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

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

¹School of Psychology, University of Sussex

²Sussex Centre for Consciousness Science, University of Sussex

³School of Social Sciences, Nanyang Technological University

⁴LKC Medicine, Nanyang Technological University

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

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

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

The Patient Health Questionnaire-4 (PHQ-4) is an ultra ²⁸ brief measurement of core signs of depression and anxiety ²⁹ (Kroenke et al., 2009). It consists of two items for depres- ³⁰ sion (PHQ-2, Kroenke et al., 2003) and anxiety (GAD-2, ³¹ Kroenke et al., 2007), each corresponding to DSM-5 diagnos- ³² tic symptoms for major depressive disorder (MDD) and gen- ³³ eralized anxiety disorder (GAD). It has been validated across ³⁴ many languages and populations (Christodoulaki et al., 2022; ³⁵ Materu et al., 2020; Mendoza et al., 2022), becoming one of ³⁶ the most popular screening instruments for depression and ³⁷ anxiety (Maurer et al., 2018).

10

12

14

16

17

18

20

21

22

25

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 39 discriminate and identify potential MDD/GAD patients. This discriminative goal materializes in the scale's design and the 40 existence of categorical cut-offs, which does not necessary entail a focus on the sensitivity to milder mood alterations. In 41 particular, the gap between the two lowest possible answers, 42 "Not at all" and "Several days", is quite large and possibly 43 leaves out the possibility of more subtle occurrences. While 44 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 quantification.

This brief report aims at testing the possibility of enhanc- $_{49}$ ing - with minimal changes to the original scale - the PHQ-4 $_{50}$

sensitivity to mild mood level inflections. In the first study, we will evaluate whether the new response option is prevalently used by participants, and whether it does capture a specific part of the latent measure. In the second study, we will compare the refined PHQ-4 version to the original one in terms of sensitivity to subclinical variability in depression, using the Beck Depression Inventory (BDI-II, Beck et al., 1996) and the State Trait Anxiety Inventory (STAI-5, Zsido et al., 2020) as our ground-truth measures of depression and anxiety.

Study 1

Method

Participants

The sample consists of 485 English-speaking participants (Mean age = 30.1 ± 10.1 [18, 73]; 50.3% females) 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. 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.

53

55

57

59

60

63

64

65

66

69

71

72

73

90

100

101

102

103

104

105

106

107

108

109

110

112

Measures

In the original PHQ-4, the instructions "Over the last 2 76 weeks, how often have you been bothered by the following 77 problems?" are followed with 4 items (A1 - Feeling nervous, 78 anxious or on edge; A2 - Not being able to stop or control 79 worrying; D1 - Little interest or pleasure in doing things; D2 80 - Feeling down, depressed, or hopeless). The original answer 81 options are "Not at all" (0), "Several days" (1), "More than half the days" (2), "Nearly every day" (3). The total score is 82 computed by summing the responses of each facet resulting 83 in a 0-6 score for depression and anxiety.

For the refined version, we added a "Once or twice" op- 85 tion between "Not at all" and "Several days"in order to bet- 86 ter capture potential mild mood inflections (see Dobson & 87 Mothersill, 1979 for the choice of the label). This new option 88 was scored as 0.5 to preserve the same scoring as the original version.

Procedure

Participants were administered the refined PHQ-4 on-93 line as part of another study, which contained additional 94 questionnaires and tasks not relevant fort the current analy-95 sis. The PHQ-4 was presented in a randomized order with 96 other questionnaires. The data is available in open-access at 97 https://github.com/RealityBending/IllusionGameReliability. 98

- Dominique Makowski
- An Shu Te
- Ana Neves
- S.H. Annabel Chen

This preprint is a non-peer-reviewed work from the **Reality Bending Lab**.



Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: 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, Writing – original draft; Ana Neves: Data curation, Formal
Analysis, Writing – review & editing; S.H. Annabel Chen: Project119
administration, Supervision, Writing – review & editing

Correspondence concerning this article should be addressed to₁₂₁ Dominique Makowski, Email: D.Makowski@sussex.ac.uk

Results

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

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 1) 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 1). 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 did cover a selective and unique portion of the latent space.

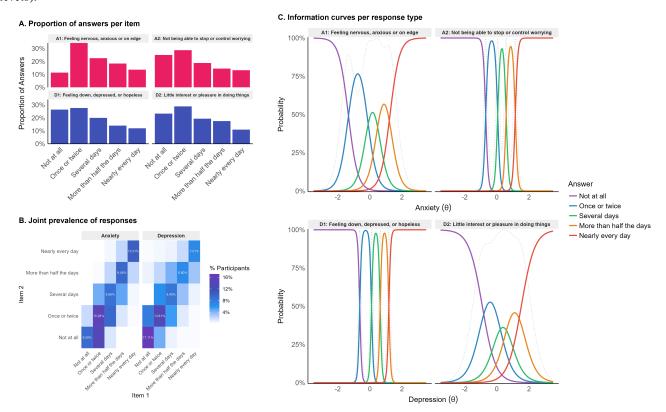
Discussion

The fact that the new "Once or twice" response option was the most prevalent response speaks to its usefulness in capturing more accurately participants' expression. The IRT analysis further revealed that this response tracks 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).

One natural methodological limitation pertains to the interpretation of the latent dimensions in an IRT framework applied to pairs of items. In this study's context, references

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).



144

145

147

148

to for instance "the latent anxiety dimension" merely corre-140 sponds to the amalgamation of the two items of the anxiety 141 subscale, and not to a more general and valid true anxiety 142 factor.

Study 2

Method

123

124

125

129

130

132

134

135

136

138

Participants

The initial sample consisted of 1053 participants, recruited 150 (181 were recruited on *Prolific*, 772 students from the Uni- 151 versity of Sussex via *SONA*, and the rest through convenience 152 sampling as part of dissertation students' data collection). We153 used attention checks as the primary target for participant ex- 154 clusion. We excluded 194 participants (18.42%) for failing at least one attention check, and 23 (2.18%) that were out- 155 liers ($|z_{robust}| > 2.58$) on measures significantly related to the probability of failing attention checks (namely, the stan- 156 dard deviation of all the items of the IAS, as well as the the157

multivariate distance obtained with the OPTICS algorithm, see Thériault et al., 2024). The experiment duration was not related to the probability of failing attention checks and was thus not used as an exclusion criterion.

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

In this sample, 51 participants (6.10%) were labelled as having Depression, as indexed by the self-reported presence of MDD together with the use of a treatment (antidepressent, anxiolytic and/or therapy), and 87 participants (10.41%) were labelled as having Anxiety, as indexed by the self-reported presence of GAD or Panic Disorder, also together with the use of a treatment.

Measures

Participants were randomly assigned to complete either the original or refined version of the PHQ-4, which included one

160

161

162

164

166

167

168

169

170

171

172

173

174

175

176

177

179

180

181

183

185

186

187

188

189

190

191

192

193

194

196

197

198

199

200

201

202

203

204

205

207

additional response option ("Once or twice") scored 0.5 (cre-209 ating more possible total scores - 0.5, 1.5, 2.5, etc.).

Beck's Depression Inventory (BDI-II, Beck et al., 1996)₂₁₁ was used as a ground truth measure of depressive symptoms.₂₁₂ It includes 21 items, each addressing a specific depression₂₁₃ symptom and offering four response options scored from 0₂₁₄ to 3. Participants are instructed to select the option that best₂₁₅ describes how they have felt over the past two weeks. The to-₂₁₆ tal score is calculated by summing the scores for all 21 items,₂₁₇ with higher scores indicating greater severity of depressive₂₁₈ symptoms.

The short version of the State-Trait Anxiety Inventory 220 (STAI-5, Zsido et al., 2020) was used as a ground truth mea-221 sure of anxiety. This abridged version of the STAI (Spiel-222 berger, 1970) includes 5 items rated on a 4-point Likert scale. 223 Changes were made in the instructions from asking "how par-224 ticipants feel right now" to "over the past 2 weeks" to keep it 225 consistent with the instructions of the PHQ-4 and BDI-II. A 226 general score of anxiety was computed by averaging all the

Participants were also asked to complete two questionnaires of interoception, namely the Interoceptive Accuracy²²⁸ Scale (IAS - 21 items rated on analog scales, Murphy et al.,²²⁹ 2020) and the Multidimensional Assessment of Interoceptive²³⁰ Awareness (MAIA-2 - 37 items, Mehling et al., 2018).

After demographic questions, participants were asked to²³² report the current presence of psychiatric issues (from a list),²³³ as well as the usage of treatment (antidepressants, mood sta-²³⁴ bilizers, anxiolytics, therapy). We indexed the presence of a²³⁵ depression when participants reported suffering from either²³⁶ Major Depressive Disorder (MDD) or Dysthymia, as well as²³⁷ undergoing a medical treatment. Similarly, we indexed the²³⁸ presence of an anxiety disorder when participants reported²³⁹ suffering from either Generalized Anxiety Disorder (GAD)²⁴⁰ or Panic Disorder, as well as undergoing a medical treatment.²⁴¹

Procedure

The original or refined version of the PHQ-4 was followed²⁴⁴ by the BDI-II, STAI-5, IAS, and MAIA-2, presented in ran-245 dom order. The IAS and the MAIA-2 were included as part246 of another study focused on interoception, and were only used²⁴⁷ in this stud as part of data quality control checks. ## Results 248 As all scripts, analysis details the at²⁴⁹ results tables are available open-access https://github.com/DominiqueMakowski/PHQ4R, we will 250 focus on reporting the main results.

PHQ-4 Depression vs. BDI-II

The linear regression predicting the BDI-II total score₂₅₄ with the PHQ-4 depression score showed no interac-₂₅₅ tion related to the PHQ-4 version ($\Delta Intercept_{refined}$ =₂₅₆ -0.13, 95% CI [-1.73, 1.47], t(832) = -0.16, p =₂₅₇ 0.871; $\Delta \beta_{refined}$ = -0.05, 95% CI [-0.70, 0.60], t(832) =₂₅₈

-0.15, p = 0.883), suggesting no differences in the relationship pattern between the two versions (see Figure 2).

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

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

PHQ-4 Anxiety vs. STAI-5

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

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

However, comparing in-between scores with adjacent scores yielded mixed results. Scoring 0.5 on the PHQ-4 anxiety was not significantly associated with a different level of STAI-5 compared to scoring 0 (BF = 1.83), but was with scores of 1 (BF > 30). Similarly, there was no evidence that scoring 1.5 was different from scoring 1 (BF = 0.605), but strong evidence that it was different from scoring 2 (BF > 30).

Correlation Differences

242

243

251

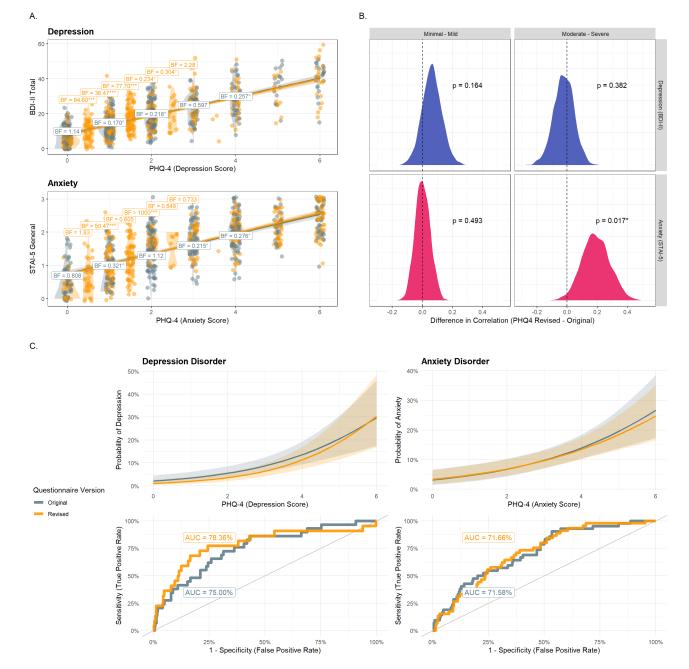
252

253

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

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 fo 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°, 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).



to severe > 18) and the STAI-5 subsamples (minimal to mild $_{308}$ < 2; moderate to severe >= 2).

The results suggested that in the subclinical range of the ³⁰⁹ BDI-II, the correlation between its score and the PHQ-4 De-³¹⁰ pression score was marginally higher (although not signifi-³¹¹ cantly, $p_{one-sided} = 0.161$) for the refined version compared ³¹² to the original one. No correlation differences were observed ³¹³ in the moderate to severe range of the BDI-II.

For the STAI-5, there was no difference in the correlation ³¹⁵ between the refined and the original version in the subclini-³¹⁶ cal range of the STAI-5. Surprisingly, we observed a stronger ³¹⁷ correlation between the refined PHQ-4 Anxiety score and the ³¹⁸ STAI-5 in the moderate to severe range compared to the original version ($p_{one-sided} = 0.027$).

Predictive Power

Finally, we tested the predictive power of the PHQ-4 de-³²³ pression and anxiety scores on the presence of a depression or anxiety disorder, respectively. We modeled the relation-³²⁴ ship with a logistic regression. While the PHQ-4 was overall³²⁵ a strong predictor of the outcome, there was no significant difference between the two PHQ-4 versions.

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

Discussion

These results suggest that the new "Once or twice" re-336 sponse option to the PHQ-4 does help capturing more fine-337 grained variations of depressive symptoms, particularly in the 338 subclinical range. Importantly, adding this new response op-339 tion with the scoring of 0.5 does not disrupt the quality of the 340 scale, which scores remain comparable to that of the original 341 version.

The results for the anxiety subscale appear more mixed,343 with less evident benefits. However, this might have been344 partly caused by our design decision regarding the question-345 naire used for the ground-truth measure of anxiety. Indeed,346 we used the abridged version of the STAI, which only in-347 cluded 5 items, arguably limiting the sensitivity of the anxiety 348 measure in the first place.

Finally, although we used a stricter criterion for classify-350 ing participants as having a depression or an anxiety disor-351 der by restricting it to participants also reporting undergoing 352 a medical treatment, it was still based on self-reported data.353 Studies in controlled clinical settings are needed to confirm354 the potential benefits of the refined PHQ-4 in mood disorders355 detection accuracy.

General 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. In the first study, we showed that the new response option was used prevalently by participants and did capture a unique portion of the depression and anxiety underlying dimensions. In the second study, we showed that the refined version of the PHQ-4 was able to better differentiate lower levels of depression compared to the original version, while remaining comparable. Although the benefits of this refinement appear to be fairly minor, and particularly marked for the depression score compared to anxiety, this cost-free improvement appear useful to implement when measuring depression and anxiety using the PHQ-4 ultra-short screening questionnaire.

Acknowledgements

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

References

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

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 socioeconomic 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 Affective₄₁₀ Disorders, 122(1-2), 86-95.
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D.412 359 (2020). Extracting, computing and exploring the parame-413 360 ters of statistical models using r. Journal of Open Source414 361 Software, 5(53), 2445. 362

358

363

364

365

368

374

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

397

398

399

401

403

405

407

- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., &416 Makowski, D. (2021). Performance: An r package for₄₁₇ assessment, comparison and testing of statistical models.418 Journal of Open Source Software, 6(60).
- Lüdecke, D., Waggoner, P. D., & Makowski, D. (2019),420 367 Insight: A unified interface to access information from₄₂₁ model objects in r. Journal of Open Source Software, 422 369 4(38), 1412.
- Materu, J., Kuringe, E., Nyato, D., Galishi, A., Mwanam-424 371 sangu, A., Katebalila, M., Shao, A., Changalucha, J., 372 Nnko, S., & Wambura, M. (2020). The psychometric 373 properties of PHQ-4 anxiety and depression screening scale among out of school adolescent girls and young 375 women in tanzania: A cross-sectional study. BMC Psy-376 chiatry, 20(1), 1-8. 377
- Maurer, D. M., Raymond, T. J., & Davis, B. N. (2018). 378 Depression: Screening and diagnosis. American Family 379 Physician, 98(8), 508-515. 380
 - Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The multidimensional assessment of interoceptive awareness, version 2 (MAIA-2). PloS One, 13(12), e0208034.
 - 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.
 - Morey, R. D., & Rouder, J. N. (2024). BayesFactor: Computation of bayes factors for common designs. https: //CRAN.R-project.org/package=BayesFactor
 - Murphy, J., Brewer, R., Plans, D., Khalsa, S. S., Catmur, C., & Bird, G. (2020). Testing the independence of selfreported interoceptive accuracy and attention. Quarterly *Journal of Experimental Psychology*, 73(1), 115–133.
 - Patil, I., Makowski, D., Ben-Shachar, M. S., Wiernik, B. M., Bacher, E., & Lüdecke, D. (2022). Datawizard: An r package for easy data preparation and statistical transformations. Journal of Open Source Software, 7(78), 4684.
 - 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.
 - R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
- Samejima, F. (1997). Graded response model. In *Handbook* 408 of modern item response theory (pp. 85–100). Springer. 409

- Spielberger, C. D. (1970). Manual for the state-trait anxiety inventory (self-evaluation questionnaire). (No Title).
- Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdecke, D., Wiernik, B. M., & Makowski, D. (2024). Check your outliers! An introduction to identifying statistical outliers in r with easystats. Behavior Research Methods, 56(4), 4162-4172.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.
- Zsido, A. N., Teleki, S. A., Csokasi, K., Rozsa, S., & Bandi, S. A. (2020). Development of the short version of the spielberger state—trait anxiety inventory. Psychiatry Research, 291, 113223.