Journal of Affective Disorders 347 (2024) 335–344

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

A psychometric evaluation of the 16-item PHQ-ADS concomitant anxiety and depression scale in the UK biobank using item response theory

Chris Patrick Pflanz *, John Gallacher, Sarah Bauermeister

Dementias Platform UK, Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford OX3 7JX, United Kingdom

ARTICLE INFO

Keywords:
Patient Health Questionnaire Anxiety and Depression Scale
Depression
Affective disorder
Psychometrics
Cohort study
Anxiety disorder

ABSTRACT

Background: The Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS) provides a reliable and valid measure of concomitant depression and anxiety. However, research on its psychometric efficiency and optimal scale length using item-response theory (IRT) has not been reported. This study aimed to optimize the length of the PHQ-ADS scale without losing information by discarding items that were a poor fit to the IRT model.

Methods: The UK Biobank is a large cohort study designed to investigate risk factors for a broad range of disease. PHQ-ADS data were available from $n = 152{,}826$ participants (age = 55.87 years; SD = 7.73; 56.4 % female), 30.4 % of the entire UK Biobank sample. Psychometric properties of the PHQ-ADS were investigated using a 2-parameter IRT and Mokken analysis. Item statistics included discrimination, difficulty and Loevinger H coefficients of monotonicity.

Results: In the entire 16-item scale, item discrimination ranged from 1.40 to 4.22, with the item 'worrying' showing the highest level of discrimination and the item 'sleep disturbance' showing the lowest. Mokken analysis showed that the 16-item PHQ-ADS scale could be reduced to a 7-item scale without loss of test information. The reduced scale comprised mainly items measuring cognitive-affective symptoms of anxiety/depression, whereas items measuring somatic symptoms were discarded. The revised scale showed high discrimination and scalability.

Limitations: Findings are limited by the use of cross-sectional data that only included the baseline online questionnaire, but not other waves.

Conclusions: IRT is a useful technique for scale reductions which serve the clinical and epidemiological need to optimize screening questionnaires to reduce redundancy and maximize information. A reduced-item 7-item PHQ-ADS scale reduces the response burden on participants in epidemiological research settings, without loss of information.

1. Introduction

Anxiety and depression show high comorbidity (Groen et al., 2020; Hanel et al., 2009; ter Meulen et al., 2021). Evidence from genetics (Ohi et al., 2020; Purves et al., 2020), neuroimaging (van Tol et al., 2021), and psychopharmacology (Bandelow, 2020; Nash and Nutt, 2007; Smith et al., 2023) suggests that a common trait of emotional vulnerability underlies both disorders. Psychometric data confirms this hypothesis with anxiety and depression scales being highly intercorrelated (Newman, 2022; Ryan et al., 2013). This is considered by some to represent an underlying common construct (Cosco et al., 2012; Kroenke et al., 2016). Understanding the emotional vulnerability underlying common mental disorders has important public health implications at the population level. It is important, therefore, to be able to assess emotional vulnerability at-scale. The Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS) has been developed as a composite measure of depression and anxiety (Kroenke et al., 2016). The PHQ-ADS combines the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder (GAD-7) scale with the overall objective of assessing concomitant symptoms of depression and anxiety (Kroenke et al., 2016). The

Abbreviations: DPUK, Dementias Platform UK; GAD-7, Generalized Anxiety Disorder 7 scale; IRT, Item-response theory; PHQ-9, Patient Health Questionnaire 9; PHQ-ADS, Patient Health Questionnaire Anxiety and Depression Scale; REC, Research Ethics Committee.

* Corresponding author at: Department of Psychiatry, Warneford hospital, Oxford OX3 7JX, United Kingdom.
E-mail address: patrick.pflanz@psych.ox.ac.uk (C.P. Pflanz).

https://doi.org/10.1016/j.jad.2023.11.067
Received 6 June 2023; Received in revised form 7 November 2023; Accepted 18 November 2023
Available online 22 November 2023

PHQ-ADS is a suitable candidate, therefore, as a measure of emotional vulnerability. The PHQ-9 was designed based on the Diagnostic and Statistical Manual (DSM-IV) diagnostic criteria of depression (Kroenke et al., 2001). The GAD-7 was developed as a clinical measure for assessing generalized anxiety disorder in the primary care setting (Spitzer et al., 2006). Although initially designed as a screening tool for generalized anxiety, the GAD-7 also performs well as a screening tool for three other common anxiety disorders including social anxiety disorder, panic disorder, and posttraumatic stress disorder (Kroenke et al., 2007).

As a combination of the PHQ-9 and the GADS-7, unsurprisingly, the 16 item PHQ-ADS has good psychometric properties. Data from three trials show the PHQ-ADS to have high internal reliability (Cronbach's α = 0.8–0.9) and high construct and convergent validity (Kroenke et al., 2016). Of interest is that factor analysis of these data shows a unidimensional structure to provide the best fit (Kroenke et al., 2016). However, these correlational metrics, drawn from classical test theory (CTT), assume that all items contribute equally to the underlying trait. As such they are poor indicators of test efficiency. Test efficiency is an important issue for assessing common mental disorders, particularly at the population level. Optimising for scale length reduces participant burden and assessment time; leading to lower attrition rates and better-quality data (Herzog and Bachman, 1981; Lavrakas, 2008). Shorter, simpler tests, whilst not sufficient for diagnosis, are important for the conduct of large studies and as screening tools. The PHQ-4 may be used as such a shorter screening tool for anxiety and depression and consists of the first two items from the GAD-7 and PHQ-9 (Kroenke et al., 2009). However, it was designed using the diagnostic categorical approach by using the two main diagnostic criteria from the DSM-5 (American Psychiatric Association, 2013) for depression (i.e. anhedonia and low mood) and anxiety (nervousness/anxiety/tension and inability to stop/control worrying). The PHQ-4 has, therefore, also not been optimized for psychometric efficiency.

To optimize the PHQ-ADS for scale length, we used item response theory (IRT) to identify the relative informativeness of specific items to the overall score. Unlike CTT, IRT does not assume additivity of item ratings; allowing items to vary in their contribution to the latent trait. In IRT the informativeness of each item for different levels of the latent trait may be assessed allowing less informative (redundant) items to be identified and omitted. Using data from UK Biobank, the psychometric qualities of the PHQ-ADS were evaluated using CTT, dimensionality was assessed by factor analysis, and scale efficiency was investigated using IRT.

2. Methods

2.1. Design

Details of the design, participants, procedure and ethics of UK Biobank are available elsewhere (Sudlow et al., 2015). In brief, UK Biobank is a large, population-based study of 502,665 participants designed to study a wide range of risk factors for a wide range of diseases of middle and old age. Ethical approval was granted to Biobank from the Research Ethics Committee - REC reference 11/NW/0382 (Sudlow et al., 2015). As part of an online follow-up mental health questionnaire, the PHQ-ADS was administered to all participants with known email addresses.

2.2. Materials

The PHQ-ADS is a combination of the items from the PHQ-9 and GAD-7 and assesses recent symptoms of depression/anxiety over the last two weeks (Kroenke et al., 2016). The PHQ-ADS consists of 16 items scored on a 4-point Likert scale: *Not at all (0), Several days (1), More than half the days (2), Nearly every day (3)*. PHQ-ADS scores range from 0 to 48 with higher scores indicating more severe depression/anxiety (Kroenke et al., 2016).

2.3. Statistical analysis

CTT-based psychometric properties of the PHQ-ADS were assessed to confirm previous reports. An exploratory factor analysis was used to investigate dimensionality of the PHQ-ADS scale. Eigen values were used to estimate the variance explained by a factor. The Kaiser criterion (eigen value>1.00) was used to retain factors. IRT and Mokken analyses were used for scale optimization using item discrimination α and Loevinger H monotonicity coefficients as indicators of item efficiency. Items were considered to have good item discrimination when $\alpha > 2$ (Baker, 2001) and to be reasonably monotonic when $H > 0.5$ (Crichton, 1999). Further details on the IRT model specification and Mokken scale analysis can be found in the supplementary methods.

Evidence for the validity of the optimized scale was obtained from the IRT analysis; compared its discrimination with the original PHQ-ADS scale and with the PHQ-4 scale when predicting mental health outcomes. Logistic regression models were fit with the respective scale as continuous predictor and the mental health outcomes as binary dependent variables. Quality metrics of the resulting classification were computed including the area under the curve (AUC) of the receiver operating characteristic (ROC) that serves as an overall indicator of classification. The mental health outcomes used for this analysis were: “Ever felt worried, tense, or anxious for most of a month or longer?”, “Seen GP for nerves, anxiety, tension, or depression?”, “Ever been offered/sought treatment for anxiety?”, “Ever been offered/sought treatment for depression?” and a diagnosis of major depressive disorder (MDD) and/or bipolar disorder.

We also conducted a secondary IRT analysis in a clinical subsample of the UK Biobank that only included participants who had experienced at least one depressive episode ($N = 11.653$) including participants with a single episode of major depressive disorder (MDD), recurrent MDD, bipolar disorder 1, or bipolar disorder 2. The purpose of this analysis was to evaluate the optimized PHQ-ADS scale in a clinical sample.

UK Biobank data for this analysis (application 15,008) were uploaded onto the Dementias Platform UK (DPUK) Data Portal (Bauermeister et al., 2019) and analysed using STATA SE 17.0 (StataCorp, 2021).

3. Results

3.1. Sample characteristics

Complete PHQ-ADS data were available for 152,826 UK Biobank participants. This sample was aged 38 to 72 years (M = 55.87 years; SD = 7.73), and was 56.4 % female. The sample was 97 % European and 49 % were educated to degree level (Table 1). This sample differed from the overall UK Biobank sample with included participant being on average younger (0.95 years $p < 0.001$), leaving fulltime education earlier (0.33 years, $p < 0.001$) and being less socioeconomically deprived (TDI = 0.50 p ⟨001) than excluded participants.

3.2. Preliminary analysis

The preliminary analysis, using classical test theory, found the internal reliability of the PHQ-ADS was high with a Cronbach's alpha of α = 0.92. This was comparable to previous research reporting values of between α = 0.8 to α = 0.9 (Kroenke et al., 2016).

An exploratory factor analysis (Table 2) found the 16-item PHQ-ADS scale to be unidimensional with only the first factor meeting the Kaiser criterion of an eigen value ≥1.00. Items from both the PHQ-9 (depression) and GAD-7 (anxiety) subscales showed high loadings on this factor. The scree plot of the factor loadings shows this unidimensionality clearly (Fig. 1).

3.3. IRT analysis

For the 16-item PHQ-ADS, items Loevinger's H coefficient for

Table 1
Descriptive statistics and percentages of sample characteristics and items from the PHQ-ADS scale.

Sample characteristic	Case numbers/percentages					
Sex	86,196 (56.4 %) female			66,630 (43.6 %) male		
Ethnicity	147,993	802	1254	1089	349	836
	97.16 %	0.53 %	0.82 %	0.71 %	0.23 %	0.55 %
	European	Mixed	Asian	African	Chinese	Other
Qualifications	69,504	20,497	30,128	5613	7665	7697
	49.26 %	14.53 %	21.35 %	3.98 %	5.43 %	5.45 %
	degree	A-level	O-Level	CSE	NVQ	Other

	Mean	SD	Min	Max	Not at all	Several days	Half the days	Nearly every day
Age	55.87	7.73	NA	NA	–			
Age completed full-time education	18.90	2.56	5	35	–			
TDI	−1.71	2.83	−6.26	11.00	–			
Anhedonia	0.24	0.57	0	3	124,460	22,594	3319	2453
					81.44 %	14.78 %	2.17 %	1.61 %
Low mood	0.27	0.56	0	3	119,396	28,338	3021	2071
					78.13 %	18.54 %	1.98 %	1.36 %
Sleep	0.71	0.90	0	3	78,342	51,816	10,728	11,940
					51.26 %	33.91 %	7.02 %	7.81 %
Tired	0.66	0.81	0	3	76,608	59,630	8303	8285
					50.13 %	39.02 %	5.43 %	5.42 %
Appetite	0.26	0.63	0	3	124,882	19,983	4222	3739
					81.72 %	13.08 %	2.76 %	2.45 %
Inadequacy	0.25	0.59	0	3	123,460	23,051	3216	3099
					80.78 %	15.08 %	2.10 %	2.03 %
Concentration	0.23	0.56	0	3	125,470	21,851	3069	2436
					82.10 %	14.30 %	2.01 %	1.59 %
Movement	0.07	0.35	0	3	144,386	6371	1176	893
					94.48 %	4.17 %	0.77 %	0.58 %
Suicidality	0.05	0.29	0	3	146,285	5321	679	541
					95.72 %	3.48 %	0.44 %	0.35 %
Nervousness	0.35	0.64	0	3	110,007	36,009	3432	3378
					71.98 %	23.56 %	2.25 %	2.21 %
Control worrying	0.31	0.63	0	3	116,933	28,798	3391	3704
					76.51 %	18.84 %	2.22 %	2.42 %
Multiple worries	0.40	0.67	0	3	104,382	40,388	3932	4124
					68.30 %	26.43 %	2.57 %	2.70 %
Relaxing	0.37	0.68	0	3	109,483	34,498	4325	4520
					71.64 %	22.57 %	2.83 %	2.96 %
Restlessness	0.15	0.47	0	3	134,691	14,484	2002	1649
					88.13 %	9.48 %	1.31 %	1.08 %
Irritability	0.32	0.59	0	3	111,199	36,412	3028	2187
					72.76 %	23.83 %	1.98 %	1.43 %
Foreboding	0.22	0.55	0	3	127,283	20,401	2541	2601
					83.29 %	13.35 %	1.66 %	1.70 %

	Mean	SD	Min	Max	None	Mild	Moderate	Severe
PHQ-ADS Sum score	4.87	6.46	0	48	127,547	19,050	4405	1824
					83.46 %	12.47 %	2.87 %	1.20 %
PHQ-9 sum score	2.75	3.69	0	27	121,671	22,426	5680	3049
					79.62 %	14.67 %	3.72 %	2.01 %
GAD-7 sum score	2.11	3.36	0	21	125,662	20,700	4149	2315
					82.22 %	13.56 %	2.71 %	1.52 %
12-item PHQ-ADS Theta	0.00	0.90	−1.07	3.76	–			
7-item PHQ-ADS Theta	0.02	0.83	−0.70	3.05	–			

Note: CSE: Certificate of Secondary Education or equivalent, NVQ: National Vocational Qualification, TDI: Townsend-deprivation index, theta: the underlying individual trait of emotional vulnerability from the IRT analysis, PHQ-9 scores of 5, 10, and 15 were used as cut-points for mild, moderate, and severe depression, respectively (Kroenke and Spitzer, 2002). GAD-7 scores of 5, 10, and 15 were used as cut-points for mild, moderate, and severe anxiety, respectively(Spitzer et al., 2006). PHQ-ADS cut-points of 10, 20, and 30 indicated mild, moderate, and severe levels of depression/anxiety, respectively (Kroenke et al., 2016).

monotonicity ranged between 0.42 and 0.58, indicating all items contributed to θ (Table 3). The discrimination of these estimates was weak to strong ranging between $\alpha = 1.40$ and $\alpha = 4.22$. For item difficulty, all items were most informative at the upper end of θ, with $\beta > 0$ for all cut points (k). These results indicate a relatively homogenous scale with redundant measurement at the upper end of θ. Of interest is that items indicating cognitive-affective symptoms of mood disorders and anxiety tended to cluster with higher monotonicity ($H > 0.5$) and higher discrimination ($\alpha > 3$). For example, the item measuring "Not being able to stop or control worrying" showed the highest discrimination value at $\alpha = 4.22$ ($H = 0.56$). In contrast, the item "Trouble falling or staying asleep, or sleeping too much", $\alpha = 1.40$ ($H = 0.44$), was the lowest. The items measuring anhedonia, low mood, inadequacy, nervousness, control worrying, multiple worries, difficulty relaxing all had high discrimination values ($\alpha > 2$) and monotonicity ($H > 0.5$). These items suggest a clustering of items around the latent trait of emotional vulnerability.

To assess the potential for more efficient assessment of emotional

Table 2
Exploratory factor analysis table showing eigenvalue characteristics and factor loadings for each item from the 16-item PHQ-ADS scale.

Item/statistic		Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Factor loadings	Eigenvalue	6.86	0.93	0.39	0.30	0.10	0.01
	Cumulative	0.89	1.01	1.07	1.10	1.12	1.12
Item loadings	Anhedonia	**0.70**	**0.33**	−0.15	−0.04	−0.14	−0.03
	Low mood	**0.75**	0.26	−0.23	−0.04	−0.13	0.00
	Sleep	**0.51**	0.15	0.23	−0.22	0.04	0.01
	Tired	**0.60**	0.26	0.19	−0.23	0.03	0.00
	Appetite	**0.56**	0.25	0.11	−0.07	0.08	0.00
	Inadequacy	**0.69**	0.18	−0.17	0.06	0.09	0.05
	Concentrating	**0.65**	0.22	0.11	0.11	0.05	−0.03
	Movement	**0.48**	0.15	0.12	0.24	0.07	−0.05
	Suicidality	**0.50**	0.20	−0.17	0.18	0.08	0.03
	Nervousness	**0.74**	**−0.30**	−0.09	−0.06	0.01	−0.02
	Control worrying	**0.79**	**−0.36**	−0.11	−0.09	0.05	−0.03
	Worrying about different things	**0.78**	**−0.34**	−0.07	−0.11	0.03	0.00
	Relaxing	**0.75**	−0.23	0.18	0.04	−0.11	0.02
	Restlessness	**0.56**	−0.15	0.25	0.25	−0.08	0.00
	Irritability	**0.64**	−0.09	0.07	0.04	−0.07	0.05
	Foreboding	**0.67**	−0.22	−0.08	0.06	0.08	0.01

Note: $N = 152{,}826$. Only factors with positive eigenvalues have been retained. The extraction method was principle factors and unrotated results are displayed. Item labels correspond to the following items: Anhedonia: "Little interest or pleasure in doing things", Low mood: "Feeling down, depressed, or hopeless", Sleep: "Trouble falling or staying asleep, or sleeping too much", Tired: "Feeling tired or having little energy", Appetite: "Poor appetite or overeating", Inadequacy: "Feeling bad about yourself or that you are a failure or have let yourself or your family down", Concentrating: "Trouble concentrating on things, such as reading the newspaper or watching television", Movement: "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", Suicidality: "Thoughts that you would be better off dead or of hurting yourself in some way", Nervousness: "Feeling nervous, anxious or on edge", Control worrying: "Not being able to stop or control worrying", Worrying about different things: "Worrying too much about different things", Relaxing: "Trouble relaxing", Restlessness: "Being so restless that it is hard to sit still", Irritability: "Becoming easily annoyed or irritable", Foreboding: "Feeling afraid as if something awful might happen".

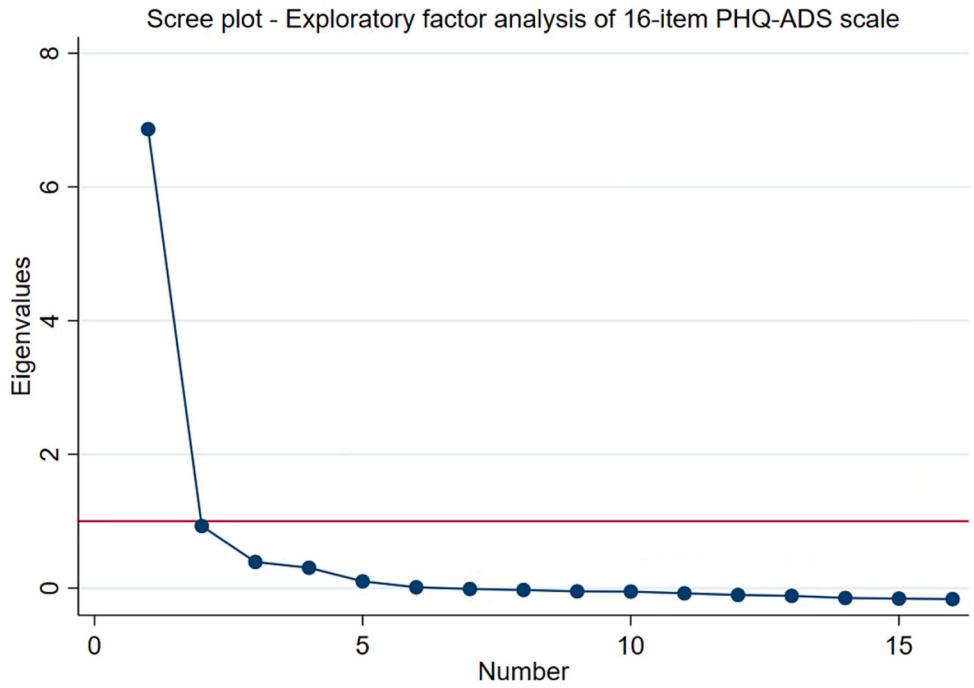

Fig. 1. Scree plot from the exploratory factor analysis including all items from the 16-item PHQ-ADS scale.

vulnerability i.e. optimize the psychometric properties of the scale, items with low monotonicity/scalability ($H < 0.50$) and low discrimination ($\alpha < 2$) were omitted and the analysis repeated. For all items Loevinger's H increased in value (Table 3). For the items emphasising mood (anhedonia, low mood, inadequacy) discrimination dropped marginally. For example, for anhedonia discrimination dropped from $\alpha = 2.6$ to $\alpha = 2.16$. For items emphasising anxiety, however (nervousness, control worrying, multiple worries, trouble relaxing), discrimination values increased markedly. For example, for control worrying, discrimination increased from $\alpha = 4.22$ to $\alpha = 6.14$. The item difficulty parameter (β) was largely unchanged with marginal increases to the mood items and marginal reductions in the anxiety items.

The comparison between scale lengths can be visualised using item information function (IIF) curves (Fig. 2). The IIF curves display the relationship between discrimination α and difficulty β. The IIF curves indicate how much each item contributes to the overall test information. The sub-optimal performance of questions assessing sleep and appetite can be easily appreciated by the flat nature of their curve. The bimodal peaks of the curves indicated where on the latent trait continuum each item is most informative. All items had their maximum curvature

Table 3
IRT model item parameters for the 16-item and 7-item scales.

Item	16-item scale					7-item scale				
	Monotonicity H	Discrimination α	Difficulty β			Monotonicity H	Discrimination α	Difficulty β		
			k ≥ 1	k ≥ 2	k = 3			k ≥ 1	k ≥ 2	k = 3
Anhedonia	0.51	2.60	1.07	2.11	2.58	0.56	2.16	1.14	2.27	2.78
Low mood	0.55	3.01	0.90	2.08	2.53	0.61	2.59	0.94	2.18	2.66
Sleep	0.44	1.40	0.07	1.65	2.26					
Tired	0.52	1.75	0.02	1.72	2.27					
Appetite	0.42	1.69	1.28	2.32	2.87					
Inadequacy	0.50	2.49	1.05	2.08	2.49	0.55	2.20	1.10	2.18	2.62
Concentrating	0.49	2.19	1.16	2.28	2.77					
Movement	0.44	2.05	2.08	2.94	3.42					
Suicidality	0.48	2.42	2.09	2.98	3.39					
Nervousness	0.54	3.36	0.67	1.87	2.23	0.63	3.96	0.66	1.81	2.16
Control worrying	0.56	4.22	0.79	1.76	2.08	0.66	6.14	0.76	1.69	2.00
Multiple worries	0.58	3.70	0.55	1.74	2.09	0.67	4.99	0.54	1.66	2.00
Relaxing	0.56	3.23	0.67	1.75	2.12	0.61	3.08	0.68	1.78	2.15
Restlessness	0.45	2.15	1.49	2.53	2.99					
Irritability	0.49	2.22	0.77	2.29	2.80					
Foreboding	0.49	2.75	1.12	2.11	2.48					

Note: Item labels correspond to the following items: Anhedonia: "Little interest or pleasure in doing things", Low mood: "Feeling down, depressed, or hopeless", Sleep: "Trouble falling or staying asleep, or sleeping too much", Tired: "Feeling tired or having little energy", Appetite: "Poor appetite or overeating", Inadequacy: "Feeling bad about yourself or that you are a failure or have let yourself or your family down", Concentrating: "Trouble concentrating on things, such as reading the newspaper or watching television", Movement: "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", Suicidality: "Thoughts that you would be better off dead or of hurting yourself in some way", Nervousness: "Feeling nervous, anxious or on edge", Control worrying: "Not being able to stop or control worrying", Worrying about different things: "Worrying too much about different things", Relaxing: "Trouble relaxing", Restlessness: "Being so restless that it is hard to sit still", Irritability: "Becoming easily annoyed or irritable", Foreboding: "Feeling afraid as if something awful might happen". k: cut-points of the rating scale: "Not at all (0)", "Several days (1)", "More than half the days (2)", "Nearly every day (3)". $p < 0.0001$ for all discrimination parameters.

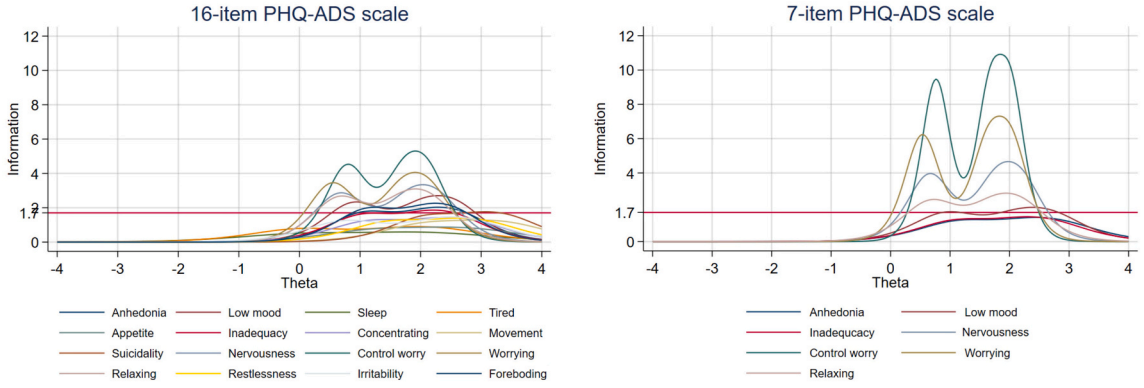

Fig. 2. Item information functions for the 16-item and 7-item PHQ-ADS scales.

positioned in the positive half (θ > 0), which indicates they provide information about the presence of anxious depressive symptomatology in the upper range. By contrast, items are less informative for respondents without anxious depressive symptomatology because no item peaked at a negative latent trait (θ). This suggests the scale is less informative for the absence of symptoms than for the presence of symptoms. The IIF curves for the revised 7-item scale show improved informativeness (height of curve) when compared to the 16-item scale.

3.4. Validation of the reduced 7-item scale using mental health outcomes

The logistic regression analyses showed that discrimination, measured using the area under the curve of the receiver operating characteristic, was satisfactory ranging from 0.63 to 0.70 for all scales including the PHQ-ADS, the reduced 7-item scale and the PHQ-4 scale (see Fig. 3). The percentage of correctly classified participants as having the mental health outcome or not ranged from 68.9 % for predicting "Seen GP for nerves" using the PHQ-4 to 78.9 % for predicting "Ever been offered/sought treatment for anxiety" using the 16-item PHQ-ADS score (see Table 4). Differences between scales in correct classifications were negligible and <0.8 % (see Table 4), thereby indicating that all scales performed equally satisfactorily and that no important information was lost through scale reduction.

3.5. Validation of the IRT analysis in a subsample of participants with at least one depressive episode

The validation of the IRT analysis in the subsample of participants with depression (N = 11,653) revealed that the items measuring sleep disturbance, feeling tired, appetite, movement, suicidality, restlessness also had poor psychometric properties as seen in either low discrimination or lack of scalability (Table 5). These items were discarded and a reduced scale with 10-items was estimated (Table 5). The 10-item scale showed good discrimination ranging from 2.15 (anhedonia) to 5.66 (inability to control worrying), as well as good scalability ranging from H = 0.52 (trouble concentrating) to H = 0.67 (multiple worries). Compared with the 7-item scale derived in the general population sample, the 10-item scale derived in the clinical sample including only

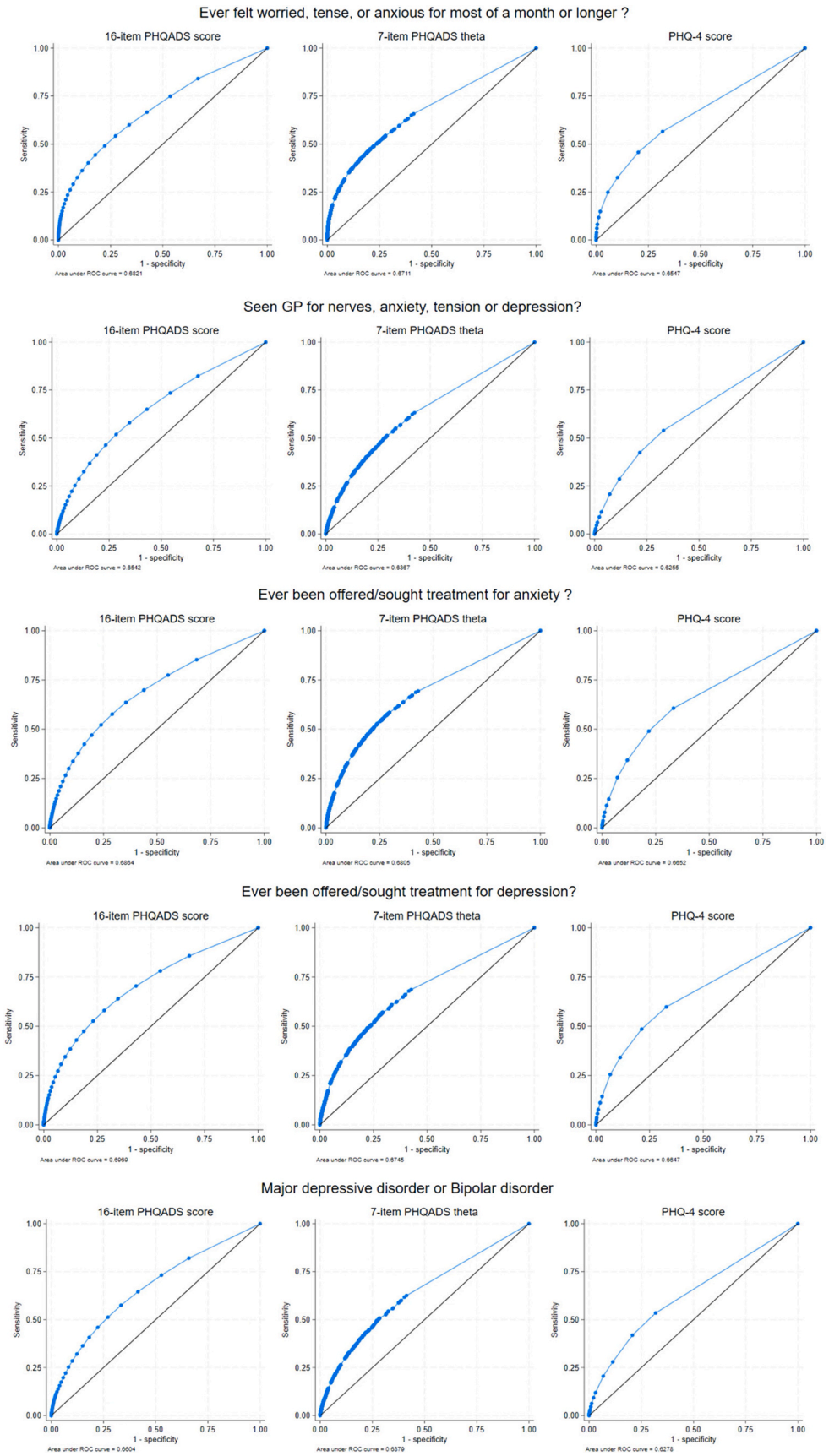

Fig. 3. Receiver operating characteristic curves for predicting mental health outcomes using three psychometric scales: the 16-item PHQ-ADS scale, the 7-item PHQ-ADS scale and the PHQ-4 scale.

Table 4
Quality indicators of classification from the logistic regression models predicting mental health outcomes using the 16-item, 7-item PHQ-ADS scores, and the PHQ-4 scores for comparison.

Outcome(O)	Predictor	Sensitivity	Specificity	PPV	NPV	FPR	FNR	P(+| ~ O)	P(−| ~ O)	Correctly classified	ROC AUC
Ever felt worried, tense, or anxious for most of a month or longer?	16-item PHQADS score	21.0 %	96.5 %	67.9 %	77.4 %	3.6 %	79.0 %	32.1 %	22.7 %	76.6 %	0.68
	7-item PHQADS theta	16.3 %	98.0 %	74.7 %	76.6 %	2.0 %	83.7 %	25.3 %	23.4 %	76.5 %	0.67
	PHQ-4	14.9 %	98.0 %	73.0 %	76.3 %	2.0 %	85.2 %	27.0 %	23.7 %	76.1 %	0.65
Seen GP for nerves, anxiety, tension or depression?	16-item PHQADS score	17.3 %	95.0 %	62.7 %	70.3 %	5.0 %	82.7 %	37.3 %	29.7 %	69.6 %	0.65
	7-item PHQADS theta	17.2 %	94.7 %	61.1 %	70.2 %	5.3 %	82.8 %	38.9 %	29.8 %	69.3 %	0.64
	PHQ-4	11.5 %	96.8 %	63.4 %	69.2 %	3.2 %	88.5 %	36.6 %	30.8 %	68.9 %	0.63
Ever been offered/sought treatment for anxiety?	16-item PHQADS score	13.2 %	97.5 %	60.1 %	79.8 %	2.5 %	86.8 %	39.9 %	20.2 %	78.9 %	0.69
	7-item PHQADS theta	10.1 %	98.3 %	62.4 %	79.4 %	1.7 %	89.9 %	37.7 %	20.6 %	78.8 %	0.68
	PHQ-4	11.3 %	97.9 %	60.1 %	79.5 %	2.1 %	88.7 %	39.9 %	20.5 %	78.7 %	0.67
Ever been offered/sought treatment for depression?	16-item PHQADS score	17.1 %	96.9 %	63.7 %	78.7 %	3.1 %	82.9 %	36.3 %	21.3 %	77.8 %	0.70
	7-item PHQADS theta	11.0 %	97.9 %	61.8 %	77.7 %	2.2 %	89.0 %	38.2 %	22.3 %	77.0 %	0.67
	PHQ-4	14.4 %	97.1 %	61.4 %	78.3 %	2.9 %	85.6 %	38.6 %	21.8 %	77.3 %	0.66
Major depressive disorder or Bipolar disorder	16-item PHQADS score	13.9 %	96.7 %	62.2 %	74.1 %	3.3 %	86.1 %	37.8 %	25.9 %	73.3 %	0.66
	7-item PHQADS theta	10.0 %	97.5 %	61.1 %	73.4 %	2.5 %	90.1 %	38.9 %	26.6 %	72.8 %	0.64
	PHQ-4	12.0 %	96.9 %	60.1 %	73.7 %	3.1 %	88.0 %	39.9 %	26.3 %	72.9 %	0.63

Note: PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False-positive rate, FNR: False-negative rate, P(+| ~ O): Conditional probability of being classified as positive as a true negative, P(+| ~ O): Conditional probability of being classified as negative as a true positive, ROC AUC: Area under the curve of the receiver operating characteristic, an indicator of overall discrimination.

participants who had at least one depressive episode also included the items measuring concentrating, irritability and feelings of foreboding that showed good discrimination and scalability in the clinical sample, but not in the general population. Finally, an IRT analysis was conducted using items from the 7-item scale in the clinical sample (Table 5). This analysis confirmed that all 7 items had good discrimination and scalability in the clinical sample of participants with depression.

4. Discussion

Here we investigated psychometric properties of the PHQ-ADS scale in a subsample of 152,826 participants from the UK Biobank using IRT. The 16-item PHQ-ADS was found to have a limited range, unscalable items (sleep disturbance, appetite, trouble concentrating, movement, suicidality, restlessness, irritability, feelings of foreboding; Table 3), and items with poor discrimination (sleep disturbance, appetite, feeling tired). When applying scale reductions, it was found that a reduced 7-item scale provided optimum discrimination and monotonicity. Although the 7-item version of the scale possessed good reliability, this was focussed at the upper end of the latent trait, rather than the extremes of the distribution. The scale is suitable for assessing population levels but not for accurately assessing very low or extremely high levels, such as respondents with either no symptoms or severe clinical symptoms.

These findings suggest that a reduced-item 7-item PHQ-ADS scale would reduce response burden without loss of information. Such a reduced scale is useful in a variety of settings, e.g. to reduce the burden on patients in lengthy clinical assessments, to reduce attrition rates or missing data in cohort studies with a lengthy protocol, and to improve data quality by reducing response bias due to patient fatigue. Moreover, a psychometric scale with optimized information is useful in large-scale trials. Previous research demonstrated that IRT can be used for assessing the efficiency of psychometric instruments. Indeed, IRT has been used to propose item reductions in screening tools, including the 19-item feelings scale for depression (Edelen and Reeve, 2007), 16-item Anxiety Sensitivity Index (Zvolensky et al., 2009), the 9-item Simple Clinical Colitis Activity Index (Walsh et al., 2021), and the 12-item Eysenck Neuroticism scale (Bauermeister et al., 2022). The present study adds to the growing body of research that provides improved tools for research, particularly for projects conducted at scale.

Whilst this is the first IRT study of the PHQ-ADS, IRT has been applied to the PHQ-9 and GAD-7 separately. A study of the PHQ-9 using IRT in patients with major depressive disorder and found that the discrimination values ranged from 1.45 to 2.80 with “low mood” showing the highest discrimination and sleep disturbance lowest discrimination (Ma et al., 2021). This is similar to our findings on the entire 16-item PHQ-ADS scale, where, when considering only items from the PHQ-9, “low mood” was found to have the highest discrimination, and sleep was found to have the lowest discrimination. Ma et al. (2021) also showed that reliability fell below <0.9 at the extreme end of the latent trait continuum, which is similar to our findings on the entire PHQ-ADS scale.

A study of the GAD-7 using IRT in a sample of primary care patients found that the discrimination values ranged from 1.69 to 3.55 with the two items assessing worry showing the highest discrimination and items assessing the symptoms of irritability and restlessness showing lowest discrimination (Jordan et al., 2017). This is similar to our findings on the entire 16-item PHQ-ADS scale, where, when considering only items from the GAD-7 the two items assessing worry were found to have the highest discrimination, whereas restlessness and irritability were found to have the lowest discrimination.

Jordan et al. (2017) conducted a Mokken analysis of the GAD-7 finding monotonicity throughout the encitytire scale. This is at variance with our findings from the entire 16-item PHQ-ADS scale, where 3 items did not fulfill the monotonicity criterion. A possible explanation for this discrepancy could be that the present study was conducted in the general population whereas the sample investigated by Jordan et al. (2017) was drawn from a somatoform disorders network. Since the items with low monotonicity were somatic symptoms of irritability and restlessness that are not disease-specific to anxiety disorders, this could

Table 5
IRT model item parameters for the 16-item, 10-item, and 7-item scales in a subsample of participants with depression from the UK Biobank.

Item	16-item scale					10-item scale					7-item scale				
	Monotonicity H	Discrimination α	Difficulty β			Monotonicity H	Discrimination α	Difficulty β			Monotonicity H	Discrimination α	Difficulty β		
			k ≥ 1	k ≥ 2	k = 3			k ≥ 1	k ≥ 2	k = 3			k ≥ 1	k ≥ 2	k = 3
Anhedonia	0.55	2.55	0.72	1.76	2.24	0.56	2.15	0.78	1.88	2.41	0.59	2.01	0.81	1.95	2.50
Low mood	0.58	2.83	0.53	1.73	2.20	0.59	2.45	0.57	1.82	2.32	0.63	2.31	0.59	1.88	2.39
Sleep	**0.47**	**1.40**	−0.32	1.25	1.89										
Tired	0.54	**1.72**	−0.39	1.31	1.85										
Appetite	**0.45**	1.56	0.84	1.91	2.46										
Inadequacy	0.54	2.40	0.65	1.66	2.09	0.56	2.17	0.68	1.74	2.19	0.58	2.03	0.71	1.79	2.26
Concentrating	0.53	2.19	0.74	1.86	2.38	0.52	1.85	0.81	2.02	2.58					
Movement	**0.47**	1.92	1.70	2.61	3.19										
Suicidality	**0.49**	2.16	1.74	2.62	3.14										
Nervousness	0.57	3.30	0.32	1.51	1.86	0.63	3.86	0.33	1.47	1.80	0.65	3.89	0.34	1.48	1.81
Control worrying	0.59	4.15	0.42	1.40	1.72	0.66	5.66	0.43	1.35	1.66	0.69	6.79	0.43	1.34	1.65
Multiple worries	0.60	3.77	0.20	1.37	1.70	0.67	4.92	0.21	1.32	1.64	0.69	5.46	0.22	1.32	1.64
Relaxing	0.58	3.13	0.29	1.36	1.75	0.62	3.18	0.30	1.37	1.75	0.63	3.08	0.31	1.39	1.78
Restlessness	**0.48**	2.04	1.12	2.15	2.62										
Irritability	0.52	2.25	0.44	1.87	2.39	0.54	2.21	0.46	1.89	2.41					
Foreboding	0.51	2.53	0.74	1.77	2.16	0.57	2.72	0.74	1.74	2.11					

Note: N = 11,653. Only participants who had a depressive episode (Single MDD, Recurrent MDD, bipolar 1, bipolar 2) were included in the analysis.

have led to monotonicity violations in the general population, whereas the scale was monotone when only including participants with somatoform disorders. A further study investigated psychometric properties of the GAD-7 in a sample of antepartum women from the Côte d'Ivoire using Rasch scale analysis, an IRT model for ordered data to evaluate monotonicity, and found that items 5, 6, and 7 had disordered thresholds, thereby violating the monotonicity criterion (Barthel et al., 2014). These are the same items from the GAD-7 that showed violations against monotonicity in our Mokken scale analysis and were therefore discarded from our revised 7-item PHQADS scale. Barthel et al. (2014) also showed that the scale violated assumptions of unidimensionality and found that all items of GAD-7 showed a lack of monotonicity in a sample from Ghana. The cross-cultural difference might come into consideration here.

The items that were removed from the scale to yield the 7-item scale had in common that most of them measured somatic symptoms of anxiety/depression, whereas the items that were retained in the 7-item scale measured affective and cognitive symptoms. A possible explanation for this might be that cognitive-affective symptoms of anxiety/depression and somatic symptoms of anxiety/depression are two distinct latent constructs and might be best measured using different scales. Since the study was conducted in the general population, the study included participants with somatic symptoms that were not necessarily suffering from anxiety/depression. Screening tools for anxiety/depression should, therefore, not include items assessing somatic symptoms of depression/anxiety as they are not disease-specific as seen with low monotonicity and therefore not informative in the general population.

Similar to our findings, previous research distinguished cognitive-affective and somatic items in the PHQ-9 and GAD-7. Previous research showed that the PHQ-9 was characterized by a two-factor structure in confirmatory factor analytic studies that distinguished between a cognitive-affective and a somatic factor (e.g. Keum et al., 2018; Patel et al., 2019). However, between-factor correlations were high enough ($\geq$ 0.80) to suggest a unidimensional structure (Keum et al., 2018; Patel et al., 2019). Similarly, previous research using confirmatory factor analysis showed that the GAD-7 had a two-dimensional factor structure in some samples that represented cognitive-affective and somatic symptoms (e.g. Moreno et al., 2019). However, the factor structure of the GAD-7 was found to be unidimensional in most populations (Dear et al., 2011; García-Campayo et al., 2010; Hinz et al., 2017; Löwe et al., 2008; Mills et al., 2014; Rodebaugh et al., 2008; Ryan et al., 2013; Spitzer et al., 2006). The present study adds to these findings from previous research that the combined GAD-7 and PHQ-9 scales, the PHQ-ADS scale, was also unidimensional as shown in the exploratory factor analysis.

4.1. Strengths

The contribution of this study is to optimize a psychometric screening scale for emotional vulnerability to be useful for transdiagnostic dimensional research. Screening for anxiety and depression in epidemiological research is typically conducted using psychometric scales (such as the GAD-7 scale and PHQ-9) which have been largely derived from the DSM-IV diagnostic criteria (Kroenke et al., 2001; Spitzer et al., 2006). However, for much research and clinical practice, this diagnostic approach is simplistic as there is increasing evidence that the biopsychosocial processes underlying mental health are transdiagnostic; suggesting the symptom space is not categorical but dimensional (Dalgleish et al., 2020). As a result, current screening tools for depression and anxiety based on DSM-IV criteria are sub-optimal as they recognise heterogeneous groups rather than the commonality of symptoms (Dalgleish et al., 2020). The transdiagnostic dimensional approach cuts across existing categorical symptoms to improve the way we classify mental disorders (e.g. Fusar-Poli et al., 2019). Although DSM-5 attempted to add superordinate dimensions that cross diagnostic boundaries in its conceptional development (Regier et al., 2009), the

dimensional approach was rejected as premature (American Psychiatric Association, 2013; Paris and Phillips, 2013). Nevertheless, there is a scientific need to further develop the transdiagnostic dimensional approach by improving psychometric scales for use across transdiagnostic categories to measure underlining common constructs such as emotional vulnerability. This study optimized the psychometric properties of the PHQ-ADS as a transdiagnostic scale for a unidimensional construct of emotional vulnerability.

Another strength of this study was the rigorous quantitative analysis of the PHQ-ADS using IRT in a large and diverse cohort including participants at various points of the latent trait continuum θ. IRT has several advantages over other methods: (a) improved precision through identification and deletion of misfitting items, (b) investigation of boundary effects in difficulty, (c) less time is needed for accurate measurement (Henning, 1984). Another strength was that the PHQ-ADS scale was adapted for use in epidemiological research by shortening the scale length without losing important information. Short, yet informative scales are important for epidemiological research for various reasons: First, it is desirable to avoid overly lengthy assessments to reduce the burden on research participants. Second, shorter scales can help to lower attrition rates or missing data by preventing from participants dropping out or not responding. Third, scale reductions can improve data quality by reducing respondent fatigue and associated response errors such as misreading items. Fourth, shorter scales can help assess a broader range of risk factors in epidemiological research within the limits of time constrained study protocols.

4.2. Limitations

Limitations in the study include that psychometric properties were only investigated in population with a limited age range. It is important to recognise that extrapolation of these findings to younger age groups requires caution. A second issue is representativeness. UK Biobank does not claim to be representative, but it is heterogeneous in the sense that the full range of variable scores is represented. Consequently, whilst the distributions of variables will not reflect that of the UK population, associations between risk factors will reflect those found more widely.

4.3. Future research

Further research is needed to validate the 7-item scale in the clinical population, and to investigate its construct validity and test-retest reliability. Future research should also focus on the development of a scale that assesses the entire continuum of emotional vulnerability, by also including items measuring the opposites of low mood such as happiness, and by also including items that measure symptoms only found in severely depressed/anxious patients. Such a scale that is reliable along a broad θ range is desirable as an improved outcome metric in longitudinal research.

5. Conclusions

Our study demonstrated that the PHQ-ADS could be reduced from the current 16-item scale to a 7-item scale without losing discrimination. This reduced 7-item scale can be used to reduce response burden while retaining optimal test information in epidemiological research, and is therefore a valuable alternative to other reduced scales such as the PHQ-4. Similar to the PHQ-4, the reduced PHQ-ADS scale comprised mainly items measuring cognitive-affective symptoms of anxiety/depression, whereas items measuring somatic symptoms were excluded due to a lack of disease-specificity as shown by low monotonicity. Overall, our findings demonstrate that IRT is a useful technique for scale reductions and that there is a need to optimize clinical screening questionnaires.

Funding

The Medical Research Council supports DPUK through grant MR/T0333771. SB and CPP are supported by DPUK.

CRediT authorship contribution statement

CPP, SB and JG conceptualised the idea. CPP and SB analysed and interpreted the data, and wrote the manuscript. CPP and JG edited and proofread the manuscript. All authors contributed to and have approved the final manuscript.

Declaration of competing interest

SB, CPP and JG declare no competing interests.

Data availability

The dataset(s) supporting the conclusions of this article is(are) available in the Dementias Platform UK (DPUK) Data Portal repository, https://portal.dementiasplatform.uk/.

Acknowledgements

All analyses were conducted on the Dementias Platform (DPUK) Data Portal using UK Biobank application 15697 PI John Gallacher for DPUK project 0169. The Medical Research Council supports DPUK through grant MR/T0333771. Sarah Bauermeister and Chris Patrick Pflanz are supported by DPUK. Access to the data can be requested through UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2023.11.067.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Publishing.

Baker, F.B., 2001. The Basics of Item Response Theory. ERIC.

Bandelow, B., 2020. In: Kim, Y.-K. (Ed.), Current and Novel Psychopharmacological Drugs for Anxiety Disorders BT-Anxiety Disorders: Rethinking and Understanding Recent Discoveries. Springer Singapore, Singapore, pp. 347–365. https://doi.org/10.1007/978-981-32-9705-0_19.

Barthel, D., Barkmann, C., Ehrhardt, S., Bindt, C., 2014. Psychometric properties of the 7-item generalized anxiety disorder scale in antepartum women from Ghana and Côte d'Ivoire. J. Affect. Disord. 169, 203–211. https://doi.org/10.1016/j.jad.2014.08.004.

Bauermeister, S., Orton, C., Thompson, S., Barker, R.A., Bauermeister, J.R., Ben-Shlomo, Y., Brayne, C., Burn, D.J., Campbell, A.I., Calvin, C., Chandran, S., Chaturvedi, N., Chêne, G., Chessell, I.P., Corbett, A., Davis, D.H.J., Denis, M., Dufouil, C., Elliott, P., Fox, N.C., Hill, D., Hofer, S., Hu, M.T.M., Jindra, C., Kee, F., Kim, C.H., Kim, C.Y., Kivimaki, M., Koychev, I., Kwan, J., Lawson, R.A., Leroi, I., Linden, G.J., Love, S., Lovestone, S., Lyons, R.A., Mackay, C.E., Matthews, P.M., Mcguiness, B., Middleton, L.T., Moody, C., Moore, K.M., Na, D.L., O'Brien, J.T., Ourselin, S., Paranjothy, S., Park, K.S., Porteous, D.J., Richards, M., Ritchie, C.W., Rohrer, J.D., Rossor, M.N., Rowe, J.B., Scahill, R.I., Schnier, C., Schott, J.M., Seo, S.W., South, M., Steptoe, A., Tabrizi, S.J., Tales, A., Thomas, A.J., Tillin, T., Timpson, N.J., Toga, A.W., Visser, P.J., Wade-Martins, R., Wilkinson, T., Williams, J., Wong, A., Gallacher, J.E., 2019. The dementias platform UK (DPUK) data portal. Eur. J. Epidemiol. 35, 601–611.

Bauermeister, S., Pflanz, C.P., Gallacher, J., 2022. Adapting the Eysenck Personality Questionnaire-Revised Neuroticism scale for use in epidemiologic studies: A psychometric evaluation using item response theory in the UK Biobank. bioRxiv 741249. https://doi.org/10.1101/741249.

Cosco, T.D., Doyle, F., Ward, M., McGee, H., 2012. Latent structure of the hospital anxiety and depression scale: a 10-year systematic review. J. Psychosom. Res. 72, 180–184. https://doi.org/10.1016/j.jpsychores.2011.06.008.

Crichton, N., 1999. Information point Mokken scale analysis. J. Clin. Nurs. 8, 388.

Dalgleish, T., Black, M., Johnston, D., Bevan, A., 2020. Transdiagnostic approaches to mental health problems: current status and future directions. J. Consult. Clin. Psychol. https://doi.org/10.1037/ccp0000482.

Dear, B.F., Titov, N., Sunderland, M., McMillan, D., Anderson, T., Lorian, C., Robinson, E., 2011. Psychometric comparison of the generalized anxiety disorder scale-7 and the Penn State worry questionnaire for measuring response during treatment of generalised anxiety disorder. Cogn. Behav. Ther. 40, 216–227. https://doi.org/10.1080/16506073.2011.582138.

Edelen, M.O., Reeve, B.B., 2007. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual. Life Res. 16, 5. https://doi.org/10.1007/s11136-007-9198-0.

Fusar-Poli, P., Solmi, M., Brondino, N., Davies, C., Chae, C., Politi, P., Borgwardt, S., Lawrie, S.M., Parnas, J., McGuire, P., 2019. Transdiagnostic psychiatry: a systematic review. World Psychiatry 18, 192–207. https://doi.org/10.1002/wps.20631.

García-Campayo, J., Zamorano, E., Ruiz, M.A., Pardo, A., Pérez-Páramo, M., López-Gómez, V., Freire, O., Rejas, J., 2010. Cultural adaptation into Spanish of the generalized anxiety disorder-7 (GAD-7) scale as a screening tool. Health Qual. Life Outcomes 8, 1–11.

Groen, R.N., Ryan, O., Wigman, J.T.W., Riese, H., Penninx, B.W.J.H., Giltay, E.J., Wichers, M., Hartman, C.A., 2020. Comorbidity between depression and anxiety: assessing the role of bridge mental states in dynamic psychological networks. BMC Med. 18, 308. https://doi.org/10.1186/s12916-020-01738-z.

Hanel, G., Henningsen, P., Herzog, W., Sauer, N., Schaefert, R., Szecsenyi, J., Löwe, B., 2009. Depression, anxiety, and somatoform disorders: vague or distinct categories in primary care? Results from a large cross-sectional study. J. Psychosom. Res. 67, 189–197.

Henning, G., 1984. Advantages of latent trait measurement in language testing. Lang. Test. 1, 123–133. https://doi.org/10.1177/026553228400100201.

Herzog, A.R., Bachman, J.G., 1981. Effects of questionnaire length on response quality. Public Opin. Q. 45, 549–559. https://doi.org/10.1086/268687.

Hinz, A., Klein, A.M., Brähler, E., Glaesmer, H., Luck, T., Riedel-Heller, S.G., Wirkner, K., Hilbert, A., 2017. Psychometric evaluation of the generalized anxiety disorder screener GAD-7, based on a large German general population sample. J. Affect. Disord. 210, 338–344.

Jordan, P., Shedden-Mora, M.C., Löwe, B., 2017. Psychometric analysis of the generalized anxiety disorder scale (GAD-7) in primary care using modern item response theory. PLoS One 12, e0182162.

Keum, B.T., Miller, M.J., Inkelas, K.K., 2018. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol. Assess. https://doi.org/10.1037/pas0000550.

Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. 32 (9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613. https://doi.org/10.1046/j.1525-1497.2001.0160 09606.x.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., Monahan, P.O., Löwe, B., 2007. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann. Intern. Med. 146, 317–325. https://doi.org/10.7326/0003-4819-146-5-200703060-00004.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., Löwe, B., 2009. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 50, 613–621. https://doi.org/10.1176/appi. psy.50.6.613.

Kroenke, K., Wu, J., Yu, Z., Bair, M.J., Kean, J., Stump, T., Monahan, P.O., 2016. The patient health questionnaire anxiety and depression scale (PHQ-ADS): initial validation in three clinical trials. Psychosom. Med. 78, 716.

Lavrakas, P.J., 2008. Encyclopedia of Survey Research Methods. SAGE Publications.

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., Herzberg, P.Y., 2008. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. Med. Care 266–274.

Ma, S., Yang, J., Yang, B., Kang, L., Wang, P., Zhang, N., Wang, W., Zong, X., Wang, Y., Bai, H., Guo, Q., Yao, L., Fang, L., Liu, Z., 2021. The patient health Questionnaire-9 vs. the Hamilton rating scale for depression in assessing major depressive disorder. Front. Psychol. 12, 747139 https://doi.org/10.3389/fpsyt.2021.747139.

ter Meulen, W.G., Draisma, S., van Hemert, A.M., Schoevers, R.A., Kupka, R.W., Beekman, A.T.F., Penninx, B.W.J.H., 2021. Depressive and anxiety disorders in concert–a synthesis of findings on comorbidity in the NESDA study. J. Affect. Disord. 284, 85–97. https://doi.org/10.1016/j.jad.2021.02.004.

Mills, S.D., Fox, R.S., Malcarne, V.L., Roesch, S.C., Champagne, B.R., Sadler, G.R., 2014. The psychometric properties of the generalized anxiety disorder-7 scale in Hispanic Americans with English or Spanish language preference. Cult. Divers. Ethn. Minor. Psychol. 20, 463.

Moreno, E., Muñoz-Navarro, R., Medrano, L.A., González-Blanch, C., Ruiz-Rodríguez, P., Limonero, J.T., Moretti, L.S., Cano-Vindel, A., Moriana, J.A., 2019. Factorial invariance of a computerized version of the GAD-7 across various demographic groups and over time in primary care patients. J. Affect. Disord. 252, 114–121. https://doi.org/10.1016/j.jad.2019.04.032.

Nash, J., Nutt, D., 2007. Psychopharmacology of anxiety. Psychiatry 6, 143–148. https://doi.org/10.1016/j.mppsy.2007.02.001.

Newman, M.W., 2022. Value added? A pragmatic analysis of the routine use of PHQ-9 and GAD-7 scales in primary care. Gen. Hosp. Psychiatry 79, 15–18. https://doi.org/10.1016/j.genhosp-psych.2022.09.005.

Ohi, K., Otowa, T., Shimada, M., Sasaki, T., Tanii, H., 2020. Shared genetic etiology between anxiety disorders and psychiatric and related intermediate phenotypes. Psychol. Med. 50, 692–704. https://doi.org/10.1017/S003329171900059X.

Paris, J., Phillips, J. (Eds.), 2013. Making the DSM-5: Concepts and Controversies., Making the DSM-5: Concepts and Controversies. Springer Science + Business Media, New York, NY, US. https://doi.org/10.1007/978-1-4614-6504-1.

Patel, J.S., Oh, Y., Rand, K.L., Wu, W., Cyders, M.A., Kroenke, K., Stewart, J.C., 2019. Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. Depress. Anxiety 36, 813–823. https://doi.org/10.1002/da.22940.

Purves, K.L., Coleman, J.R.I., Meier, S.M., Rayner, C., Davis, K.A.S., Cheesman, R., Bækvad-Hansen, M., Børglum, A.D., Wan Cho, S., Jürgen Deckert, J., Gaspar, H.A., Bybjerg-Grauholm, J., Hettema, J.M., Hotopf, M., Hougaard, D., Hübel, C., Kan, C., McIntosh, A.M., Mors, O., Bo Mortensen, P., Nordentoft, M., Werge, T., Nicodemus, K.K., Mattheisen, M., Breen, G., Eley, T.C., 2020. A major role for common genetic variation in anxiety disorders. Mol. Psychiatry 25, 3292–3303. https://doi.org/10.1038/s41380-019-0559-1.

Regier, D.A., Narrow, W.E., Kuhl, E.A., Kupfer, D.J., 2009. The conceptual development of DSM-V. Am. J. Psychiatry 166, 645–650. https://doi.org/10.1176/appi.ajp.2009.09020279.

Rodebaugh, T.L., Holaway, R.M., Heimberg, R.G., 2008. The factor structure and dimensional scoring of the generalized anxiety disorder questionnaire for DSM-IV. Assessment 15, 343–350.

Ryan, T.A., Bailey, A., Fearon, P., King, J., 2013. Factorial invariance of the patient health questionnaire and generalized anxiety disorder questionnaire. Br. J. Clin. Psychol. 52, 438–449.

Smith, A.L.W., Harmer, C.J., Cowen, P.J., Murphy, S.E., 2023. The serotonin 1A (5-HT1A) receptor as a pharmacological target in depression. CNS Drugs 37, 571–585. https://doi.org/10.1007/s40263-023-01014-7.

Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166, 1092–1097. https://doi.org/10.1001/archinte.166.10.1092.

StataCorp, 2021. Stata Statistical Software College Station.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (e1001), 779.

van Tol, M.J., van der Wee, N.J.A., Veltman, D.J., 2021. Fifteen years of NESDA neuroimaging: an overview of results related to clinical profile and bio-social risk factors of major depressive disorder and common anxiety disorders. J. Affect. Disord. 289, 31–45. https://doi.org/10.1016/j.jad.2021.04.009.

Walsh, A., Cao, R., Wong, D., Kantschuster, R., Matini, L., Wilson, J., Kormilitzin, A., South, M., Travis, S., Bauermeister, S., 2021. Using item response theory (IRT) to improve the efficiency of the simple clinical colitis activity index (SCCAI) for patients with ulcerative colitis. BMC Gastroenterol. 21, 132. https://doi.org/10.1186/s12876-021-01621-y.

Zvolensky, M.J., Strong, D., Bernstein, A., Vujanovic, A.A., Marshall, E.C., 2009. Evaluation of anxiety sensitivity among daily adult smokers using item response theory analysis. J. Anxiety Disord. 23, 230–239. https://doi.org/10.1016/j.janxdis.2008.07.005.