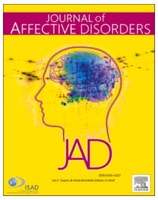
Contents lists available at ScienceDirect

Journal of Affective Disorders

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



Research paper

Measurement invariance and differential item functioning of the PHQ-9 and GAD-7 between working age and older adults seeking treatment for common mental disorders

H. Delamain[a,*], J.E.J. Buckman[a,b], J. Stott[c], A. John[c], S. Singh[d], S. Pilling[a,e], R. Saunders[a]

[a] *CORE Data Lab, Centre for Outcomes Research and Effectiveness (CORE), Research Department of Clinical, Educational, and Health Psychology, UCL, London, United Kingdom*
[b] *iCope - Camden and Islington Psychological Therapies Services, Camden & Islington NHS Foundation Trust, London, United Kingdom*
[c] *ADAPT Lab, Research Department of Clinical, Educational and Health Psychology, UCL, London, United Kingdom*
[d] *Waltham Forest Talking Therapies, North East London NHS Foundation Trust, London, United Kingdom*
[e] *Camden and Islington NHS Foundation Trust, London, United Kingdom*

ARTICLE INFO

Keywords:
Measurement invariance
Differential item functioning
GAD-7
PHQ-9
Older adults

ABSTRACT

Background: The nine-item Patient Health Questionnaire (PHQ-9) and seven-item Generalised Anxiety Disorder (GAD-7) scale are widely used clinically and within research, and so it is important to determine how the measures, and individual items within the measures, are answered by adults of differing ages. This study sought to evaluate measurement invariance and differential item functioning (DIF) of the PHQ-9 and GAD-7 between working age and older adults seeking routine psychological treatment.

Methods: Data of working age (18–64 years old) and older (≥65) adults in eight Improving Access to Psychological Therapies (IAPT) services were used. Confirmatory factor analysis (CFA) was used to establish unidimensionality of the PHQ-9 and GAD-7, with multiple-group CFA to test measurement invariance and The Multiple Indicators, Multiple Causes Models approach to assess DIF. The employed methods were applied to a propensity score matched (PSM) sample in sensitivity analyses to control for potential confounding.

Results: Data from 166,816 patients (159,325 working age, 7491 older) were used to show measurement invariance for the PHQ-9 and GAD-7, with limited evidence of DIF and similar results found with a PSM sample ($n = 5868$).

Limitations: The localised sample creates an inability to detect geographical variance, and the potential effect of unmeasured confounders cannot be ruled out.

Conclusions: The findings support the use of the PHQ-9 and GAD-7 measures for working age and older adults, both clinically and in research settings. This study validates using the measures for these age groups to assess clinically significant symptom thresholds, and monitor treatment outcomes between them.

1. Introduction

Depression and anxiety disorders are some of the most commonly presenting mental health problems (Craske and Stein, 2016; Malhi and Mann, 2018). As disorders, they cause substantial individual impairment, are associated with increased direct and indirect healthcare costs, and reduced productivity (Trautmann et al., 2016). For older adults (≥65 years of age), common mental disorders, defined as depression and anxiety conditions, are particularly problematic as they have been associated with increased disability and use of physical health services (Beekman et al., 2002). As populations age (Harper, 2014), the problematic effect of common mental disorders will only continue to worsen.

Meta-analyses of controlled trials have suggested that psychological interventions are equally effective for depression for both older and working-age adults (Cuijpers et al., 2018; Haigh et al., 2018), and that older adults are less likely to benefit from intervention for anxiety disorders (Gould et al., 2012). Despite that, evidence from routine psychological treatment services suggests that older adults are more likely

[*] Corresponding author at: Research Department of Clinical, Educational and Health Psychology, University College London (UCL), 1-19 Torrington Place, London, WC1E 7HB, United Kingdom

E-mail address: henry.delamain@ucl.ac.uk (H. Delamain).

https://doi.org/10.1016/j.jad.2023.11.048
Received 16 June 2023; Received in revised form 9 November 2023; Accepted 15 November 2023
Available online 20 November 2023

to benefit from them than working age adults (Saunders et al., 2021). Whilst there are likely to be some differences in the characteristics of older adults taking part in randomised controlled trials compared to attending routine treatment, there may also be differences in how different age groups interpret measures used to assess treatment effectiveness, resulting in artefactual rather than actual differences in treatment outcomes.

Compared to working age adults, the prevalence of such disorders are reported to generally be less for older adults (Volkert et al., 2013; Wolitzky-Taylor et al., 2010). Whilst this may reflect real differences, there is potential that these two groups interpret commonly used screening tools to measure common mental disorder symptom severity, which if true would cause issues comparing scores between age groups. The Patient Health Questionnaire 9-item depression scale (PHQ-9; Kroenke et al., 2001) and 7-item Generalised Anxiety Disorder scale (GAD-7; Spitzer et al., 2006) are two of the most validated and widely used measures for screening common mental disorders, and evaluating treatment efficacy in research (Kroenke et al., 2010) and clinical practice (Clark, 2018). However, to use such measures in group comparisons it is important to establish that the scales measure their respective constructs consistently across different groups of people.

Measurement invariance (Chen, 2008) and differential item functioning (DIF; Ellis, 1989) are tools that can be used to establish this consistency. Measurement invariance assesses whether an instrument or scale is consistently interpreted between different groups of individuals. DIF will identify whether a given item on a scale is answered differently for one group, compared to another, when the same two groups have the same level of the underlying trait of interest. For example, anhedonia appears to be a more common symptom of depression experienced by older adults, compared to tearfulness or sadness (Gum et al., 2010). The PHQ-9 has been shown to exhibit DIF determined by age group (although the cut-off was $\geq$54 years of age), specifically on items addressing anhedonia, fatigue and low mood (Cameron et al., 2013). Previous research has shown measurement invariance for the PHQ-9 (Patel et al., 2019) and GAD-7 (Shevlin et al., 2022), although not by age group or in a large routine clinical sample. Given the comorbidity of depression and anxiety (Tiller, 2013), there is notably limited research testing DIF of the PHQ-9 and GAD-7 together in a clinical sample, despite the widespread use of these measures.

The aim of this study was to assess the measurement invariance of the PHQ-9 and GAD-7 and DIF of the individual scale items between working age (18–64 years old) and older adults ($\geq$65 years old) seeking psychological therapy for common mental disorders.

2. Method

2.1. Participants

Eight Improving Access to Psychological Therapies (IAPT; now known as NHS Talking Therapies, for anxiety and depression) services provided data on referrals received between January 2011 and August 2020. These services were all members of the North and Central East London IAPT Service Improvement and Research Network (NCEL IAPT SIRN; Buckman et al., 2021; Saunders et al., 2020). These services, grouped together geographically and managed by local NHS Trusts, support the provision of evidence-based psychological treatments for common mental disorders using a stepped-care model (Clark, 2018).

For this analysis, only scores recorded at the initial assessment were included, regardless of whether individuals received formal treatment by the services at later contacts. Further, individuals were included if they had item-level data available for both the PHQ-9 and GAD-7 at their assessment and were at least 18 years of age. Those whose diagnosis (referred to as 'problem descriptor' by services) was recorded as a severe mental illness, such as schizophrenia or substance misuse problems, were also excluded. This is because these primary-care based services do not have standardised treatment protocols for severe mental illness (Clark, 2018), although they can support people with depression or anxiety in the context of a severe mental illness where it is safe to provide care without the input or oversight of a multidisciplinary team. Therefore, if the problem descriptor is recorded as a severe mental illness for people (indicating it is the focus of treatment) then they would have had a different pathway into the services and so will be different from the main analytic sample. To determine the age group comparison in all the analyses, individuals who were 18–64 were recorded as being 'working age' and $\geq$ 65 considered as 'older'.

2.2. Measures

The Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001) is used to measure the degree of depression symptom severity. The nine items within the measure address: anhedonia, low mood, sleep fatigue, appetite, low self-esteem, concentration, psychomotor disturbance and suicidal ideation. The questions are scored between 0 ('not at all') and 3 ('nearly every day') so total scores range between 0 and 27.

The Generalised Anxiety Disorder-7 (GAD-7; Spitzer et al., 2006) is a measure used to assess the severity of symptoms of generalised anxiety disorder, as classified by DSM-IV. The seven items address: nervousness, uncontrollability of worrying, pervasiveness of worrying, issues relaxing, restlessness, irritability and anticipatory fear. The items are scored in the same manner as the PHQ-9, with total scores ranging between 0 and 21.

Both the PHQ-9 and GAD-7 are used by services to measure symptom severity at assessment to identify clinical need, but are also collected on a sessional basis as part of routine outcome measurement to monitor treatment progress. Within the initial assessment, patients answered a range of additional questions to provide sociodemographic and clinical information. As part of this, their age, gender, ethnicity, employment status and whether they are currently prescribed or taking psychotropic medication is recorded. The Index of Multiple Deprivation (IMD; Noble et al., 2006) was also calculated based on the lower layer super output area (LSOA) and collapsed into quintiles, where a lower quintile indicated greater local area deprivation.

2.3. Analysis

To explore the latent factors of depression and anxiety, the evidenced unidimensional structures of each measure were considered (Bianchi et al., 2022; Rutter and Brown, 2017). This is how the factors are commonly considered in clinical practice and research, as positive summative correlations between them exist (Boothroyd et al., 2018; Smith et al., 2020).

In the first instance, confirmatory factor analysis (CFA) was undertaken using the R package 'lavaan' (Rosseel, 2012) for the proposed model (see Fig. 1). Two latent variables that distinctly represented depression and anxiety were constructed using both the PHQ-9 and GAD-7, as well as their correlation (Shevlin et al., 2022).

Three commonly used metrics were calculated to estimate the fit of the CFA model. This included the comparative fit index (CFI) with a threshold for 'good' fit defined as a value of at least 0.95 and an 'acceptable' fit as $>$0.90 (Hu and Bentler, 1999). The root mean squared error of approximation (RMSEA) was also used whereby a threshold of $<$0.05 was taken to indicate a 'close' fit, 0.05–0.08 as an 'acceptable' fit and 0.08–0.10 as a 'moderate' fit (Schermelleh-Engel et al., 2003). Finally, the standardised root mean square residual (SRMR) was estimated with $<$0.05 used to indicate 'good' fit (Hu and Bentler, 1999) and $<$ 0.10 as 'acceptable' (Schermelleh-Engel et al., 2003).

Multiple-group CFA (MGCFA) was used to assess measurement invariance across different groups (Chen, 2008). To do so, several models with increasing strictness were constructed to estimate the level of invariance:

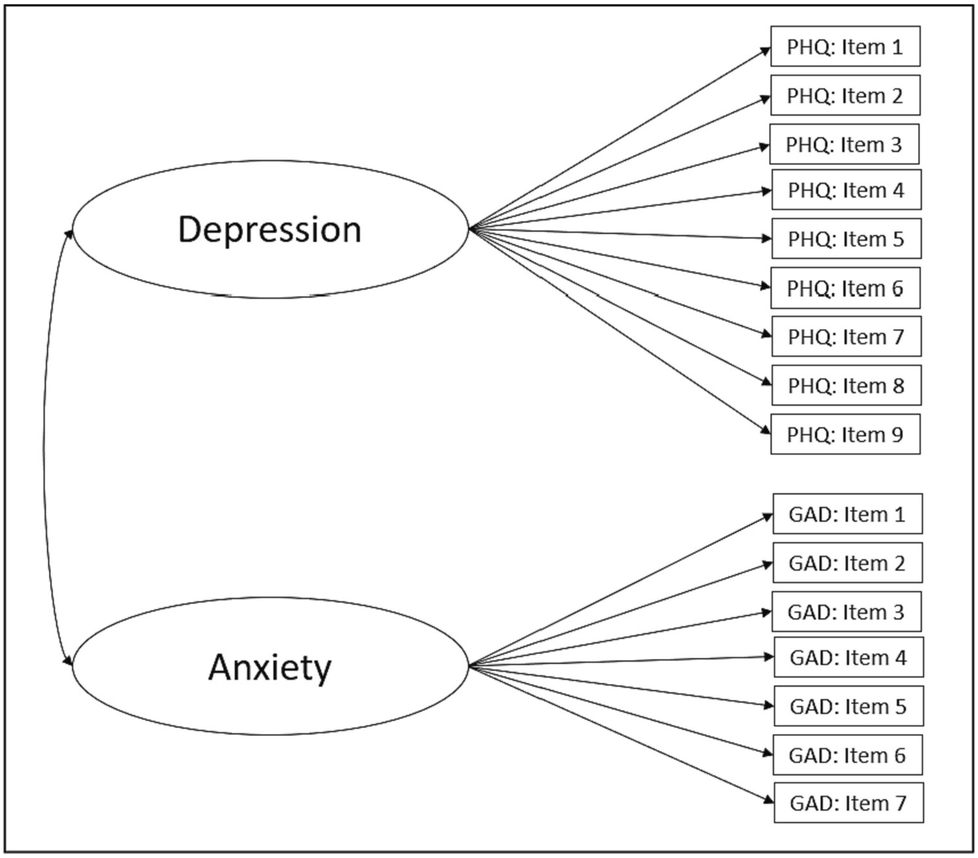


Fig. 1. Proposed model structure.

- M1: Configural invariance (structural equivalence), with the same model structure between groups and free parameters.
- M2: Metric invariance (measurement unit equivalence), with invariance in factor loadings between groups.
- M3: Scalar invariance (full score equivalence), with invariance in factor loadings and intercepts.
- M4: Residual invariance, with invariance in factor loadings, intercepts and residuals.
- M5: Residual invariance, with invariance in factor loadings, intercepts, residuals and factor means.
- M6: Residual invariance, with invariance in factor loadings, intercepts, residuals, factor means and variances.

Evidence of measurement invariance was determined by comparing the change in model fit statistics between the given model (M) and the preceding model (M-1). When measurement invariance had been established, the model could be adopted. To determine this, models were considered within predetermined tolerated ranges: CFI value change was (ΔCFI) <0.01, ΔRMSEA <0.015 and ΔSRMR <0.03 (Cheung and Rensvold, 2002; Shevlin et al., 2022). χ^2 values were recorded, but not used in deciding to adopt a model or not due to issues with doing so when using larger samples (Cheung and Rensvold, 2002).

Differential item functioning (Ellis, 1989) determines differences in individual item scores among certain groups (i.e. age groups) or levels of a variable (such as a depression item as individual PHQ-9 question score, or an anxiety item as individual GAD-7 question score) while considering the overall construct being measured. The Multiple Indicators, Multiple Causes Models (MIMIC; Jöreskog and Goldberger, 1975) approach has been applied to assess DIF in similar previous research (MacIntosh and Hashim, 2003; Shevlin et al., 2022). MIMIC is adapted in the current analyses to explore individual item differences between working age and older adults.

The MIMIC models seek to provide information on (1) the factor loadings of the PHQ-9/GAD-7, (2) the regression coefficients between the predictor variables and the latent variables (that show means differences in the latent variable, based on different levels of the predictor variables) and (3) the direct effects between predictor variables and PHQ-9/GAD-7 items, unaffected by latent variable variability (significant direct effects suggest the presence of DIF). The MIMIC model included two correlated latent variables (depression and anxiety), 16 covariates (the individual items of the PHQ-9/GAD-7) and a single predictor (age group).

Modification indices (MIs) and standardised expected parameter change (SEPC) values were used to decide which direct effects to include within the model. MIs provide an indication of which path could substantially improve the model fit if it was freely estimated, indicated by a reduction of chi-square by >3.84 (the critical value for one degree of freedom, $p < .05$). However, to avoid adding insignificant parameters, a more moderate value of 10 was used to determine potential direct effects based on MI scores. SEPC indicated the estimated value of a fixed parameter and reflected the expected standardised regression coefficient. Since MIs are partially a product of sample size (Chou and Bentler, 1990), the SEPC was used in combination to determine which parameters should be added to the model (Kaplan, 1989). The criteria for adding a direct effect to the model was: MI > 10 and SEPC>0.20 (Shevlin et al., 2022). The model was repeatedly estimated by adding the path with the largest MI/SEPC until there were no MIs/SEPCs >10/0.2. The package 'MplusAutomation' (Hallquist and Wiley, 2018) was used for the MIMIC/DIF related analysis.

Parameters of the model were estimated using robust maximum likelihood estimation (MLR; Tucker and Lewis, 1973), and the same fit statistics as the MGCFA. Chi-square tests of independence were used for categorical variables (with Cramér's v to indicate the magnitude), and independent samples *t*-tests for continuous variables (with Hedge's g to indicate the magnitude).

Differences have been observed in sociodemographic and clinical variables between older and working age individuals attending psychological treatment services at assessment in previous research (Saunders et al., 2021). As such, sensitivity analyses were conducted in which older adults were matched on covariates (excluding age) to working age individuals using propensity score matching (Austin, 2011). Matching was conducted using the ethnicity, gender, mental health service trust, psychotropic medication, IMD quintile, referral year and problem descriptor variables. The PHQ-9 and GAD-7 scores were not used for matching on to avoid impacting the measurement invariance analyses (Saunders et al., 2023). Individuals with missing data on matching covariates were excluded from these sensitivity analyses and matching with replacement was employed, using a narrow caliper of 0.0001 (Gruber et al., 2022). The MGCFA and DIF procedures described above was then repeated for the matched control sample of older adults and their working age matches.

2.4. Ethics

NHS ethical approval was not needed for this study, as confirmed by the Health Research Authority July 2020 #81/81. The IAPT services provided data as part of a service improvement project, and the research adhered to procedures specified by the data hosting providers and was registered with the relevant NHS Trusts overseeing the IAPT services (project reference: 00519-IAPT).

3. Results

3.1. Descriptive statistics

There were 173,578 people with PHQ-9 and GAD-7 item-level data available. Of these, 1315 individuals were <18 years of age or did not have any age data available. Additionally, 5447 were treated for a mental health disorder where there was no standardised IAPT treatment protocol and so were then also excluded. The analytic sample was 166,816 individuals, with 159,325 (95.5 %) of working age adults (18–64 years old) and 7491 (4.5 %) that were older adults (≥65 years old). This is shown in Fig. 2.

The descriptive statistics for the sample used within the analysis are presented in Table 1, separated by age category. Significant group differences were observed on all baseline variables except for gender.

3.2. Confirmatory factor analysis

CFA was applied to the whole sample and then stratified by age range groups. Within the whole sample, model fits were within the acceptable range (RMSEA = 0.079, CFI = 0.907, SRMR = 0.049). Similarly, acceptable metrics were obtained for the working age and older adult age range groups (working age: RMSEA = 0.079, CFI = 0.906, SRMR = 0.049; older adult: RMSEA = 0.074, CFI = 0.917, SRMR = 0.045). Consequently, unidimensionality of the GAD-7 and of the PHQ-9 (as independent scales) was indicated within the models.

3.3. Multiple-group confirmatory factor analysis

The results from the MGCFA are presented in Table 2. Measurement invariance was tested with incremental increases of model strictness from M1 to M6, with changes to model fit statistics being below the criteria values. In the initial model to be tested, configural invariance (M1), similar fit statistics to those seen within the CFA conducted on the whole sample were found. In the metric invariance (M2) model there were sub-criteria changes in fit statistics, indicating that loadings were similar between the working and older age categories. In the scalar invariance (M3) and residual invariance (M4) models, minimal changes were observed in fit statistics. Residual invariance (M4) and factor mean (M5) models led to minimal changes in model fit statistics, with the same for factor variances (M6) included. Consequently, measurement invariance of the GAD-7 and PHQ-9 between working age and older adults was indicated within the model.

3.4. Matched sample multiple-group confirmatory factor analysis

Propensity score matching was undertaken to create a matched sample of working age and older adults who did not have missing covariate data. This led to $n = 24{,}940$ (15.7 %) working age and $n = 1525$ (20.4 %) older adults being excluded. Within the sample of 5966 older adults, there were 98 individuals (1.6 %) for whom adequate matches could not be found and so they were subsequently excluded from these analyses. Therefore, there were 5868 older adults with matched controls (aged 18–64) included as part of the analysis. The significant group differences that were found pre-matching (see Table 1) were non-significant post-matching are shown in Supplementary Materials 1 Table 1.

The measurement invariance results of the matched sample are presented within Supplementary Materials 1 Table 2 and show variable

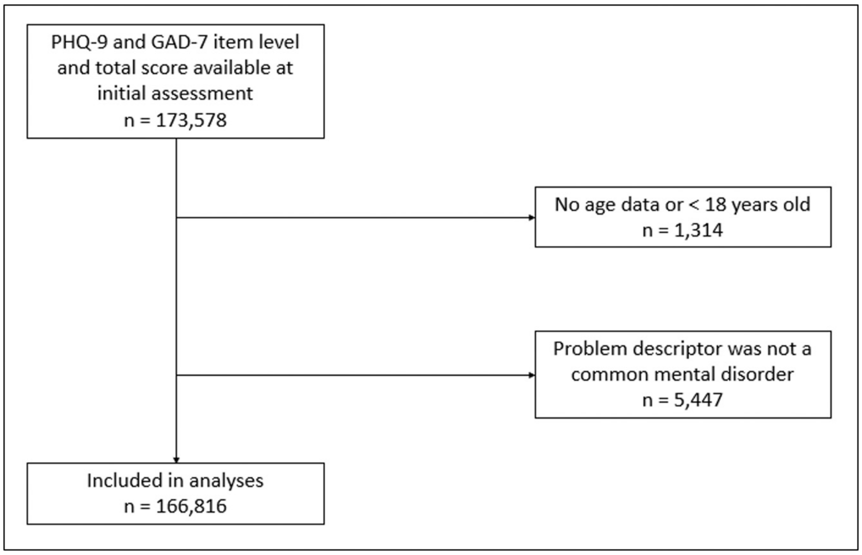

PHQ-9 and GAD-7 item level and total score available at initial assessment
n = 173,578
No age data or < 18 years old
n = 1,314
Problem descriptor was not a common mental disorder
n = 5,447
Included in analyses
n = 166,816

Fig. 2. Patient flow diagram.

Table 1
Sample demographics with group differences.

Sample characteristics		Working age (%)	Older adults (%)	Comparison (*p*-value | Cramér's V)
Overall		159,325 (95.5)	7491 (4.5)	
Age	18–24	25,874 (16.2)	–	–
	25–34	56,414 (35.4)	–	
	35–44	35,907 (22.5)	–	
	45–54	26,503 (16.6)	–	
	55–64	14,627 (9.2)	–	
	65–74	–	5149 (68.7)	
	75–84	–	1942 (25.9)	
	85–94	–	388 (5.2)	
	95+	–	12 (0.2)	
Ethnicity	Asian	18,265 (11.5)	644 (8.6)	<0.001 | 0.061
	Black	19,496 (12.2)	612 (8.2)	
	Chinese	1147 (0.7)	30 (0.4)	
	Mixed	10,125 (6.4)	124 (1.7)	
	Other	6818 (4.3)	212 (2.8)	
	White	95,220 (59.8)	5225 (69.8)	
	Missing	8254 (5.2)	644 (8.6)	
Gender	Female	105,900 (66.5)	4933 (65.9)	0.527 | 0
	Male	52,551 (33.0)	2488 (33.2)	
	Missing	874 (0.5)	70 (0.9)	
Local NHS trust	Trust 1	64,734 (40.6)	3602 (48.1)	<0.001 | 0.059
	Trust 2	32,196 (20.2)	1133 (15.1)	
	Trust 3	40,645 (25.5)	1280 (17.1)	
	Trust 4	21,750 (13.7)	1476 (19.7)	
Psychotropic Medication	Not prescribed	86,873 (54.5)	904 (12.1)	<0.001 | 0.009
	Prescribed and taking	52,625 (33.0)	3830 (51.1)	
	Prescribed not taking	7001 (4.4)	2486 (33.2)	
	Missing	12,826 (8.1)	271 (3.6)	
IMD quintile	1	56,226 (35.3)	1781 (23.8)	<0.001 | 0.093
	2	54,799 (34.4)	2119 (28.3)	
	3	26,828 (16.8)	1593 (21.3)	
	4	15,320 (9.6)	1312 (17.5)	
	5	3895 (2.4)	512 (6.8)	
	Missing	2257 (1.4)	174 (2.3)	
Year of referral	2011	5230 (3.3)	189 (2.5)	<0.001 | 0.017
	2012	8863 (5.6)	326 (4.4)	
	2013	13,159 (8.3)	637 (8.5)	
	2014	15,726 (9.9)	819 (10.9)	
	2015	17,102 (10.7)	818 (10.9)	
	2016	19,236 (12.1)	885 (11.8)	
	2017	21,553 (13.5)	1042 (13.9)	
	2018	25,297 (15.9)	1229 (16.4)	
	2019	25,246 (15.8)	1235 (16.5)	
	2020	7913 (5.0)	311 (4.2)	
Problem descriptor†	Depression	64,116 (40.2)	3171 (42.3)	<0.001 | 0.043
	GAD	20,740 (13.0)	1006 (13.4)	
	Mixed A&D	8946 (5.6)	352 (4.7)	
	OCD	2473 (1.6)	52 (0.7)	
	Other phobia & panic	7144 (4.5)	331 (4.4)	
	PTSD	5309 (3.3)	91 (1.2)	
	Social Phobia	3718 (2.3)	25 (0.3)	
	Unspecified anxiety	999 (0.6)	62 (0.8)	
	Not specified / missing	45,880 (28.8)	2401 (32.1)	

	Mean (±SD)	Mean (±SD)	Comparison (p-value | Hedge's g)
GAD-7 item 1: nervousness	2.12 (0.94)	1.78 (1.09)	<0.001 | -0.356
GAD-7 item 2: uncontrollability of worrying	2.12 (0.98)	1.82 (1.13)	<0.001 | -0.305
GAD-7 item 3: pervasiveness of worrying	2.20 (0.95)	1.84 (1.12)	<0.001 | -0.375
GAD-7 item 4: issues relaxing	1.99 (1.00)	1.54 (1.15)	<0.001 | -0.451
GAD-7 item 5: restlessness	1.23 (1.11)	0.93 (1.10)	<0.001 | -0.268
GAD-7 item 6: irritability	1.81 (1.03)	1.32 (1.12)	<0.001 | -0.483
GAD-7 item 7: anticipatory fear	1.66 (1.13)	1.38 (1.20)	<0.001 | -0.244
GAD-7 (pre-treatment) total score	13.25 (5.29)	10.68 (5.99)	<0.001 | -0.483
PHQ-9 item 1: anhedonia	1.44 (1.04)	1.25 (1.10)	<0.001 | -0.184
PHQ-9 item 2: low mood	1.92 (0.98)	1.70 (1.07)	<0.001 | -0.224
PHQ-9 item 3: sleep	2.02 (1.06)	1.78 (1.18)	<0.001 | -0.228
PHQ-9 item 4: fatigue	2.09 (0.97)	1.86 (1.09)	<0.001 | -0.229
PHQ-9 item 5: appetite	1.58 (1.14)	1.19 (1.19)	<0.001 | -0.343
PHQ-9 item 6: low self-esteem	1.91 (1.06)	1.37 (1.19)	<0.001 | -0.514
PHQ-9 item 7: concentration	1.68 (1.09)	1.24 (1.16)	<0.001 | -0.402
PHQ-9 item 8: psychomotor disturbance	1.05 (1.09)	0.86 (1.08)	<0.001 | -0.173
PHQ-9 item 9: suicidal ideation	0.64 (0.91)	0.47 (0.84)	<0.001 | -0.196
PHQ-9 (pre-treatment) total score	14.75 (6.40)	12.02 (6.72)	<0.001 | -0.425

Notes. †GAD = Generalised anxiety disorder, Mixed A&D = Mixed anxiety and depression, OCD = Obsessive-compulsive disorder, PTSD = Post-traumatic stress disorder.

Table 2
Multiple-group confirmatory factor analysis and fit indices (full sample).

Model	χ2	df	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
M1: Configural invariance	107,741	206	0.906	0.079	0.047	–	–	–
M2: Metric invariance	108,035	220	0.906	0.077	0.047	0	−0.002	0.001
M3: Scalar invariance	109,763	234	0.905	0.075	0.047	−0.001	−0.002	0
M4: Residual invariance	112,637	250	0.902	0.073	0.048	−0.002	−0.001	0
M5: M4 + factor means	113,845	252	0.901	0.074	0.051	−0.001	0	0.003
M6: M5 + factor variances	114,540	254	0.9	0.073	0.051	−0.001	0	0

patterns of change. The greatest contrast is the difference between Scalar Invariance (M3) and Residual Invariance (M4), with ΔCFI -0.018 in the matched sample which exceeds the tolerance range. Further, the move from Residual Invariance (M4) to the restriction of factor means led to a ΔCFI -0.008 and ΔSRMR 0.023 which while below their respective measure tolerance ranges, are both greater differences than observed anywhere in the non-matched sample. These results indicate that there is not sufficient evidence of measurement invariance between the working age and older samples when they are matched on covariates.

3.5. *Differential item functioning*

Initially, the greatest MI and SEPC was a direct effect between age group and the sixth PHQ-9 item (*'Feeling bad about yourself - or that you are a failure or have let yourself or your family down':* MI = 593.561, SEPC = −0.266). The direct effect was added and then the model was re-estimated. This showed the next greatest MI/SEPC values to be between the age group variable and the sixth GAD-7 item (*'Becoming easily annoyed or irritable'*, MI = 443.709, SEPC = −0.241). Once this direct effect had also been added to the model, no variables met the MI/SEPC criteria.

The model with all direct effects added indicated that they all had small magnitudes (Age - > PHQ-9 item 6 = −0.051, p < .001; Age - > GAD-7 item 6 = −0.048, *p* < .001). Additionally, the overall model differences in R-square with the two items included as direct effects was small. The R-square for PHQ-9 item 6 increased from 0.461 to 0.462, suggesting that DIF accounted for 0.001 % of the variance in that item. For GAD-7 item 6, the R-square increased from 0.311 to 0.312, indicating that DIF account for 0.001 % of the variance for that item. Table 3 shows the DIF Model fit statistics.

As part of a sensitivity analysis, the same DIF process was undertaken within the propensity score matched sample and produced similar results. PHQ-9 item 6 and GAD-7 item 6 were identified as items in which there was evidence of DIF; these results are shown in Supplementary Materials 1 Table 3.

Table 3
Differential item functioning model fit statistics for depression and anxiety.

Model	χ2	df	p	CFI	RMSEA	SRMR
Baseline	92,413.52	117	<0.001	0.904	0.069 [0.068 0.069]	0.044
Age - > PHQ-9 item 6	91,917.96	116	<0.001	0.904	0.069 [0.069 0.069]	0.044
Age - > GAD-7 item 6	91,506.68	115	<0.001	0.904	0.069 [0.069 0.069]	0.044

4. Discussion

This study has demonstrated measurement invariance of the PHQ-9 and GAD-7 between working age and older adults in a large sample of individuals seeking treatment for common mental disorders. Differential item functioning has been shown for one item of the PHQ-9 (item 6) and one item of the GAD-7 (item 6), although the effect was minimal. The findings indicate that the same underlying constructs (depression and anxiety) exist for working age and older adults, supporting their use within clinical practice. In the matched sample analysis some potential measurement invariance was detected as was differential item functioning in the same items that were identified in the primary analysis.

The demonstrated measurement invariance of the PHQ-9 and GAD-7 in the unmatched sample provides further validation of the measures and supports their use for adults in clinical practice. This validation is important for the use of these measures as screening and outcome monitoring tools (necessary for service evaluation and performance estimation), as well as their use within research. However, it is noteworthy that in the matched sample analysis, fit statistics exceeded the tolerance ranges, indicating a degree of measurement variance. While limited comparable research exists for the specific age split (≥65 years), similar research with predominantly clinical samples has established measurement invariance for the PHQ-9 (Lamela et al., 2020) and GAD-7 (Moreno et al., 2019).

The results indicate that when controlling for the overall level of anxiety, older adults scored lower on PHQ-9 item 6 (low self-esteem) compared to working age adults. However, the effect size was small and the minimal variance suggests that DIF would not be the likely explanation for group differences for responses to the item. Additionally, when controlling for the overall level of depression older adults scored lower on GAD-7 item 6 (irritability), but the effect size and variance was minimal and not a likely cause for DIF. The detection of potential DIF of these specific items within the measures could be possibly be partly explained by the greater significant group effect size differences, relative to other measure items, recorded for GAD-7 item 6 and PHQ-9 item 6. The findings are generally supported by research showing measurement invariance (and absence of DIF) across multiple sociodemographic variables (Lamela et al., 2020; Moreno et al., 2019).

4.1. *Limitations*

This study should be considered within the context of some limitations. Despite the substantial size of the sample, it is drawn from a localised area and so other sources of geographical variance may not have been detected, limiting the generalisability of the findings. Additionally, although propensity score matching was undertaken, the effect of potential confounding by unmeasured factors cannot be ruled out. Further, the sample was matched on problem descriptors and that may have indirectly constrained GAD-7/PHQ-9 scores, although significant differences were still reported for each continuous variable. The analyses also only tested for uniform rather than non-uniform DIF (where the effect of the independent variable on the item varies depending on the level of the latent variable), but this is in-line with prior research that has consistently demonstrated unidimensionality of the measures.

4.2. Implications

Measurement invariance of the PHQ-9 and GAD-7 in the unmatched sample supports their clinical use for adults of all ages. However, the variance that arose in the matched sample would indicate that there is the potential for bias when comparing scores across groups of working and older adults. Despite this, the very limited magnitude of the effect sizes and minimal variability from the measurement invariance and DIF analysis may not be clinically or individually meaningful to any given patient (Bauer-Staeb et al., 2021), although such thresholds have not been tested across age groups. As tools used in clinical decision making, with the presence of measurement invariance and absence of DIF in the PHQ-9 and GAD-7, the use of uniform measures thresholds between groups can be undertaken with greater confidence. The limited magnitude of any potential effect would also suggest that the measures are suitable tools for routine outcome measuring, as the results do not indicate that alternative measures are needed to compare across age groups.

5. Conclusions

This study observed measurement invariance in an unmatched sample for the PHQ-9 and GAD-7 between working age and older adults, with potential variance detected in a propensity score matched sample. Differential item functioning was minimally detected for two items of both measures and the findings were replicated in a matched sample. These results support their use within clinical practice and research, although future work would possibly seek to test differential item functioning with different covariates such as ethnicity or gender, where intersectionality may impact findings. In addition, to increase the robustness of the findings it would valuable to use geographically different samples, both nationally (in the UK) and internationally (Cromarty et al., 2016; Knapstad et al., 2018).

Financial support

This work was supported by University College London (UCL).

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

CRediT authorship contribution statement

H. Delamain: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Project administration. **J.E.J. Buckman:** Methodology, Investigation, Writing – review & editing. **J. Stott:** Methodology, Writing – review & editing. **A. John:** Methodology, Writing – review & editing. **S. Singh:** Data curation, Writing – review & editing. **S. Pilling:** Data curation, Writing – review & editing. **R. Saunders:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft.

Declaration of competing interest

All authors declare that there are no conflicts of interest.

Acknowledgements

The North and Central East London (NCEL) IAPT Service Improvement and Research Network (SIRN) includes: Ana Cardoso, Andre Lynam-Smith, Catherine Simpson, Evi Aresti, Helmi Van Leur, John Cape, Jon Wheatley, Joshua Cane, Joshua E J Buckman, Judy Leibowitz, Laura Fontaine, Maria Perez, Mina Spatha, Mirko Cirkovic, Renuka Jena, Rob Saunders, Sarah Ellard, Satwant Singh, Stephen Pilling, Syed Ali Naqvi, Tania Knight & Tina Cross. We would like to thank all clinicians and patients from NCEL IAPT services. We are grateful to the service leads for their support with the NCEL project and to the local data managers for their time and dedication. In addition, we would like to thank Adam Kapelner for his support and advice.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2023.11.048.

References

Austin, P.C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar. Behav. Res. 46 (3), 399–424. https://doi.org/10.1080/00273171.2011.568786.

Bauer-Staeb, C., Kounali, D.-Z., Welton, N.J., Griffith, E., Wiles, N.J., Lewis, G., Faraway, J.J., Button, K.S., 2021. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J. Clin. Epidemiol. 137, 200–208. https://doi.org/10.1016/j.jclinepi.2021.04.002.

Beekman, A.T.F., Penninx, B.W.J.H., Deeg, D.J.H., de Beurs, E., Geerling, S.W., van Tilburg, W., 2002. The impact of depression on the well-being, disability and use of services in older adults: a longitudinal perspective. Acta Psychiatr. Scand. 105 (1), 20–27. https://doi.org/10.1034/j.1600-0447.2002.10078.x.

Bianchi, R., Verkuilen, J., Toker, S., Schonfeld, I.S., Gerber, M., Brähler, E., Kroenke, K., 2022. Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. Psychol. Assess. 34 (6), 595–603. https://doi.org/10.1037/pas0001124.

Boothroyd, L., Dagnan, D., Muncer, S., 2018. Psychometric analysis of the generalized anxiety disorder scale and the patient health questionnaire using Mokken scaling and confirmatory factor analysis. Health and Primary Care 2 (4). https://doi.org/10.15761/HPC.1000145.

Buckman, J.E.J., Saunders, R., Cape, J., Pilling, S., 2021. Establishing a service improvement network to increase access to care and improve treatment outcomes in community mental health: a series of retrospective cohort studies. Lancet 398, S28. https://doi.org/10.1016/S0140-6736(21)02571-X.

Cameron, I.M., Crawford, J.R., Lawton, K., Reid, I.C., 2013. Differential item functioning of the HADS and PHQ-9: an investigation of age, gender and educational background in a clinical UK primary care sample. J. Affect. Disord. 147 (1–3), 262–268. https://doi.org/10.1016/j.jad.2012.11.015.

Chen, F., 2008. What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. J. Pers. Soc. Psychol. 95, 1005–1018. https://doi.org/10.1037/a0013193.

Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equ. Model. 9, 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

Chou, C.-P., Bentler, P.M., 1990. Model modification in covariance structure modeling: a comparison among likelihood ratio, Lagrange multiplier, and Wald tests. Multivar. Behav. Res. 25 (1), 115–136. https://doi.org/10.1207/s15327906mbr2501_13.

Clark, D.M., 2018. Realizing the mass public benefit of evidence-based psychological therapies: the IAPT program. Annu. Rev. Clin. Psychol. 14 (1), 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833.

Craske, M.G., Stein, M.B., 2016. Anxiety. Lancet 388 (10063), 3048–3059. https://doi.org/10.1016/S0140-6736(16)30381-6.

Cromarty, P., Drummond, A., Francis, T., Watson, J., Battersby, M., 2016. New access for depression and anxiety: adapting the UK improving access to psychological therapies program across Australia. Australian Psychiatry. https://doi.org/10.1177/1039856216641310 in press.

Cuijpers, P., Reijnders, M., Karyotaki, E., de Wit, L., Ebert, D.D., 2018. Negative effects of psychotherapies for adult depression: a meta-analysis of deterioration rates. J. Affect. Disord. 239, 138–145. https://doi.org/10.1016/j.jad.2018.05.050.

Ellis, B.B., 1989. Differential item functioning: implications for test translations. J. Appl. Psychol. 74, 912–921. https://doi.org/10.1037/0021-9010.74.6.912.

Gould, R.L., Coulson, M.C., Howard, R.J., 2012. Efficacy of cognitive behavioral therapy for anxiety disorders in older people: a Meta-analysis and Meta-regression of randomized controlled trials. J. Am. Geriatr. Soc. 60 (2), 218–229. https://doi.org/10.1111/j.1532-5415.2011.03824.x.

Gruber, J., Lordan, G., Pilling, S., Propper, C., Saunders, R., 2022. The impact of mental health support for the chronically ill on hospital utilisation: evidence from the UK. Soc. Sci. Med. 294, 114675 https://doi.org/10.1016/j.socscimed.2021.114675.

Gum, A.M., McDougal, S.J., McIlvane, J.M., Mingo, C.A., 2010. Older adults are less likely to identify depression without sadness. J. Appl. Gerontol. 29 (5), 603–621. https://doi.org/10.1177/0733464809343106.

Haigh, E.A.P., Bogucki, O.E., Sigmon, S.T., Blazer, D.G., 2018. Depression among older adults: a 20-year update on five common myths and misconceptions. The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry 26 (1), 107–122. https://doi.org/10.1016/j.jagp.2017.06.011.

Hallquist, M.N., Wiley, J.F., 2018. MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus. Struct. Equ. Model. 621–638 https://doi.org/10.1080/10705511.2017.1402334.

Harper, S., 2014. Economic and social implications of aging societies. Science (New York, N.Y.) 346 (6209), 587–591. https://doi.org/10.1126/science.1254405.

Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. Multidiscip. J. 6 (1), 1–55. https://doi.org/10.1080/10705519909540118.

Jöreskog, K.G., Goldberger, A.S., 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Am. Stat. Assoc. 70 (351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485.

Kaplan, D., 1989. Model modification in covariance structure analysis: application of the expected parameter change statistic. Multivar. Behav. Res. 24 (3), 285–305. https://doi.org/10.1207/s15327906mbr2403_2.

Knapstad, M., Nordgreen, T., Smith, O.R.F., 2018. Prompt mental health care, the Norwegian version of IAPT: clinical outcomes and predictors of change in a multicenter cohort study. BMC Psychiatry 18 (1). https://doi.org/10.1186/s12888-018-1838-0. Article 1.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., Löwe, B., 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen. Hosp. Psychiatry 32 (4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006.

Lamela, D., Soreira, C., Matos, P., Morais, A., 2020. Systematic review of the factor structure and measurement invariance of the patient health questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. J. Affect. Disord. 276, 220–233. https://doi.org/10.1016/j.jad.2020.06.066.

MacIntosh, R., Hashim, S., 2003. Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. Appl. Psychol. Meas. 27 (5), 372–379. https://doi.org/10.1177/0146621603256021.

Malhi, G.S., Mann, J.J., 2018. Depression. Lancet 392 (10161), 2299–2312. https://doi.org/10.1016/S0140-6736(18)31948-2.

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.

Noble, M., Wright, G., Smith, G., Dibben, C., 2006. Measuring multiple deprivation at the small-area level. Environment and Planning A: Economy and Space 38 (1). https://doi.org/10.1068/a37168. Article 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 (9), 813–823. https://doi.org/10.1002/da.22940.

Rosseel, Y., 2012. Lavaan: an R package for structural equation modeling. J. Stat. Softw. 48 (2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Rutter, L.A., Brown, T.A., 2017. Psychometric properties of the generalized anxiety disorder Scale-7 (GAD-7) in outpatients with anxiety and mood disorders. J. Psychopathol. Behav. Assess. 39 (1), 140–146. https://doi.org/10.1007/s10862-016-9571-9.

Saunders, R., Cape, J., Leibowitz, J., Aguirre, E., Jena, R., Cirkovic, M., Wheatley, J., Main, N., Pilling, S., Buckman, J.E.J., 2020. Improvement in IAPT outcomes over time: are they driven by changes in clinical practice? Cognitive Behaviour Therapist 13, e16. https://doi.org/10.1017/S1754470X20000173.

Saunders, R., Buckman, J.E.J., Stott, J., Leibowitz, J., Aguirre, E., John, A., Lewis, G., Cape, J., Pilling, S., NCEL network., 2021. Older adults respond better to psychological therapy than working-age adults: evidence from a large sample of mental health service attendees. J. Affect. Disord. 294, 85–93. https://doi.org/10.1016/j.jad.2021.06.084.

Saunders, R., Moinian, D., Stott, J., Delamain, H., Naqvi, S.A., Singh, S., Wheatley, J., Pilling, S., Buckman, J.E.J., 2023. Measurement invariance of the PHQ-9 and GAD-7 across males and females seeking treatment for common mental health disorders. BMC Psychiatry 23 (1), 298. https://doi.org/10.1186/s12888-023-04804-x.

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., 2003. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods of Psychological Research 8 (2), 23–74.

Shevlin, M., Butter, S., McBride, O., Murphy, J., Gibson-Miller, J., Hartman, T.K., Levita, L., Mason, L., Martinez, A.P., McKay, R., Stocks, T.V., Bennett, K.M., Hyland, P., Vallieres, F., Valiente, C., Vazquez, C., Contreras, A., Peinado, V., Trucharte, A., Bentall, R.P., 2022. Measurement invariance of the patient health questionnaire (PHQ-9) and generalized anxiety disorder scale (GAD-7) across four European countries during the COVID-19 pandemic. BMC Psychiatry 22 (1), 154. https://doi.org/10.1186/s12888-022-03787-5.

Smith, M., Francq, B., McConnachie, A., Wetherall, K., Pelosi, A., Morrison, J., 2020. Clinical judgement, case complexity and symptom scores as predictors of outcome in depression: an exploratory analysis. BMC Psychiatry 20 (1), 125. https://doi.org/10.1186/s12888-020-02532-0.

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 (10) https://doi.org/10.1001/archinte.166.10.1092. Article 10.

Tiller, J.W.G., 2013. Depression and anxiety. Med. J. Aust. 199 (6). https://www.mja.com.au/journal/2013/199/6/depression-and-anxiety.

Trautmann, S., Rehm, J., Wittchen, H., 2016. The economic costs of mental disorders: do our societies react appropriately to the burden of mental disorders? EMBO Rep. 17 (9), 1245–1249. https://doi.org/10.15252/embr.201642951.

Tucker, L.R., Lewis, C., 1973. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 38 (1), 1–10. https://doi.org/10.1007/BF02291170.

Volkert, J., Schulz, H., Härter, M., Wlodarczyk, O., Andreas, S., 2013. The prevalence of mental disorders in older people in Western countries—a meta-analysis. Ageing Res. Rev. 12 (1), 339–353. https://doi.org/10.1016/j.arr.2012.09.004.

Wolitzky-Taylor, K.B., Castriotta, N., Lenze, E.J., Stanley, M.A., Craske, M.G., 2010. Anxiety disorders in older adults: a comprehensive review. Depress. Anxiety 27 (2), 190–211. https://doi.org/10.1002/da.20653.