

**1 Measuring Depression and Anxiety with 4 items? Adaptation of the PHQ-4 to increase its
2 Sensitivity to Subclinical Variability**

3 Dominique Makowski^{1,2}, An Shu Te³, Ana Neves¹, and S.H. Annabel Chen^{3,4,5}

4 ¹School of Psychology, University of Sussex

5 ²Sussex Centre for Consciousness Science, University of Sussex

6 ³School of Social Sciences, Nanyang Technological University

7 ⁴LKC Medicine, Nanyang Technological University

8 ⁵Centre for Research and Development in Learning, Nanyang Technological University

Author Note

Dominique Makowski  <https://orcid.org/0000-0001-5375-9967>

An Shu Te <https://orcid.org/0000-0002-9312-5552>

Ana Neves <https://orcid.org/0009-0006-0020-7599>

S.H. Annabel Chen  <https://orcid.org/0000-0002-1540-5516>

Author roles were classified using the Contributor Role Taxonomy (CRediT;

¹⁶ <https://credit.niso.org/>) as follows: Dominique Makowski: Conceptualization, Data curation,

17 Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration,

18 Resources, Software, Supervision, Validation, Visualization, Writing – original draft; An Shu Te:

¹⁹ Project administration, Resources, Writing – original draft; Ana Neves: Data curation, Formal

²⁰ Analysis, Writing – original draft, Writing – review & editing; S.H. Annabel Chen: Project

²¹ administration, Supervision, Writing – review & editing

Correspondence concerning this article should be addressed to Dominique Makowski,

²³ Email: D.Makowski@sussex.ac.uk

24

Abstract

25 The PHQ-4 is an ultra-brief (4 items) screening questionnaire for depression and anxiety. In this
26 brief report, we test the benefits of adding one additional response option (“Once or twice”, in
27 between “Not at all” and “Several days”) to improve the scale’s sensitivity to milder alterations,
28 and thus increase its usefulness in subclinical populations. In study 1 (N=485), we provide
29 evidence using Item Response Theory (IRT) that the new response option does improve the
30 scale’s psychometric quality and extends the sensitivity to the measured constructs on the lower
31 end of the spectrum. In study 2 (N=836), we show that the refined version offers an improved
32 sensitivity to subclinical variability in depression (indexed by the BDI-II) as compared to the
33 original version. In conclusion, adding the “once or twice” response option is a low-cost
34 no-downsides way of increasing the PHQ-4’s sensitivity to subclinical variability, making it a tool
35 of choice for general population research.

36 *Keywords:* PHQ-4, depression, anxiety, brief questionnaire validation, ultra short scale

37 Measuring Depression and Anxiety with 4 items? Adaptation of the PHQ-4 to increase its 38 Sensitivity to Subclinical Variability

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

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

17 This brief report aims at testing the possibility of enhancing - with minimal changes to the
18 original scale - the PHQ-4 sensitivity to mild mood level inflections. In the first study, we will
19 evaluate whether the new response option is prevalently used by participants, and whether it does
20 capture a specific part of the latent measure. In the second study, we will compare the refined
21 PHQ-4 version to the original one in terms of sensitivity to subclinical variability in depression,
22 using the Beck Depression Inventory (BDI-II, [Beck et al., 1996](#)) and the State Trait Anxiety
23 Inventory (STAI-5, [Zsido et al., 2020](#)) as our ground-truth measures of depression and anxiety.

24

Study 1

25 **Method**26 **Participants**

27 The sample consists of 485 English-speaking participants (Mean age = 30.1 ± 10.1 [18,
28 73]; 50.3% females) from the general population recruited via *Prolific*, a crowd-sourcing platform
29 recognized for providing high quality data (Peer et al., 2022). The only inclusion criterion was a
30 fluent proficiency in English to ensure that the task instructions would be well-understood. This
31 study was approved by the NTU Institutional Review Board (NTU IRB-2022-187). All
32 participants provided their informed consent prior to participation and were incentivized after
33 completing the study.

34 **Measures**

35 In the original PHQ-4, the instructions “*Over the last 2 weeks, how often have you been
36 bothered by the following problems?*” are followed with 4 items (A1 - *Feeling nervous, anxious
37 or on edge*; A2 - *Not being able to stop or control worrying*; D1 - *Little interest or pleasure in
38 doing things*; D2 - *Feeling down, depressed, or hopeless*). The original answer options are “Not at
39 all” (0), “Several days” (1), “More than half the days” (2), “Nearly every day” (3). The total score
40 is computed by summing the responses of each facet resulting in a 0-6 score for depression and
41 anxiety.

42 For the refined version, we added a “Once or twice” option between “Not at all” and
43 “Several days” in order to better capture potential mild mood inflections (see Dobson &
44 Mothersill, 1979 for the choice of the label). This new option was scored as 0.5 to preserve the
45 same scoring as the original version.

46 **Procedure**

47 Participants were administered the refined PHQ-4 online as part of another study, which
48 contained additional questionnaires and tasks not relevant for the current analysis. The PHQ-4
49 was presented in a randomized order with other questionnaires. The data is available in

50 open-access at <https://github.com/RealityBending/IllusionGameReliability>.

51 **Results**

52 The analysis was carried out using *R* 4.4 (R Core Team, 2023), the *tidyverse* (Wickham et
53 al., 2019), and the *easystats* collection of packages (Lüdecke et al., 2019, 2020, 2021; Patil et al.,
54 2022). All reproducible scripts and complimentary analyses are available open-access at
55 <https://github.com/DominiqueMakowski/PHQ4R>

56 **Descriptive Statistics**

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

61 **Item Response Theory**

62 Item Response Theory (IRT) provides insights into how well items and responses capture
63 an underlying latent trait θ . For each of the subscales, we fitted a unidimensional graded response
64 model (GRM, Samejima, 1997). For anxiety, the two items captured 89.2% of the variance of the
65 latent anxiety dimension ($\theta_{anxiety}$). The discrimination parameters suggested that the first item
66 was less precise (α = 3.42) than the second item (α = 12.55) in its ability to discriminate between
67 various levels of anxiety (i.e., each response on the second item covers a more exclusive range of
68 $\theta_{anxiety}$, as can be seen in Figure 1). The two depression items captured 82.8% of the variance of
69 its latent trait ($\theta_{depression}$), and the opposite pattern was found: the first item had a higher
70 precision (α = 16.46) than the first (α = 2.41). However, it is important to note that the "less
71 precise" items were also the ones covering a larger portion of the latent space (being more
72 sensitive especially on the lower end of the spectrum), offering an interesting trade-off between
73 sensitivity and precision. Importantly for our objective, the added "Once or twice" option did
74 cover a selective and unique portion of the latent space.

75 **Discussion**

76 The fact that the new “Once or twice” response option was the most prevalent response
77 speaks to its usefulness in capturing more accurately participants’ expression. The IRT analysis
78 further revealed that this response tracks with precision a unique portion of the variability in the
79 latent factors measured by the instrument. Taken together, our results suggest that adding this
80 option response increases the scale’s potential to discriminate average mood levels (which are
81 superior to zero) from lower-end extremes (the true zero).

82 One natural methodological limitation pertains to the interpretation of the latent
83 dimensions in an IRT framework applied to pairs of items. In this study’s context, references to
84 for instance “the latent anxiety dimension” merely corresponds to the amalgamation of the two
85 items of the anxiety subscale, and not to a more general and valid true anxiety factor.

86 **Study 2**

87 **Method**

88 **Participants**

89 The initial sample consisted of 1053 participants, recruited (181 were recruited on
90 *Prolific*, 772 students from the University of Sussex via *SONA*, and the rest through convenience
91 sampling as part of dissertation students’ data collection). We used attention checks as the
92 primary target for participant exclusion. We excluded 194 participants (18.42%) for failing at least
93 one attention check, and 23 (2.18%) that were outliers ($|z_{robust}| > 2.58$) on measures significantly
94 related to the probability of failing attention checks (namely, the standard deviation of all the
95 items of the IAS, as well as the the multivariate distance obtained with the OPTICS algorithm, see
96 [Thériault et al., 2024](#)). The experiment duration was not related to the probability of failing
97 attention checks and was thus not used as an exclusion criterion.

98 The final sample included 836 participants (Mean age = 25.1 ± 11.3 [18, 76]; 73.8%
99 women). This study was approved by the University of Sussex’ Ethics Committee (ER/ASF25/4).

100 In this sample, 51 participants (6.10%) were labelled as having Depression, as indexed by
101 the self-reported presence of MDD together with the use of a treatment (antidepressent, anxiolytic

102 and/or therapy), and 87 participants (10.41%) were labelled as having Anxiety, as indexed by the
103 self-reported presence of GAD or Panic Disorder, also together with the use of a treatment.

104 **Measures**

105 Participants were randomly assigned to complete either the original or refined version of
106 the PHQ-4, which included one additional response option (“Once or twice”) scored 0.5 (creating
107 more possible total scores - 0.5, 1.5, 2.5, etc.).

108 Beck’s Depression Inventory (BDI-II, [Beck et al., 1996](#)) was used as a ground truth
109 measure of depressive symptoms. It includes 21 items, each addressing a specific depression
110 symptom and offering four response options scored from 0 to 3. Participants are instructed to
111 select the option that best describes how they have felt over the past two weeks. The total score is
112 calculated by summing the scores for all 21 items, with higher scores indicating greater severity of
113 depressive symptoms.

114 The short version of the State-Trait Anxiety Inventory (STAI-5, [Zsido et al., 2020](#)) was
115 used as a ground truth measure of anxiety. This abridged version of the STAI ([Spielberger, 1970](#))
116 includes 5 items rated on a 4-point Likert scale. Changes were made in the instructions from
117 asking “*how participants feel right now*” to “*over the past 2 weeks*” to keep it consistent with the
118 instructions of the PHQ-4 and BDI-II. A general score of anxiety was computed by averaging all
119 the items.

120 Participants were also asked to complete two questionnaires of interoception, namely the
121 Interoceptive Accuracy Scale (IAS - 21 items rated on analog scales, [Murphy et al., 2020](#)) and the
122 Multidimensional Assessment of Interoceptive Awareness (MAIA-2 - 37 items, [Mehling et al.,
2018](#)).

124 After demographic questions, participants were asked to report the current presence of
125 psychiatric issues (from a list), as well as the usage of treatment (antidepressants, mood
126 stabilizers, anxiolytics, therapy). We indexed the presence of a depression when participants
127 reported suffering from either Major Depressive Disorder (MDD) or Dysthymia, as well as
128 undergoing a medical treatment. Similarly, we indexed the presence of an anxiety disorder when

129 participants reported suffering from either Generalized Anxiety Disorder (GAD) or Panic
130 Disorder, as well as undergoing a medical treatment.

131 **Procedure**

132 The original or refined version of the PHQ-4 was followed by the BDI-II, STAI-5, IAS,
133 and MAIA-2, presented in random order. The IAS and the MAIA-2 were included as part of
134 another study focused on interoception, and were only used in this study as part of data quality
135 control checks. ## Results

136 As all the scripts, analysis details and results tables are available open-access at
137 <https://github.com/DominiqueMakowski/PHQ4R>, we will focus on reporting the main results.

138 **PHQ-4 Depression vs. BDI-II**

139 The linear regression predicting the BDI-II total score with the PHQ-4 depression score
140 showed no interaction related to the PHQ-4 version
141 ($\Delta\text{Intercept}_{\text{refined}} = -0.13$, 95% CI $[-1.73, 1.47]$, $t(832) = -0.16$, $p = 0.871$;
142 $\Delta\beta_{\text{refined}} = -0.05$, 95% CI $[-0.70, 0.60]$, $t(832) = -0.15$, $p = 0.883$), suggesting no differences
143 in the relationship pattern between the two versions (see Figure 2).

144 Moreover, Bayesian t -tests (using *BayesFactor*'s `ttestBF()` function with default priors,
145 [Morey & Rouder, 2024](#)) comparing the BDI-II scores between the refined and the original version
146 at each integer score (0, 1, 2, 3) yielded no evidence in favour of a significant difference ($\text{BF} > 3$).
147 In other words, having the same score on the refined version as on the original version was related
148 to the same outcome on the BDI-II.

149 However, the low in-between scores from the refined version are overall capturing
150 significantly different levels of depression compared to the adjacent scores. Scoring 0.5 was
151 associated with a higher BDI-II score than scoring 0 ($\text{BF} > 30$), and lower scores than scoring 1
152 ($\text{BF} > 30$). Similarly, scoring 1.5 was associated with a higher BDI-II score than scoring 1 ($\text{BF} >$
153 30), but not lower scores than scoring 2 ($\text{BF} = 0.234$).

154 ***PHQ-4 Anxiety vs. STAI-5***

155 The linear regression predicting the STAI-5 general score with the PHQ-4 anxiety score
156 showed no interaction related to the PHQ-4 version
157 ($\Delta\text{Intercept}_{\text{refined}} = -0.02$, 95% CI $[-0.15, 0.11]$, $t(832) = -0.32$, $p = 0.750$;
158 $\Delta\beta_{\text{refined}} = 0.01$, 95% CI $[-0.03, 0.05]$, $t(832) = 0.56$, $p = 0.576$), suggesting no differences in
159 the relationship pattern between the two versions.

160 Moreover, Bayesian *t*-tests comparing the STAI-5 scores between the refined and the
161 original version at each integer score yielded no evidence in favour of a significant difference. In
162 other words, having the same score on the refined version as on the original version was related to
163 the same outcome on the STAI-5.

164 However, comparing in-between scores with adjacent scores yielded mixed results.

165 Scoring 0.5 on the PHQ-4 anxiety was not significantly associated with a different level of STAI-5
166 compared to scoring 0 ($\text{BF} = 1.83$), but was with scores of 1 ($\text{BF} > 30$). Similarly, there was no
167 evidence that scoring 1.5 was different from scoring 1 ($\text{BF} = 0.605$), but strong evidence that it
168 was different from scoring 2 ($\text{BF} > 30$).

169 ***Correlation Differences***

170 While the relationship pattern (i.e., the slope of the linear relationship) was not affected by
171 the PHQ-4 version, we focused next on testing the difference in the strength (i.e., the precision) of
172 the relationship, in particular at the lower end of the spectrum (i.e., for sub-clinical threshold
173 scores of the BDI-II and STAI-5). We bootstrapped (2000 iterations) the difference in correlation
174 between the refined and the original version for each of the two ground-truth measures, separately
175 for the BDI-II subsamples (minimal to mild ≤ 18 ; moderate to severe > 18) and the STAI-5
176 subsamples (minimal to mild < 2 ; moderate to severe ≥ 2).

177 The results suggested that in the subclinical range of the BDI-II, the correlation between
178 its score and the PHQ-4 Depression score was marginally higher (although not significantly,
179 $p_{\text{one-sided}} = 0.164$) for the refined version compared to the original one. No correlation
180 differences were observed in the moderate to severe range of the BDI-II.

181 For the STAI-5, there was no difference in the correlation between the refined and the
182 original version in the subclinical range of the STAI-5. Surprisingly, we observed a stronger
183 correlation between the refined PHQ-4 Anxiety score and the STAI-5 in the moderate to severe
184 range compared to the original version ($p_{one-sided} = 0.017$).

185 ***Predictive Power***

186 Finally, we tested the predictive power of the PHQ-4 depression and anxiety scores on the
187 presence of a depression or anxiety disorder, respectively. We modeled the relationship with a
188 logistic regression. While the PHQ-4 was overall a strong predictor of the outcome, there was no
189 significant difference between the two PHQ-4 versions.

190 However, the ROC curves for the refined and the original version of the PHQ-4, suggested
191 that the refined version had a better sensitivity / specificity trade-off (AUC = 78.36%) compared to
192 the original version (AUC=75%), in particular on the lower end of the spectrum. The difference
193 was negligible for anxiety.

194 **Discussion**

195 These results suggest that the new “Once or twice” response option to the PHQ-4 does
196 help capturing more fine-grained variations of depressive symptoms, particularly in the
197 subclinical range. Importantly, adding this new response option with the scoring of 0.5 does not
198 disrupt the quality of the scale, which scores remain comparable to that of the original version.

199 The results for the anxiety subscale appear more mixed, with less evident benefits.
200 However, this might have been partly caused by our design decision regarding the questionnaire
201 used for the ground-truth measure of anxiety. Indeed, we used the abridged version of the STAI,
202 which only included 5 items, arguably limiting the sensitivity of the anxiety measure in the first
203 place.

204 Finally, although we used a stricter criterion for classifying participants as having a
205 depression or an anxiety disorder by restricting it to participants also reporting undergoing a
206 medical treatment, it was still based on self-reported data. Studies in controlled clinical settings
207 are needed to confirm the potential benefits of the refined PHQ-4 in mood disorders detection

208 accuracy.

209 **General Discussion**

210 The objective of this study was to test the introduction of a “Once or twice” response
211 option to the PHQ-4 to enhance its sensitivity to milder mood fluctuations. In the first study, we
212 showed that the new response option was used prevalently by participants and did capture a
213 unique portion of the depression and anxiety underlying dimensions. In the second study, we
214 showed that the refined version of the PHQ-4 was able to better differentiate lower levels of
215 depression compared to the original version, while remaining comparable. Although the benefits
216 of this refinement appear to be fairly minor, and particularly marked for the depression score
217 compared to anxiety, this cost-free improvement appear useful to implement when measuring
218 depression and anxiety using the PHQ-4 ultra-short screening questionnaire.

219 **Acknowledgements**

220 We would like to thank the dissertation students from the University of Sussex for their
221 help in data collection.

222 **References**

223 Beck, A. T., Steer, R. A., Brown, G. K., et al. (1996). *Beck depression inventory*.

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

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

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

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

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

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

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

244 Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and
245 exploring the parameters of statistical models using r. *Journal of Open Source Software*,
246 5(53), 2445.

247 Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance:
248 An r package for assessment, comparison and testing of statistical models. *Journal of Open
249 Source Software*, 6(60).

250 Lüdecke, D., Waggoner, P. D., & Makowski, D. (2019). Insight: A unified interface to access
251 information from model objects in r. *Journal of Open Source Software*, 4(38), 1412.

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

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

258 Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The multidimensional
259 assessment of interoceptive awareness, version 2 (MAIA-2). *PloS One*, 13(12), e0208034.

260 Mendoza, N. B., Frondozo, C. E., Dizon, J. I. W. T., & Buenconsejo, J. U. (2022). The factor
261 structure and measurement invariance of the PHQ-4 and the prevalence of depression and

262 anxiety in a southeast asian context amid the COVID-19 pandemic. *Current Psychology*, 1–10.

263 Morey, R. D., & Rouder, J. N. (2024). *BayesFactor: Computation of bayes factors for common*
264 *designs*. <https://CRAN.R-project.org/package=BayesFactor>

265 Murphy, J., Brewer, R., Plans, D., Khalsa, S. S., Catmur, C., & Bird, G. (2020). Testing the
266 independence of self-reported interoceptive accuracy and attention. *Quarterly Journal of*
267 *Experimental Psychology*, 73(1), 115–133.

268 Patil, I., Makowski, D., Ben-Shachar, M. S., Wiernik, B. M., Bacher, E., & Lüdecke, D. (2022).
269 Datawizard: An r package for easy data preparation and statistical transformations. *Journal of*
270 *Open Source Software*, 7(78), 4684.

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

273 R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for
274 Statistical Computing. <https://www.R-project.org/>

275 Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp.
276 85–100). Springer.

277 Spielberger, C. D. (1970). Manual for the state-trait anxiety inventory (self-evaluation
278 questionnaire). (*No Title*).

279 Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdecke, D., Wiernik, B. M., & Makowski, D. (2024).
280 Check your outliers! An introduction to identifying statistical outliers in r with easystats.
281 *Behavior Research Methods*, 56(4), 4162–4172.

282 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G.,
283 Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open*
284 *Source Software*, 4(43), 1686.

285 Zsido, A. N., Teleki, S. A., Csokasi, K., Rozsa, S., & Bandi, S. A. (2020). Development of the
286 short version of the spielberger state—trait anxiety inventory. *Psychiatry Research*, 291,
287 113223.

Figure 1

A) Proportion of answers of each type to the four items. 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).

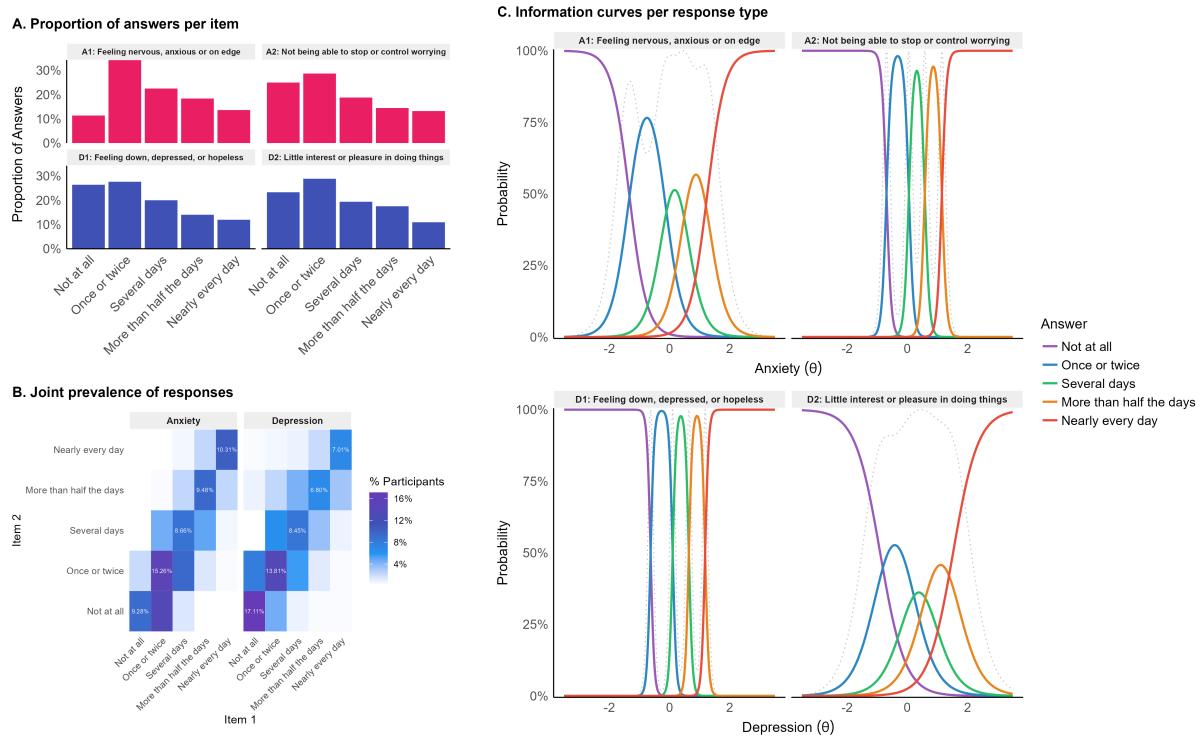


Figure 2

A) *PHQ-4 depression and anxiety scores against their respective ground-truth measures, the BDI-22 and the STAI-5*. Bayes factors in grey tell if there is a difference, for the same *PHQ-4* score, between the original and the refined version ($BFs < 1$ suggest no difference and thus evidence for a comparability of the refined version with respect to the original scale. Bayes factors in yellow represent how new in-between scores (0.5, 1.5, 2.5, ...) available with refined version differ from the adjacent scores ($BFs > 3$ suggest that half a point of difference on the refined *PHQ-4* relates to a significant difference on the ground truth measure). $BF < 1/3^\circ$, $BF > 3^*$, $BF > 10^{**}$, $BF > 30^{***}$. B) Bootstrapped distributions of the difference of correlation between the revised *PHQ-4* scores and the original one for sub-clinical threshold scores of depression and anxiety. Positive differences suggest that the correlation between the ground-truth measure and the refined *PHQ-4* score was stronger compared to the original version. C) Predictive power of the *PHQ-4* scores on the presence of a depression or anxiety disorder. The upper plots show the relationship modelled by a logistic regression, while the above plots represent the ROC curves (in which a line further away from the diagonal represents a higher combination of sensitivity and specificity).

