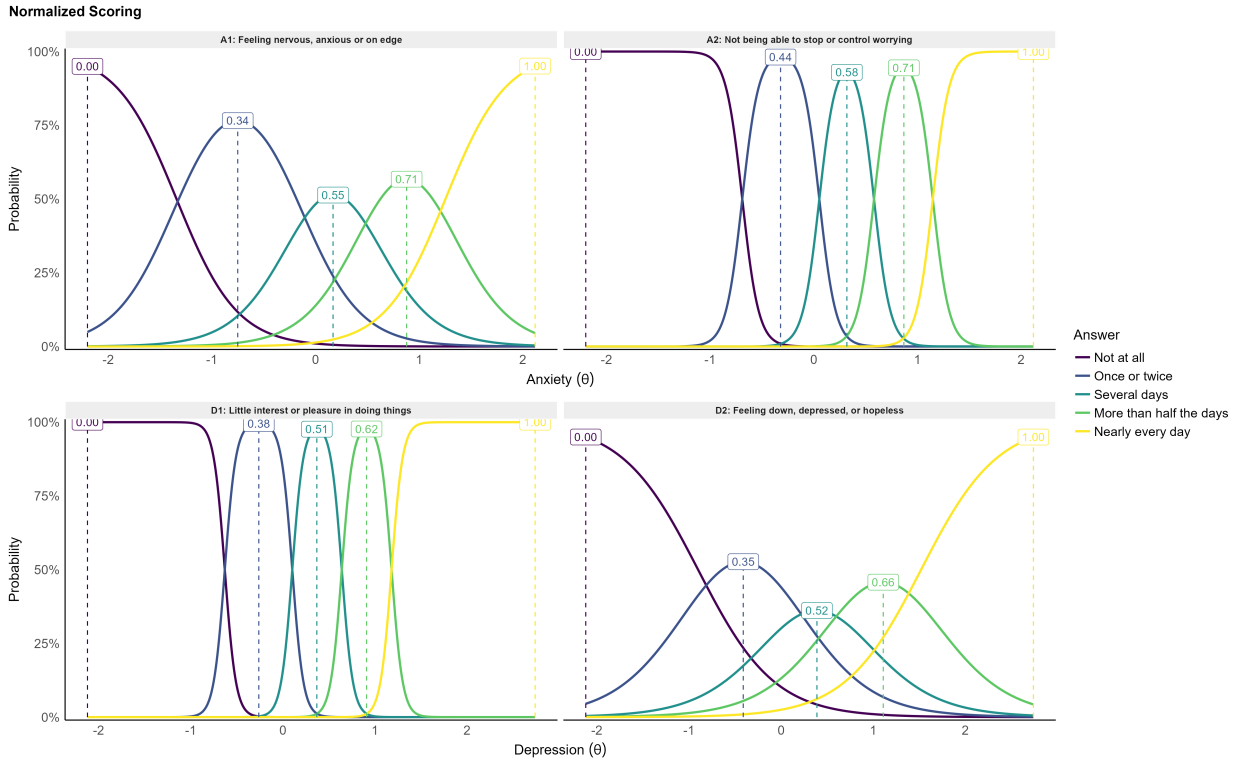

Figure 2. Normalized scoring of the items taking into account the non-linear spacing between responses (based on the peak of their measurement information).

Discussion

The objective of this study was to test the introduction of a "Once or twice" response option to the PHQ-4 to enhance its sensitivity to milder mood fluctuations. The fact that this new response option was the most prevalent response made by participants in our study is in itself evidence for its usefulness. IRT analysis further revealed that this response captures with precision a unique portion of the variability in the latent factors measured by the instrument. Taken together, our results suggest that adding this option response increases the scale's potential to discriminate average mood levels (which are superior to zero) from lower-end extremes (the true zero).

It is important to emphasize that our study is a statistical validation in the general population, which analysed how different responses relate to one another. It is a first step providing supporting evidence for the proposed change. However, the structural gains that

Table 1

Refined Patient Health Questionnaire-4 (PHQ-4R). The instructions are "Over the last 2 weeks, how often have you been bothered by the following problems?". The 0.5 scoring is introduced to preserve the compatibility with the original version, and enables to use its norms and cut-offs. The normalized scoring takes into account the non-linear spacing between item responses.

Facet	Item	Response	Basic Scoring	Normalized Scoring
Anxiety	Feeling nervous, anxious or on edge	*Not at all*	0	0
		Once or twice	0.5	0.34
		Several days	1	0.55
		More than half the days	2	0.71
		Nearly every day	3	1
	Not being able to stop or control worrying	*Not at all*	0	0
		Once or twice	0.5	0.44
		Several days	1	0.58
		More than half the days	2	0.71
		Nearly every day	3	1
Depression	Little interest or pleasure in doing things	*Not at all*	0	0
		Once or twice	0.5	0.38
		Several days	1	0.51
		More than half the days	2	0.62
		Nearly every day	3	1
	Feeling down, depressed, or hopeless	*Not at all*	0	0
		Once or twice	0.5	0.35
		Several days	1	0.52
		More than half the days	2	0.66
		Nearly every day	3	1

we observed in terms off finer sensitivity should be confirmed 1) by cross-validating the PHQ-4R with more comprehensive measures of anxiety and depression, and 2) testing the usefulness of this revision on more targeted samples (e.g., clinically characterized populations) to investigate the impact on existing cut-offs and on the detection of mood disorders. Additionally, evidence comparing the refined version against the original PHQ-4 would be useful to better understand whether the new “Once or twice” option pulls participants that would have otherwise answer “Not at all” or “Several days”.

Data Availability

The dataset analysed during the current study are available in the GitHub repository https://github.com/DominiqueMakowski/PHQ4R. We have created an R function for easy scoring according to the two methods described, which can be used as follows:

```
# Load the function
source("https://raw.githubusercontent.com/DominiqueMakowski/PHQ4R/main/score_PHQ4.R")


# Enter the responses
score_PHQ4(A1 = "Not at all", A2 = "Once or twice",
           D1 = "Once or twice", D2 = "several days",
           method = "basic")
##   A1  A2  D1 D2 Anxiety Depression
## 1  0 0.5 0.5  1     0.5        1.5
```

Funding

This work was supported by the Presidential Postdoctoral Fellowship Grant (NTU-PPF-2020-10014) from Nanyang Technological University (awarded to DM).

References

Christodoulaki, A., Baralou, V., Konstantakopoulos, G., & Touloumi, G. (2022). Validation of the patient health questionnaire-4 (PHQ-4) to screen for depression and anxiety in the greek general population. *Journal of Psychosomatic Research*, *160*, 110970.

Dobson, K. S., & Mothersill, K. J. (1979). Equidistant categorical labels for construction of likert-type scales. *Perceptual and Motor Skills*, *49*(2), 575–580.

Hajek, A., & König, H.-H. (2020). Prevalence and correlates of individuals screening positive for depression and anxiety on the phq-4 in the german general population: Findings from the nationally representative german socio-economic panel (GSOEP). *International Journal of Environmental Research and Public Health*, *17*(21), 7865.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2003). The patient health questionnaire-2: Validity of a two-item depression screener. *Medical Care*, 1284–1292.

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ–4. *Psychosomatics*, *50*(6), 613–621.

Kroenke, K., Spitzer, R. L., Williams, J. B., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, *146*(5), 317–325.

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., & Brähler, E. (2010). A 4-item measure of depression and anxiety: Validation and standardization of the patient health questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders*, *122*(1-2), 86–95.

Materu, J., Kuringe, E., Nyato, D., Galishi, A., Mwanamsangu, A., Katebalila, M., Shao, A., Changalucha, J., Nnko, S., & Wambura, M. (2020). The psychometric properties of PHQ-4 anxiety and depression screening scale among out of school adolescent girls and young women in tanzania: A cross-sectional study. *BMC Psychiatry*, *20*(1), 1–8.

Maurer, D. M., Raymond, T. J., & Davis, B. N. (2018). Depression: Screening and diagnosis. *American Family Physician*, *98*(8), 508–515.

Mendoza, N. B., Frondozo, C. E., Dizon, J. I. W. T., & Buenconsejo, J. U. (2022). The factor structure and measurement invariance of the PHQ-4 and the prevalence of depression and anxiety in a southeast asian context amid the COVID-19 pandemic. *Current Psychology*, 1–10.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.