Supplementary material

Supplementary Methods

IRT model

IRT is a technique used to evaluate the psychometric performance of test instruments by estimating where on the scale each item is most informative. IRT achieves this by describing the latent trait underlying the scale (θ) as a standardised distribution with a mean of 0 and a standard deviation of 1. Typically, the range of θ lies between -4 and +4. The distribution of θ, is used to estimate the location (β) along the scale where each item is most informative, i.e. shows most discrimination between high and low latent trait levels. Beta is commonly referred to as item difficulty. Regardless of difficulty, some items provide more precise estimates of the latent trait than others (α). Alpha is commonly referred to as item discrimination. In practice, $\alpha = 0$ indicates no discrimination, whilst $\alpha > 2$ indicates high discrimination (Baker, 2001).

The PHQ-ADS scale was analysed using a graded response IRT model for ordinal data. In graded-response models, items vary in their difficulty β and discrimination α. The item difficulty β indicates the level of the latent trait needed to score high on the item (DeMars, 2010). Discrimination measures how well an item can distinguish between contiguous latent trait levels near the inflection point of a boundary characteristic curve. The respondent's individual latent trait levels being measured are referred to as theta (θ) in IRT. In the graded-response IRT model, each item was modelled with its own discrimination parameter α, difficulty parameter β and cutpoints k that identify boundaries between the ordered item ratings (Samejima, 1969). For example, the probability p of observing rating k or higher for item i and person j is given by the following equation:

$$p\left(Y_{ij} \geq k \mid \theta\right) = \frac{exp^{\alpha_i(\theta_j - \beta_{ik})}}{1 + exp^{\alpha_i(\theta_j - \beta_{ik})}} \quad \theta \sim N(0,1)$$

where α_i is the discrimination of item i, β_{ik} is the k^{th} cutpoint for item i, and θ_j is the level of the latent trait of person j. The β_{ik} corresponds to the difficulty of responding with rating k or higher for item i. The dependent variable is the ordinal response (0,1,2,3), and the independent variables are the person's trait level θ, the item difficulty β_{ik}, the cutpoint k, and the item discrimination parameter α_i. The independent variables combine accumulatively and each item's difficulty β_{ik} is subtracted from θ. That is the ratio of the probability of rating an item at the cutpoint k or higher for a person to the probability of rating an item less than the cutpoint, where a logistic function provides the probability that rating any item i at the cutpoint k or higher is independent from the outcome of any other item, controlling for person parameters θ and item parameters, i.e. the item difficulty β_{ik}, and the item discrimination α_i.

For each item, an item response function (IRF) was calculated which calibrates the responses of an individual against each item at each cutpoint. A calibrated standardized score for trait severity θ was returned for each item and each cutpoint and plotted as a boundary characteristic curve (BCC) along a standardized scale with a mean of 0. Boundary characteristic curves for all items from the PHQ-ADS scale are depicted in Figure 3. From the BCC two parameters were estimated: The first is the value of θ at which the likelihood of rating an item at or above the cutpoint is 0.5, interpreted as 'expressed latent symptom severity'. The second is the slope of the curve from the point at which the likelihood of rating an item at or above the cutpoint is 0.5, interpreted as 'expressed item discrimination' i.e., the ability to discriminate between greater and lesser severity scores. The IRF may also be expressed as an item information function (IIF) curve which displays the relationship between severity θ and discrimination α. The IIFs for each item from the PHQ-ADS scale is depicted in Figure 2. The apices of the curve for any IIF indicate the values of θ at which there is maximum discrimination. If an item displays more than one apex, it shows high

discrimination at multiple cutpoints of the rating scale. Scales expressing a range of θ values are more informative than those with items clustering around a single value. By convention, items with a discrimination of score of > 1.7 are considered informative, although lower values are considered contributory within context (Baker, 2001). Statistical assumptions underlying the IRT principles of scalability and item-independence were examined.

Mokken scale analysis

A Mokken scale analysis was conducted to investigate the fit of the data from the PHQ-ADS scale to the monotonely homogeneous Mokken model. Mokken scale analysis is a non-parametric IRT analysis that can be used to reduce the number of questionnaire items based on assumptions of unidimensionality, local independence, and latent monotonicity (Sijtsma and van der Ark, 2017). Mokken scale analysis was used to optimize the unidimensionality of the scale i.e. to evaluate the contribution of each item to the single, underlying, latent concept (Crichton, 1999). Loevinger H scalability coefficients and the Guttman errors (observed and expected) were calculated for each item and all the other items of the PHQ-ADS scale using STATA's loevH package (Hardouin, 2004). H coefficients follow a standard normal distribution ranging between 0 and 1. By convention H ≈ 0.3 indicates weak scalability whilst H ≈ 0.6 indicates strong scalability. A Guttman error occurs when a participant rates one item high, but rates another item low. Loevinger coefficients compare the observed Guttman errors to the expected number of Guttman errors if the items would be unrelated (Crichton, 1999).

Supplementary Results

Internal reliability

Internal Reliability along Theta

16-item PHQ-ADS scale

In IRT, reliability may be calculated at multiple point values of θ along the continuum rather than a single reliability score as in CTT. Reliability may be is defined at different points of θ with the mean of θ fixed at 0 and the variance at 1, facilitating identification of the model and reliability for all points along the θ continuum, distinguishing respondents according to specific values of θ (Thissen, 2000). For the 16-item PHQ-ADS scale, there is reliable information to differentiate respondents who show no or minimal symptoms (θ = 0; Reliability = 0.893), considered very good for reliability. There is high reliability for respondents with mild symptoms (θ = 1; Reliability = 0.961), high reliability for respondents with moderate symptoms (θ = 2; Reliability = 0.971) as well as high reliability for respondents with moderate to severe symptoms (θ = 3; Reliability = 0.939). However, reliability then decreases for respondents with severe symptoms (θ = 4; 0.794). These findings suggest that the highest reliability of measuring symptoms is at the mild to moderate amount of anxious depressive symptomatology, θ = 1, 2, or 3. Thereafter, reliability reduces so that the extreme ends of the continuum, θ = 4; -1; -2; -3, -4, are no longer reliably measured (see Supplementary Table 1). In summary, the PHQ-ADS scale was not reliable for θ < 0.

Revised scales

For the revised 8-item PHQ-ADS scale, reliability was high (> 0.9) for θ = 1; 2, very good (> 0.85) for θ = 0; 1, but not acceptable (< 0.45) at the extreme ends of the continuum with θ = 4; -1; -2; -3; -4. For the revised 7-item PHQ-ADS scale, reliability was high (> 0.9) for θ = 1; 2, moderate (> 0.80) for θ = 0; 1, but not acceptable (< 0.45) at the extreme ends of the continuum with θ = 4; -1; -2; -3; -4. These findings suggest that there is poor reliability at the extremes of the scale score and the scale was not informative in participants without

symptoms and participants with severe symptoms. Reliability statistics are displayed in Supplementary Table 1.

Supplementary Table 1. Reliability of the 16-item, 8-item and 7-item PHQ-ADS scales at values of Θ

Θ	16-item PHQ-ADS scale			8-item PHQ-ADS scale			7-item PHQ-ADS scale		
	TIF	**TIF SE**	**Reliability**	**TIF**	**TIF SE**	**Reliability**	**TIF**	**TIF SE**	**Reliability**
-4	1.01	1.00	0.01	1.01	1.00	0.01	1.00	1.00	0.00
-3	1.05	0.98	0.04	1.03	0.99	0.03	1.00	1.00	0.00
-2	1.23	0.90	0.18	1.12	0.95	0.10	1.02	0.99	0.02
-1	2.16	0.68	0.54	1.58	0.80	0.37	1.22	0.90	0.18
0	9.39	0.33	0.89	7.12	0.37	0.86	5.97	0.41	0.83
1	25.85	0.20	0.96	18.77	0.23	0.95	19.24	0.23	0.95
2	34.02	0.17	0.97	27.66	0.19	0.96	30.28	0.18	0.97
3	16.55	0.25	0.94	6.74	0.39	0.85	6.02	0.41	0.83
4	4.87	0.45	0.79	1.82	0.74	0.45	1.74	0.76	0.43

Note: Θ: latent trait, TIF = Test Information Function, SE = Standard Error.

Internal Reliability using Cronbach's alpha coefficient

A reliability analysis was carried out on the 16-item PHQADS scale using Cronbach's alpha coefficient. Cronbach's alpha showed the scale to reach high reliability, $\alpha = 0.915$. Most items appeared to be worthy of retention, resulting in a decrease in the alpha if deleted. The one exception to this was the item assessing sleep, which would increase the alpha to $\alpha = 0.917$. Overall, the scale showed high internal reliability.

Statistical assumptions

Item independence

Spearman-rank order correlation coefficients were used to assess initial item independency. All items were significantly correlated ($p < .001$ after Bonferroni-correction) with correlation coefficients ranging from 0.693 for the association between "Little interest or pleasure in doing things" and "Feeling down, depressed, or hopeless" to 0.213 for the

association between “Trouble falling or staying asleep, or sleeping too much” and “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”. The majority of values were lower than 0.50, suggesting basic local item independence. A residual coefficient matrix, requested after estimation of a single-factor factor analysis showed that no residuals were too highly correlated, all $R < 0.20$ (Yen, 1993), with the highest residual correlation between tiredness and sleep ($R = 0.1996$), suggesting basic item independence.

References

Baker, F.B., 2001. The basics of item response theory. ERIC.

Crichton, N., 1999. Information point Mokken Scale Analysis. J. Clin. Nurs. 8, 388.

DeMars, C., 2010. Item response theory. Oxford University Press.

Hardouin, J.-B., 2004. LOEVH: Stata module to compute Guttman errors and Loevinger H coefficients.

Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. Psychom. Monogr. Suppl.

Thissen, D., 2000. Reliability and measurement precision.

Yen, W.M., 1993. Scaling Performance Assessments: Strategies for Managing Local Item Dependence. J. Educ. Meas. 30, 187–213. https://doi.org/https://doi.org/10.1111/j.1745-3984.1993.tb00423.x