BMJ Mental Health

ADULT MENTAL HEALTH

International Depression Questionnaire and International Anxiety Questionnaire: validation of brief ICD-11 measures for depression and generalised anxiety disorder

Johanna Schröder ,[1] Leonhard Kratzer ,[2] Thanos Karatzias,[3] Anamaria Semm,[4] Stefan Tschöke,[5,6] Sarah Biedermann,[7] Eva Schäflein,[8] Julia König,[9] Matthias Knefel,[10] Philip Hyland,[11] Mark Shevlin[12]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjment-2025-302389).

For numbered affiliations see end of article.

Correspondence to
Dr Mark Shevlin; m.shevlin@ulster.ac.uk

JS and LK contributed equally.

JS and LK are joint first authors.

Received 4 December 2025
Accepted 1 March 2026

To cite: Schröder J, Kratzer L, Karatzias T, *et al*. *BMJ Ment Health* 2026;**29**:1–8.

ABSTRACT

Background The 11th revision of the International Classification of Diseases (ICD-11) introduced revised diagnostic criteria for a depressive episode (DE) and generalised anxiety disorder (GAD). The International Depression Questionnaire (IDQ) and the International Anxiety Questionnaire (IAQ) are the first self-report measures developed to assess and screen these disorders according to the ICD-11 diagnostic rules.

Objective This study aims to validate the IDQ and the IAQ in clinical and community samples, examining internal consistency, factorial validity and construct validity.

Methods The cross-sectional, observational multicentre validation study applied internal consistency testing, confirmatory factor analyses and item response theory (IRT) in a clinical sample (n=569; age 18–73; 417 females, 118 males, 34 diverse) and a sample representative of the German general population (n=1001) by age, education and gender (500 females, 499 males, 2 diverse). Factorial and IRT model fit of the IDQ and IAQ as well as concordance with the Patient Health Questionnaire-9 (PHQ-9) and the Generalised Anxiety Disorder-7 (GAD-7) was tested.

Results Both questionnaires showed excellent internal consistency (ω=0.96 each) and strong factor loadings. A three-factor IDQ model and a one-factor IAQ model provided the best fit. In the clinical sample, 39.7% met ICD-11 DE criteria and 51.0% GAD criteria (overlap 32.2%). In the general population, prevalence was 5.9% for DE and 9.3% for GAD. Concordance with PHQ-9 and GAD-7 was partial, suggesting differences between ICD-11-based and established screening tools.

Conclusions The IDQ and IAQ are psychometrically robust self-report measures for ICD-11 DE and GAD. They are reliable, valid, brief, easy to administer, cost-free and suitable for use in primary care and research across diverse clinical and research settings. Their availability supports standardised screening of depression and GAD in both clinical and community settings.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Self-report measures are widely used in primary care to assess anxiety and depression.
⇒ The IDQ and IAQ are the first instruments aligned with ICD-11 diagnostic rules for depressive episode and generalised anxiety disorder.

WHAT THIS STUDY ADDS

⇒ The IDQ and IAQ demonstrated excellent reliability and validity in the assessment of ICD-11 depressive episodes and generalised anxiety disorder in the general population as well as clinical samples, demonstrating their potential to enhance symptom screening in primary care.
⇒ Prevalence estimates derived from the IDQ and IAQ were consistent with the Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental disorders but diverged from commonly used screening tools (Patient Health Questionnaire-9, Generalised Anxiety Disorder-7), highlighting important differences in case identification.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Using IDQ and IAQ may improve ICD-11 aligned screening in clinical research and practice.
⇒ ICD-11 aligned screening using IDQ and IAQ may reduce overestimation of DE and GAD.

BACKGROUND

Low content overlap and limited diagnostic concordance across depression and anxiety measures indicate construct heterogeneity, which can bias results, impair replication and limit comparability across studies.[1 2] The Patient Health Questionnaire-9 (PHQ-9)[3] and the Generalised Anxiety Disorder-7 (GAD-7)[4] were developed to represent the diagnostic criteria for a depressive episode (DE) and generalised anxiety disorder (GAD), respectively, in the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV).[5] The alignment of the PHQ-9 and GAD-7 items to the DSM-IV diagnostic criteria has likely been responsible for their widespread use over the last two decades, as this approach enhances both comparability across

studies and clinical relevance. They have been widely used in epidemiological studies to estimate the prevalence of DE and GAD in the general population[6] and high-risk groups,[7] and are also employed to screen individuals for clinical services and to monitor treatment outcomes.[8] The PHQ-9 and GAD-7 have become industry-standard screening tools using cut-off scores to identify probable cases for DE and GAD. However, the items of PHQ-9 and GAD-7, or any other self-report measure, no longer align with the symptom specification of DE or GAD in the DSM-5-TR and, until recently, the 11th revision of the International Classification of Diseases (ICD-11). Further, a recent study identified widespread misinterpretation of the PHQ instructions across community and clinical samples, raising doubts about its validity for both research and clinical decision-making.[9]

There were no self-report instruments that can screen DE or GAD on the basis of current ICD-11 (or DSM-5-TR) diagnostic criteria. In order to provide researchers and clinicians with self-report measures that are consistent with a major diagnostic system, the International Depression Questionnaire (IDQ) and the International Anxiety Questionnaire (IAQ) were developed to screen the symptoms of DE and GAD as per the descriptions in ICD-11[10]; these measures can provide severity scores and generate provisional caseness for DE and GAD based on the ICD-11 diagnostic rules.

The initial development of the IDQ and IAQ[10] followed a bottom-up validation strategy. This iterative process began with proof of concept studies in diverse populations like university students[11] as well as specific high-risk or high-prevalence populations including bereaved individuals[12] and trauma-exposed military samples[13] to ensure robust item performance and sensitivity in contexts where symptom severity is typically high. These preliminary studies, which included various linguistic translations,[11 13] demonstrated strong psychometric foundations: both scales showed evidence of unidimensionality, high reliability (IDQ: ω=0.96; IAQ: ω=0.96), and strong convergent validity with the PHQ-9 and GAD-7.[10]

The present study represents the next necessary phase in this validation hierarchy. Moving beyond specific subgroups, we now evaluate the instruments within large-scale, heterogeneous general population and diverse clinical samples to establish broader generalisability. Furthermore, this study addresses recent updates in the WHO's Clinical Descriptions and Diagnostic Requirements (CDDR),[14] which imply a three-factor structure (affective, cognitive-behavioural and neurovegetative) of DE rather than the unidimensional model previously tested.

The aim of this study was to assess the reliability, factorial validity and concurrent validity of the scale scores in a clinical as well as a general population sample, testing the following hypotheses: (1) The IDQ and IAQ will demonstrate excellent internal consistency in both clinical and general population samples, as indicated by McDonald's ω≥0.90. (2) In both the clinical and in the general population samples, a one-factor model of the IDQ will demonstrate at least acceptable fit as assessed by confirmatory factor analysis (CFA) and item response theory (IRT), and an additional three-factor model of the IDQ will also show at least acceptable fit. (3) In both the clinical and in the general population sample, a one-factor model of the IAQ will provide at least acceptable fit, as assessed by CFA and IRT. (4) IDQ and PHQ-9 caseness will show a strong and statistically significant association (Cramer's V≥0.40) in both samples, reflecting substantial concordance of the two measures. (5) IAQ and GAD-7 caseness will show a strong and statistically significant association (Cramer's V≥0.40) in both samples, reflecting substantial concordance of the two measures. Further, (6) comorbidity of DE and GAD based on IDQ and IAQ is explored.

METHODS

Study design and procedure

This study was conducted as a cross-sectional, observational, multicentre validation study. Two anonymous, cross-sectional online surveys were conducted via the survey platform Tivian/unipark.[15]

Participants and recruitment

A total of 569 participants were recruited at seven recruitment centres, including psychiatric, psychosomatic and psychotherapeutic inpatient and outpatient treatment centres across Germany from July 2023 until November 2024. Mental health staff, as well as flyers and posters, invited the patients of the recruitment centres to participate in the study. Inclusion criteria were an age of at least 18 years, being recruited in one of the seven recruitment centres and reporting having a psychiatric diagnosis at the time of the survey. The German general population sample, representative by gender, age and education (n=1001), was recruited with the support of the market research institute Civey in 1 week in May 2025. The inclusion criterion was an age of at least 18 years. Sample characteristics can be obtained from table 1.

The clinical sample structure reflects a typical cross-section of the German mental healthcare system. It comprises patients from various diagnostic groups and treatment intensities (inpatient and outpatient). Detailed sociodemographic and clinical characteristics, including the distribution of reported diagnoses, are summarised in table 1 and online supplemental table S1. 450 (79.1%) patients reported a diagnosis of depressive disorder and 99 (17.4%) patients reported a GAD.

Measures

The IDQ[10] is a nine-item (plus one binary functional impairment item) self-report measure designed to assess DE in accordance with the ICD-11 guidelines. Participants rate symptom frequency over the past 2 weeks on a 5-point Likert scale (0=Never to 4=Every day). Items 1 and 2 reflect essential symptoms (depressed mood or diminished interest in activities), while items 3–9 address additional symptoms. There is also a question on functional impairment across personal, family, social, educational, occupational and other important domains. To fulfil diagnostic criteria of DE, individuals must endorse (score of ≥3) five or more items including at least one essential symptom (items 1 or 2). Additionally, functional impairment due to these symptoms must be present (ie, a response of 'Yes') for diagnosis. Possible summed scores range from 0 to 36 with higher scores indicating greater levels of depression. The IDQ has demonstrated excellent internal reliability (ω=0.96).[10]

The IAQ[10] is an eight-item (plus one binary functional impairment item) self-report measure that evaluates GAD based on ICD-11 requirements. Participants report symptom frequency over the past several months using the same Likert scale the IDQ uses. Items 1 and 2 measure essential symptoms (general apprehension or excessive worry), while items 3–8 capture additional symptoms. A final item assesses functional impairment across various life domains. To meet the diagnostic criteria of GAD, at least one core symptom (item 1 or 2) and a total of four or more symptoms must have a Likert score of 3 or higher. Additionally, a separate item assessing functional impairment caused by these symptoms must be answered 'Yes' to confirm the diagnosis.

Table 1 Sociodemographic and clinical characteristics of the clinical sample (n=569) and the general population sample (n=1001)

	Clinical sample		General population sample	
Age	M=36.0	SD=13.7	M=50.86	SD=15.49
	n	%	n	%
Sex				
Female	446	78.4	500	49.95
Male	118	20.7	499	49.85
Other	5	0.9	2	0.2
Gender				
Female	417	73.3	494	49.35
Male	118	20.7	504	50.35
Diverse/non-binary	34	6.0	3	0.3
Nationality (multiple answers possible)				
German	543	95.43	968	96.7
Other	39	6.85	33	3.3
Population of place of residence				
<2000	42	7.4	101	10.09
2000–5000	53	9.3	96	9.59
5000–20 000	112	19.7	210	20.98
20 000–100 000	96	16.9	257	25.67
100 000–500 000	34	6.0	149	14.89
>500 000	232	40.8	188	18.78
Highest level of general education				
Still in school, no qualification yet	1	0.2	3	0.3
Left school without a qualification	2	0.4	6	0.6
Special education diploma	1	0.2	2	0.2
Secondary general school diploma	36	6.3	200	19.98
Intermediate secondary school diploma	154	27.1	280	27.97
High school diploma/university entrance qualification	375	65.9	510	50.95
Employment status				
Full-time	159	27.9	436	43.56
Part-time	113	19.9	143	14.29
Mini-job	52	9.1	22	2.2
Unemployed, seeking work	62	10.9	41	4.1
Unemployed, caregiving/housework	14	2.5	48	4.8
Unemployed, early retirement due to illness	69	12.1	64	6.39
Unemployed, retirement	13	2.3	218	21.78
Unemployed, student	87	15.3	29	2.9
Monthly income				
<1000 €	146	25.7	69	6.89
1000 €–2000 €	183	32.2	223	22.28
2000 €–3000 €	103	18.1	272	27.17
3000 €–4000 €	60	10.5	178	17.78
4000 €–5000 €	42	7.4	132	13.19
>5000 €	35	6.2	127	12.69
Ever received psychotherapeutic treatment				
Yes, currently and in the past	284	49.9	62	6.19
Yes, in the past	144	25.3	314	31.37
Yes, currently	100	17.6	41	4.1
No, never	41	7.2	584	58.34
Ever received psychopharmacological treatment				
Yes, currently and in the past	221	38.8	89	8.89
Yes, in the past	124	21.8	181	18.08
Yes, currently	95	16.7	58	5.79
No, never	129	22.7	673	67.23

Sex was assessed as a biological classification assigned at birth. Gender was assessed as self-identified social identity.

Possible summed scores range from 0 to 36, with higher scores reflecting more severe anxiety. The IAQ has shown excellent internal reliability (ω=0.96).[10]

The German translation of the IDQ and IAQ followed the principles of good practice for the translation and cultural adaptation process for patient-reported outcomes of the ISPOR task force for translation and cultural adaptation.[16] These principles comprise, among others, the respect of copyright, the involvement of the developers of an instrument, the recruitment of key in-country persons to the project, the development of at least two independent forward translations, the reconciliation of forward translations and back translations, the critical review, the harmonisation of translations and proofreading. For the German versions of the IDQ and IAQ, we refer to online supplemental tables S5 and S6.

The PHQ-9 is a self-administered version of the PRIME-MD screening instrument for common mental disorders.[3 17] The PHQ-9 is the depression module, which scores at each of the 9 DSM-IV criteria on a 4-point Likert scale from '0' (not at all) to '3' (nearly every day).[3] As a severity measure, scores range from 0 to 27 and represent: minimal (<5), mild (5–9), moderate (10–14), moderately severe (15–19) and severe depression (≥20). As a screening instrument, major depression is diagnosed if five or more of the nine depressive symptom criteria have been present at least 'more than half the days' in the past 2 weeks, and one of the symptoms is depressed mood or anhedonia. Using a clinical interview as criterion standard, a PHQ-9 score ≥10 had a sensitivity of 88% and a specificity of 88% for major depression, and internal consistency was excellent with α=0.89.[3]

The GAD-7[4 18] screens the DSM-IV criteria for GAD on a 4-point Likert-scale from '0' (not at all) to '3' (nearly every day). As a severity measure, scores range from 0 to 21 and represent: minimal (<5), mild (5–9), moderate (10–14) and severe anxiety (≥15). As a screening instrument, a meta-analysis found that GAD-7 achieved acceptable accuracy at a cut-off point of 8 (sensitivity: 0.83, specificity: 0.84, pooling 12 samples and 5223 participants).[19] A validation study in the general population showed excellent internal consistency with α=0.89.[6]

Data analysis

All analyses were conducted separately on the data from the clinical and general population samples. First, the distributions of all IDQ and IAQ items were examined; the percentages of each response category were reported, along with the mean score and percentage endorsement (score ≥3) as summary statistics. Item–total correlations were also calculated and expected to exceed the minimum acceptable value of ≥0.30.[20] Factorial validity was examined using both CFA and IRT. CFA was used to test theoretically specified factor structures at the scale level and to evaluate their alignment with ICD-11-based conceptualisations. IRT analyses were conducted to examine item-level properties, account for categorical response formats and potential deviations from normality, and assess measurement precision across the latent trait continuum. The IDQ and IAQ are used to assess severity using the sum score of all the items. Therefore, CFA models were used based on the full Likert-scale responses to test the factorial validity of one-factor solutions for both scales as well as a three- factor solution for the IDQ. As the scales can also be used to determine caseness based on dichotomised (≥3) item scores, 2-parameter IRT models were fitted to test one-factor models of the IDQ and IAQ, and the correlated three-factor model of the IDQ that is implied by the CDDR (factors representing affective, cognitive-behavioural, neurovegetative clusters). The CFA models were estimated using robust full information maximum likelihood based on the variance/covariance matrix of the Likert scores, and the IRT was estimated using the robust weighted least squares estimator with a non-linear probit link based on the tetrachoric correlation matrix of latent continuous response variables.

The IRT model was a 2-parameter model with discrimination and difficulty parameters estimated for all items. The discrimination parameter (α) is the probit regression that relates the latent variable, theta (*q*), to the binary indicator where higher values indicate increased discriminatory power. The difficulty parameter is the point on the underlying scale at which a participant has a probability of endorsing the item. The probit estimates were transformed into standard IRT parameters and can be interpreted according to magnitude[21]: very low (0.01<α<0.34), low (0.35<α<0.64), moderate (0.65<α<1.34), high (1.35<α<1.69) and very high (α>1.70). There was no previous evidence to support the 1-parameter model.

Model fit was assessed using common fit indices, and standard criteria were used to determine acceptable model fit: a non-significant χ^2, Comparative Fit Index (CFI) and Tucker-Lewis Index values ≥0.90, and root mean square error of approximation (RMSEA) and standardised root mean residual values ≤0.08 and ≤0.05 indicating acceptable and excellent model fit, respectively.[22] For the CFA models, the Bayesian Information Criterion (BIC) was also available and used to compare models, with lower values indicating superior fit. Reliability was assessed using McDonald's omega (ω), where values can range from 0 to 1 and higher values indicate stronger internal reliability.[23]

Third, the probable prevalence estimates for ICD-11 DE and ICD-11 GAD were calculated using the ICD-11 diagnostic algorithms. If the three-factor CFA/IRT models were found to be well-fitting models, a diagnostic algorithm would also be applied that recognised the CDDR multicluster definition of MDD. This alternative algorithm also required a total of five symptoms to be endorsed, but also that one should be from each of the three clusters. If the ICD-11 and alternative algorithm did not produce divergent estimates (the ICD-11 estimate should be within the 95% CI of the estimate of alternative algorithm), the analyses would proceed with the ICD-11-based estimates. Prevalence estimates for depression and anxiety based on the PHQ-9 and the GAD-7 cut-off score of ≥10[24] were compared with the estimates obtained from the IDQ and the IAQ. The effect sizes Cramer's V and Cohen's d were calculated using the R package effect size.[25] Effect sizes are categorised as small (0.1), medium (0.3) or large (0.5). The Venn diagram of the clinical sample was built using the R package VennDiagram.[26]

FINDINGS

Descriptive statistics

The item-level response distributions, means, SD and item–total correlations for the IDQ and IAQ for both samples are reported in tables 2 and 3. For the clinical sample, the sum scores for the IDQ covered the entire range of possible scores (0–36) with a mean of 18.58 (SD=8.43). The sum scores for the IAQ almost covered the entire range of possible scores (1–32) with a mean of 18.10 (SD=7.32). For the general population sample, the sum scores for the IDQ covered the entire range of possible scores (0–36) with a mean of 7.33 (SD=7.61). The sum scores for the IAQ covered the entire range of possible scores (1–32) with a mean of 7.63 (SD=7.55).

Table 2 Clinical sample: item response distributions, means, SD and item–total correlations for the IDQ and IAQ

	Scale value							
	0 Never	1 Only a few days	2 Half the days	3 Most days	4 Every day	% Endorsed	Mean (SD)	Item–total correlation
Depression								
Felt down or depressed for most of the day?	6.3%	30.4%	20.2%	28.3%	14.8%	43.1%	2.15 (1.18)	0.80
Experienced less interest or pleasure from normal activities for most of the day?	8.1%	25.3%	21.6%	30.9%	14.1%	45.0%	2.18 (1.19)	0.76
Have had difficulty concentrating?	4.6%	25.0%	19.7%	27.4%	23.4%	50.8%	2.40 (1.21)	0.56
Had feelings of worthlessness or guilt?	13.2%	25.1%	16.3%	24.4%	20.9%	45.3%	2.15 (1.35)	0.72
Felt hopeless?	14.8%	26.4%	17.2%	26.0%	15.6%	41.7%	2.01 (1.32)	0.75
Had recurrent thoughts of death or suicide?	40.9%	31.3%	11.1%	10.4%	6.3%	16.7%	1.10 (1.22)	0.58
Have had changes in appetite or sleep?	13.5%	22.7%	21.4%	26.5%	15.8%	42.4%	2.08 (1.28)	0.56
Moved slower or felt more restless?	14.6%	25.8%	21.4%	26.5%	11.6%	38.1%	1.95 (1.25)	0.63
Experienced reduced energy or fatigue?	4.6%	18.8%	18.6%	31.5%	26.5%	58.0%	2.57 (1.19)	0.73
Anxiety								
Felt nervous or anxious?	7.7%	27.1%	19.7%	28.6%	16.9%	45.5%	2.20 (1.22)	0.74
Worried a lot about different things?	3.2%	16.5%	16.3%	31.8%	32.2%	64.0%	2.73 (1.16)	0.67
Felt physically tense or agitated?	5.4%	18.6%	19.9%	28.5%	27.6%	56.1%	2.54 (1.22)	0.78
Felt your heart racing, difficulty breathing, stomach discomfort or dry mouth?	24.6%	31.8%	16.9%	18.3%	8.4%	26.7%	1.54 (1.27)	0.61
Felt 'on edge'?	3.5%	18.6%	18.8%	30.9%	28.1%	59.1%	2.62 (1.17)	0.77
Had difficulty concentrating?	6.0%	25.1%	19.2%	25.5%	24.3%	49.7%	2.37 (1.25)	0.57
Been easily annoyed by different things?	11.4%	33.7%	20.4%	22.3%	12.1%	34.4%	1.90 (1.22)	0.49
Experienced sleep disturbances?	14.8%	24.3%	13.5%	20.0%	27.4%	47.5%	2.21 (1.44)	0.50

IAQ, International Anxiety Questionnaire; IDQ, International Depression Questionnaire.

Reliability and factor validity

The internal consistency of the IDQ (clinical ω=0.96; population ω=0.94) and the IAQ (clinical ω=0.96; population ω=0.95) scale scores were both high.

Table 4 shows the fit statistics for the CFA and IRT models for both samples. In the clinical sample, the IDQ one-factor CFA model showed acceptable comparative fit (CFI=0.916) but suboptimal absolute fit (RMSEA=0.109). The IDQ

Table 3 General population sample: item response distributions, means, SD and item–total correlations for the IDQ and IAQ

	Scale value							
	0 Never	1 Only a few days	2 Half the days	3 Most days	4 Every day	% Endorsed	Mean (SD)	Item–total correlation
Depression								
Felt down or depressed for most of the day?	42.6%	37.6%	9.3%	8.7%	1.9%	10.6%	0.90 (1.01)	0.83
Experienced less interest or pleasure from normal activities for most of the day?	39.8%	38.9%	11.6%	8.0%	1.8%	9.8%	0.93 (1.00)	0.80
Have had difficulty concentrating?	40.8%	35.5%	12.7%	8.2%	2.9%	11.1%	0.97 (1.06)	0.75
Had feelings of worthlessness or guilt?	61.3%	21.4%	7.6%	6.4%	3.3%	9.7%	0.69 (1.07)	0.80
Felt hopeless?	55.5%	24.2%	8.6%	9.1%	2.6%	11.7%	0.79 (1.09)	0.83
Had recurrent thoughts of death or suicide?	79.2%	12.4%	4.3%	3.2%	0.9%	4.1%	0.34 (0.78)	0.62
Have had changes in appetite or sleep?	50.0%	30.5%	9.0%	7.1%	3.4%	10.5%	0.83 (1.07)	0.72
Moved slower or felt more restless?	50.9%	30.1%	11.2%	5.9%	1.9%	7.8%	0.78 (0.99)	0.79
Experienced reduced energy or fatigue?	34.0%	40.0%	12.6%	9.2%	4.3%	13.5%	1.10 (1.10)	0.79
Anxiety								
Felt nervous or anxious?	50.2%	30.5%	9.8%	6.6%	2.9%	9.5%	0.81 (1.04)	0.81
Worried a lot about different things?	28.1%	41.2%	12.7%	11.6%	6.5%	18.1%	1.27 (1.18)	0.81
Felt physically tense or agitated?	39.2%	36.3%	11.8%	9.2%	3.6%	12.8%	1.02 (1.10)	0.86
Felt your heart racing, difficulty breathing, stomach discomfort or dry mouth?	64.8%	19.9%	7.9%	5.1%	2.3%	7.4%	0.60 (0.99)	0.74
Felt 'on edge'?	38.7%	37.5%	10.9%	9.0%	4.0%	13.0%	1.02 (1.10)	0.87
Had difficulty concentrating?	49.1%	31.0%	9.4%	7.0%	3.6%	10.6%	0.85 (1.08)	0.81
Been easily annoyed by different things?	45.1%	32.8%	9.9%	8.2%	4.1%	12.3%	0.94 (1.11)	0.79
Experienced sleep disturbances?	39.8%	33.2%	9.8%	10.1%	7.2%	17.3%	1.12 (1.24)	0.73

IAQ, International Anxiety Questionnaire; IDQ, International Depression Questionnaire.

Table 4 Fit statistics for the alternative IDQ and IAQ models in the clinical sample and the general population sample

Model (clinical sample)	χ^2	df	*P*	CFI	TLI	RMSEA	SRMR	BIC
IDQ one-factor CFA	209.650	27	<0.001	0.916	0.888	0.109 (0.096, 0.123)	0.051	14365.540
IDQ three-factor CFA	139.748	24	<0.001	0.947	0.920	0.092 (0.078, 0.107)	0.051	14293.001
IDQ one-factor IRT	128.378	27	<0.001	0.977	0.969	0.081 (0.067, 0.096)	0.066	
IDQ three-factor IRT	96.626	24	<0.001	0.983	0.975	0.073 (0.058, 0.088)	0.057	
IAQ one-factor CFA	101.467	20	<0.001	0.953	0.934	0.085 (0.069, 0.101)	0.038	12953.886
IAQ one-factor IRT	41.298	20	<0.001	0.995	0.993	0.043 (0.024, 0.062)	0.039	
Model (general population sample)	χ^2	df	p	CFI	TLI	RMSEA	SRMR	BIC
IDQ one-factor CFA	218.230	27	<0.001	0.947	0.930	0.084 (0.074, 0.095)	0.032	19290.308
IDQ three-factor CFA	130.235	24	<0.001	0.971	0.956	0.066 (0.056, 0.078)	0.028	19136.484
IDQ one-factor IRT	63.017	27	<0.001	0.996	0.994	0.037 (0.025, 0.048)	0.035	
IDQ three-factor IRT	41.272	24	<0.001	0.998	0.997	0.027 (0.012, 0.040)	0.030	
IAQ one-factor CFA	83.057	20	<0.001	0.984	0.977	0.056 (0.044, 0.069)	0.019	17687.025
IAQ one-factor IRT	39.845	20	<0.001	0.998	0.998	0.031 (0.017, 0.046)	0.025	

BIC, Bayesian Information Criterion; CFA, confirmatory factor analysis; CFI, Comparative Fit Index; IAQ, International Anxiety Questionnaire; IDQ, International Depression Questionnaire; IRT, item response theory; RMSEA, root mean square error of approximation; SRMR, standardised root mean residual; TLI, Tucker-Lewis Index.

three-factor CFA model demonstrated improved fit across indices (CFI=0.947; RMSEA=0.092) and was strongly preferred based on the BIC (ΔBIC≈73). IRT analyses indicated good fit for both models, with consistently better fit for the three-factor solution (CFI=0.983 vs 0.977; RMSEA=0.073 vs 0.081). In the general population sample, the IDQ three-factor CFA model again showed superior fit (CFI=0.971; RMSEA=0.066) compared with the one-factor model (CFI=0.947; RMSEA=0.084), with a substantial BIC advantage (ΔBIC≈154). Both IRT models demonstrated excellent fit, with near-perfect indices for the three-factor model (CFI=0.998; RMSEA=0.027). For the IAQ, a one-factor structure was tested. In the clinical sample, the one-factor CFA demonstrated acceptable fit (CFI=0.953; RMSEA=0.085), while the IRT model showed excellent fit (CFI=0.995; RMSEA=0.043). In the general population sample, both CFA and IRT models indicated good to excellent fit (CFA RMSEA=0.056; IRT RMSEA=0.031).

For both samples, the factor loadings were high, positive and statistically significant in the CFA models, and for the IRT model, the discrimination parameters were all significant and 'moderate' to 'very high' in magnitude. For the clinical sample, the factor correlations for the three-factor CFA models (r=0.83–0.91) and IRT model (r=0.85–0.92) were high, but slightly lower than for the general population sample (CFA r=0.91–0.92; IRT r=0.92–0.94). Estimates for all models are available in supplementary materials (online supplemental table S2–S5).

Depression caseness and measure concordance

In the clinical sample using the ICD-11 diagnostic algorithm, 226 patients (39.7%: 95% CI 35.7% to 43.7%) met screening criteria for ICD-11 DE. The estimate of DE using the new CDDR-aligned ICD-11 diagnostic algorithm was 38.8%, 433 patients (76.1%) exceeded the PHQ-9 cut-off for depression in the clinical sample. There was a significant association of PHQ-9 and IDQ caseness (χ^2[1] = 115.58, p<0.001, *V*=0.45 (95% CI 0.37 to 0.54)). 226 (52.2%) of the 433 PHQ-9 cases were found to be IDQ cases.

In the general population sample using the ICD-11 algorithm, 93 participants (5.89%: 95% CI 4.43% to 7.36%) met diagnostic requirements for ICD-11 DE. There was a significant association of PHQ-9 and IDQ caseness (χ^2[1] = 536.10, p<0.001, *V*=0.73 (95% CI 0.70 to 0.76)). Yet, only 57 (25.6%) of the 223 PHQ-9 cases were found to be IDQ cases.

GAD caseness and measure concordance

In the clinical sample, 289 patients (51.0%: 95% CI 46.7% to 54.9%) met screening criteria for ICD-11 GAD. In the general population sample using the ICD-11 algorithm, 59 participants (9.29%: 95% CI 7.49% to 10.9%) met screening criteria for ICD-11 GAD. 361 (63.4%) patients exceeded the GAD-7 cut-off for GAD. There was a significant association of GAD-7 and IAQ caseness (χ^2[1] = 224.99, p<0.001, *V*=0.63 (95% CI 0.55 to 0.71)), with 270 (74.8%) of the 361 GAD-7 cases being found to be also IAQ cases.

In the general population sample, 186 (18.6%) participants exceeded the GAD-7 cut-off for GAD. There was a significant association of GAD-7 and IAQ caseness (χ^2[1] =328.19, p<0.001, *V*=0.57 (95% CI 0.52 to 0.62)), with 82 (44.1%) of the 186 GAD-7 cases being found to be also IAQ cases.

Comorbidity of depression and GAD based on IDQ and IAQ

A total of 183 (32.2%) patients met screening criteria for both disorders, 43 (7.6%) patients met screening criteria only for DE and 106 (18.6%) patients met screening criteria only for GAD. 237 (41.7%) patients met neither screening criteria for DE nor GAD. There was a significant association of caseness for ICD-11 DE and ICD-11 GAD (χ^2[1] = 136.65, p<0.001). Of those with a positive screening for ICD-11 DE, 81.0% also screened positive for ICD-11 GAD. Of those who screened positive for ICD-11 GAD, 63.3% also screened positive for ICD-11 DE. Online supplemental figure S1 depicts overlaps of screening caseness of GAD-7, IAQ, PHQ-9 and IDQ in the clinical sample.

In the general population sample, a total of 43 (4.3%) participants met screening criteria for both disorders, 16 (1.6%) participants met requirements only for ICD-11 DE and 50 (5.0%) participants met requirements only for ICD-11 GAD. 892 (89.1%) participants met neither screening criteria for ICD-11 DE nor ICD-11 GAD. There was a significant association of meeting screening criteria for ICD-11 DE and ICD-11 GAD (χ^2[1] = 300.82, p<0.001). Of those with a positive screening for ICD-11 DE, 72.9% also screened positive for ICD-11 GAD. Of those who screened positive for ICD-11 GAD, 46.2% also screened positive for ICD-11 DE.

DISCUSSION

Overall, the findings largely support the study hypotheses. Both the IDQ and IAQ showed excellent internal consistency (ω=0.96

in both samples), confirming reliability. As hypothesised, the IDQ demonstrated a clear factorial structure: a one-factor model fitted reasonably, supporting its use as a parsimonious screening solution. The three-factor model showed superior fit across CFA and IRT analyses, indicating that the IDQ can also accommodate a multidimensional structure for research and detailed clinical assessment. For the IAQ, a one-factor solution yielded good model fit in both samples, supporting unidimensionality. Screening caseness agreement between the IDQ and PHQ-9, and between the IAQ and GAD-7, was statistically significant in both samples. However, the agreement in caseness was moderate rather than substantial, suggesting that established PHQ-9 and GAD-7 cut-offs may identify a broader group than those meeting ICD-11 screening criteria.

Strengths of the study include using a large, representative general population sample and a well-characterised clinical sample from multiple centres, enhancing generalisability. Limitations involve online recruitment, which may exclude individuals with limited digital access, lack of independent diagnostic validation and restriction to a German population, limiting intercultural generalisability. Furthermore, the clinical sample was predominantly female (78.4%). While this aligns with known gender differences in help-seeking behaviour and service utilisation for internalising disorders in Western countries,[27] it limits the generalisability of our findings. Future studies should aim for more gender-diverse samples to ensure the invariance of the IDQ and IAQ across sexes and genders.

Because the study did not include a clinical interview reference standard, the sensitivity and specificity of the questionnaires for identifying DE and GAD could not be determined. Accordingly, conclusions are limited to screening-defined caseness and construct validity; screening accuracy relative to interview-based diagnoses remains to be established.

Future research should replicate findings using clinician-administered diagnostic interviews to assess criterion validity and longitudinal designs to assess test–retest reliability, screening accuracy and sensitivity to change. Further studies should examine predictive validity for treatment outcomes, minimal clinically important change scores and applicability across diverse cultural and linguistic contexts. Moreover, investigations into the measures' utility in primary care, digital screening environments and stepped-care models could clarify their broader clinical value. Finally, exploring dimensional cut-offs and their correspondence to functional impairment may refine their diagnostic accuracy.

CLINICAL IMPLICATIONS

The results support the reliability and structural validity of the IDQ and IAQ as brief, ICD-11-aligned screening tools for DE and GAD. The measures yield more conservative and diagnostically specific classifications than traditional screeners, ensuring that screening caseness is more closely mapped to the formal symptom requirements and thresholds defined by the ICD-11, reducing overestimation and facilitating appropriate treatment decisions. In contrast to the PHQ-9, which combines symptom frequency with subjective distress, the IDQ and IAQ are explicitly operationalised according to the ICD-11/CDDR. This design supports a more conservative and diagnostically specific classification aligned with current diagnostic standards, although formal diagnostic specificity estimates for the IDQ and IAQ against an interview reference standard are not yet available. The superior three-factor model fit (affective, cognitive-behavioural and neurovegetative symptoms) for the IDQ when compared with the one-factor model mirrors the multidimensional nature of depression described in the recent CDDRs.[14] The substantial overlap between DE and GAD, based on ICD-11 screening criteria, challenges the ICD-11 notion that 'mixed depressive and anxiety disorder' (MDAD) involves only subclinical symptoms.[28] This pattern aligns with transdiagnostic models.[29] Clinically, it suggests that the current ICD-11 definition of MDAD may underidentify patients with syndromal mixed presentations if the category is interpreted as inherently subclinical, potentially contributing to undertreatment. From a measurement perspective, applying categorical thresholds to highly correlated symptom dimensions can yield high rates of dual screening caseness despite shared underlying mechanisms. Future longitudinal research should test whether MDAD represents a distinct clinical phenotype or whether dimensional coding of concurrent DE and GAD with severity and impairment better reflects clinical reality. IDQ and IAQ comparisons with PHQ-9 and GAD-7 indicated lower prevalence rates for DE and GAD, especially in the clinical sample, suggesting that these common cut-offs are overly sensitive and may reflect distress rather than disorder. Researchers and clinicians should therefore avoid interpreting them as proxies for diagnostic caseness.

Clinical implications of the IDQ factor structure

The superiority of the three-factor model of DE over the unidimensional structure we established aligns with the conceptual architecture of the ICD-11 CDDR.[14] The CDDR categorises depressive symptoms into affective, cognitive-behavioural and somatic/neurovegetative clusters. Our findings mirror this conceptualisation and are consistent with the broader psychometric literature (eg, Seemüller *et al*).[30] While high interfactor correlations suggest that these clusters are manifestations of a highly cohesive underlying depressive construct, the statistically superior fit of the three-factor model provides significant practical advantages for clinical practice. Specifically, this structure enables 'symptom profiling'. For example, it allows clinicians to differentiate between primarily somatic, primarily affective and primarily cognitive presentations. Such a level of detail, which is typically lost in a unidimensional sum score, is vital for personalised treatment selection: for instance, a high load on the behavioural factor may indicate a specific need for behavioural activation, whereas a profile dominated by the neurovegetative factor might suggest a primary focus on sleep hygiene. Beyond clinical implications, more nuanced psychometric tools may also enhance future research regarding such clinical questions. To conclude, providing a clear mapping to the ICD-11 clusters as described in the CDDR, the IDQ serves as a brief and economic, yet multifaceted screening tool that may support the selection of targeted interventions early in the treatment process, particularly in primary care settings with little time.

Author affiliations
[1]Institute for Clinical Psychology and Psychotherapy, Department of Psychology, Medical School Hamburg, Hamburg, Germany
[2]Department of Psychotraumatology, Clinic St Irmingard, Prien am Chiemsee, Bavaria, Germany
[3]School of Health & Social Care, Edinburgh Napier University, Edinburgh, UK
[4]Department of Psychology, Division of Clinical Psychology and Psychological Treatment, Ludwig-Maximilians-Universität München, Munich, Germany
[5]Clinic for Psychiatry and Psychotherapy I (Weissenau), Ulm, Germany
[6]Centre for Psychiatry Südwürttemberg, Ravensburg, Germany
[7]Social and Emotional Neuroscience Group, Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, HH, Germany
[8]Department of Psychotherapy and Psychosomatic Medicine, TUD Dresden University of Technology, Dresden, Germany

[9]Department of Clinical and Biological Psychology, Catholic University of Eichstätt-Ingolstadt, Ingolstadt, Bavaria, Germany
[10]Landesklinikum Baden, Baden bei Wien, Austria
[11]Department of Psychology, Maynooth University, Maynooth, Ireland
[12]School of Psychology, University of Ulster, Coleraine, UK

Contributors JS: conceptualisation, methodology, translation, software, investigation, data curation, writing—original draft, writing—review and editing, project administration. LK: conceptualisation, methodology, translation, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualisation, project administration. TK: conceptualisation, methodology, translation, data curation, writing—original draft, writing—review and editing. AS: conceptualisation, methodology, investigation, writing—review and editing. ST: conceptualisation, methodology, translation, investigation, writing—review and editing. SB: conceptualisation, methodology, investigation, writing—review and editing. ES: conceptualisation, methodology, investigation, writing—review and editing. JK conceptualisation, methodology, translation, investigation, writing—review and editing. MK: conceptualisation, methodology, writing—review and editing. PH: conceptualisation, methodology, translation, writing—review and editing. MS: conceptualisation, methodology, translation, software, validation, formal analysis, data curation, writing—original draft, writing—review and editing, visualisation; guarantor; corresponding author. MS is the guarantor.

Funding Robert-Enke-Foundation (no grant number available).

Competing interests All authors have completed the ICMJE uniform disclosure form (available on request from the corresponding author) and declare: JS received funding for the present work from the Robert-Enke-Foundation. LK received payment or honoria for lectures from Deutschsprachige Gesellschaft für Psychotraumatologie, Kliniken Bayern Ost (kbo), AWO Johanna Kirchner Haus, Ludwig-Maximilians-Universität München, Verein zur Förderung der klinischen Verhaltenstherapie, Berufsverband österreichischer Psychologinnen und Psychologen. ST received funding from Bundesamt für Migration und Flüchtlinge. SB received payment or honoria for lectures from Sozialbehörde Hamburg, Elbwerkstätten Hamburg, Ärztekammer Hamburg and Blomenburg Privatklinik; no other relationships or activities that could appear to have influenced the submitted work.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval This study involves human participants. The protocol obtained ethics approval from the ethics committees at Medical School Hamburg (identifier MSH-2023/259). The study was conducted in accordance with the Declaration of Helsinki. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data are available upon reasonable request from the authors.

ORCID iDs
Johanna Schröder https://orcid.org/0000-0002-0751-4720
Leonhard Kratzer https://orcid.org/0000-0002-0174-8497

REFERENCES

1 Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *J Affect Disord* 2017;208:191–7.
2 Wall A, Lee E. What do anxiety scales really measure? an item content analysis of self-report measures of anxiety. *PsyArXiv* [Preprint] 2021.
3 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *J Gen Intern Med* 2001;16:606–13.
4 Spitzer RL, Kroenke K, Williams JBW, *et al*. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;166:1092–7.
5 American Psychiatric Association. Diagnostic and statistical manual of mental disorders, fourth edition, text revision (DSM-IV-TR). Washington, DC American Psychiatric Association; 2000.
6 Löwe B, Decker O, Müller S, *et al*. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Med Care* 2008;46:266–74.
7 Tinghög P, Malm A, Arwidson C, *et al*. Prevalence of mental ill health, traumas and postmigration stress among refugees from Syria resettled in Sweden after 2011: a population-based survey. *BMJ Open* 2017;7:e018899.
8 Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *Int Rev Psychiatry* 2011;23:318–27.
9 Panayiotou M, Razum J, Eisele G, *et al*. Interpretation Issues With the Patient Health Questionnaire Instructions. *JAMA Psychiatry* 2025;85721:1–5.
10 Shevlin M, Hyland P, Butter S, *et al*. The development and initial validation of self-report measures of ICD-11 depressive episode and generalized anxiety disorder: The International Depression Questionnaire (IDQ) and the International Anxiety Questionnaire (IAQ). *J Clin Psychol* 2023;79:854–70.
11 Alpay EH, Redican E, Hyland P, *et al*. Translation and validation of the Turkish forms of the International Depression Questionnaire (IDQ) and the International Anxiety Questionnaire (IAQ). *Acta Psychol (Amst)* 2023;238:103988.
12 Hyland P, Redican E, Karatzias T, *et al*. Assessing the validity and reliability of the International Anxiety Questionnaire and the International Depression Questionnaire in two bereaved national samples. *Clin Psychology and Psychoth* 2024;31.
13 Martsenkovskyi D, Shevlin M, Ben-Ezra M, *et al*. Mental health in Ukraine in 2023. *Eur Psychiatry* 2024;67:e27.
14 World Health Organization. *Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders*. Geneva, Switzerland: World Health Organization, 2024.
15 World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human participants. *JAMA* 2024.
16 Wild D, Grove A, Martin M, *et al*. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health* 2005;8:94–104.
17 Kliem S, Sachser C, Lohmann A, *et al*. Psychometric evaluation and community norms of the PHQ-9, based on a representative German sample. *Front Psychiatry* 2024;15:1483782.
18 Kliem S, Sachser C, Lohmann A, *et al*. Psychometric evaluation and community norms of the GAD-7, based on a representative German sample. *Front Psychol* 2025;16:1526181.
19 Plummer F, Manea L, Trepel D, *et al*. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *Gen Hosp Psychiatry* 2016;39:24–31.
20 Lamping DL, Schroter S, Marquis P, *et al*. The Community-Acquired Pneumonia Symptom Questionnaire. *Chest* 2002;122:920–9.
21 Baker FB. *The Basics of Item Response Theory*. 2nd edn. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation, 2001.
22 Schermelleh-Engel K, MoosbruggerH, Müller H. Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods Psychol Res* 2003;8:23–74.
23 McDonald RP. *Test Theory: A Unified Approach*. Mahwah, NJ: Erlbaum, 1999.
24 Levis B, Benedetti A, Thombs BD, *et al*. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476.
25 Ben-Shachar M, Lüdecke D, Makowski D. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *JOSS* 2020;5:2815.
26 Chen H. VennDiagram: Generate high-resolution Venn and Euler plots. 2014.
27 Whitley R. Why do men have low rates of formal mental health service utilization? an analysis of social and systemic barriers to care and discussion of promising male-friendly practices. In: *Men's Issues and Men's Mental Health*. Cham: Springer International Publishing, 2021: 127–49.
28 Shevlin M, Hyland P, Nolan E, *et al*. ICD-11 'mixed depressive and anxiety disorder' is clinical rather than sub-clinical and more common than anxiety and depression in the general population. *British J Clinic Psychol* 2022;61:18–36.
29 Kotov R, Krueger RF, Watson D, *et al*. The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J Abnorm Psychol* 2017;126:454–77.
30 Seemüller F, Schennach R, Musil R, *et al*. A factor analytic comparison of three commonly used depression scales (HAMD, MADRS, BDI) in a large sample of depressed inpatients. *BMC Psychiatry* 2023;23:548.