**Figure captions**

**Fig. 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).

**Fig. 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)