Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agreement of GPT-5.5 scenario scores with the PHQ-9 and GAD-7 in university students: a cross-sectional study

Mahmoud Sayed Saad mohammed

mahmoud.sayed.maghrappy@gmail.com

Cairo University

Hager Fathy Mohamed Ali Abo-El-Enin

Damietta University

Omnia Salah El-Shahate Aref

Delta University for Science and Technology

Khaled Elbarbary

Mansoura University

Kholoud ELHADIDI

Capital University (formerly Helwan University)

Nora Eldmrdash

Delta University for Science and Technology

Rana Hassan Moustafa

Alexandria University

Rana Mohamed Attallah

Delta University for Science and Technology

Layan Eltayieb

The Ohio State University

Research Article

Keywords: GPT-5.5, depression, anxiety, university students, screening, measurement agreement

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11666457/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. We examined whether GPT-5.5 scores derived from four open-ended responses agreed with self-reported depression and anxiety symptoms. In this cross-sectional study, 359 university students completed the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 between 1 April and 1 June 2026. A self-report score of 10 or higher defined a positive screen.

Results. Paired scores were available for 355 participants for depression and 353 for anxiety. Spearman correlations were 0.317 and 0.323, respectively, and absolute-agreement ICCs were 0.169 and 0.176. Self-report scores exceeded model scores by a mean of 6.07 points for depression and 5.04 for anxiety. Areas under the curve were 0.662 (95% CI 0.604–0.720) and 0.628 (95% CI 0.570–0.686), respectively. Of 139 students who endorsed PHQ-9 item 9, 75 received a model score of zero for that item.

Conclusions. The scenario scores underestimated symptom severity and showed poor agreement with both reference instruments. These findings do not support their use for stand-alone screening in this population.

Background

University students commonly experience symptoms of depression and anxiety. An umbrella review estimated prevalences of 35.4% and 40.2%, respectively [1]. The PHQ-9 and GAD-7 provide brief standardized measures of these symptoms [2, 3]. Their performance depends on the population and language of administration [4, 5, 6, 7].

Large language models can generate questions and interpret free-text responses [8]. Interest in these tools among university students [9, 10] and their use for mental health advice among young people [11] motivate a practical question. Can a model infer symptom frequency from a student's description of everyday situations as reliably as the student reports it on a standardized questionnaire?

Liu et al. [12] found moderate correlations between ChatGPT-derived scores and the PHQ-9 and GAD-7 among college students. Correlation alone cannot establish agreement. A model may rank students similarly while assigning scores low enough to miss students who meet a screening threshold.

We evaluated GPT-5.5 score agreement and screening performance against the PHQ-9 and GAD-7 using responses to four scenarios about university life. We also examined severity classification, agreement on PHQ-9 item 9, and associations with response length and survey version.

Methods

Study design and setting

This was a cross-sectional study evaluating the agreement and screening performance of a ChatGPT-adapted scenario questionnaire against the validated Patient Health Questionnaire-9 (PHQ-9) and

Generalized Anxiety Disorder-7 (GAD-7) among university students. Data were collected between 1 April 2026 and 1 June 2026. All data were collected anonymously through an electronic self-administered questionnaire delivered via Google Forms. The study was approved by the Research Ethics Committee of Mansoura University (approval number 36265PR126/6/26) and was conducted in accordance with the Declaration of Helsinki. Electronic informed consent was obtained from each participant on the first page of the questionnaire before any item was displayed. Reporting was organized with reference to the Standards for Reporting Diagnostic Accuracy Studies (STARD) 2015 statement [13].

Participants

A convenience sample of students currently enrolled at a university was recruited. Eligibility required current university enrolment and consent to participate. No exclusion was applied on the basis of age, nationality, or field of study. Participants completed the survey before the model-derived scores were generated.

Recruitment used an Arabic-language version (A) and an English-language version (B) of the survey, with identical content. Participants could answer the free-text questions in either language.

Sample size

The sample size was determined by the number of responses obtained during the recruitment window rather than by an a priori calculation. The final dataset contained 359 participants; paired complete-case analyses included 355 participants for depression and 353 for anxiety. Precision is reported through confidence intervals for the principal estimates.

ChatGPT-adapted questionnaire

The scenario questionnaire comprised four open-ended questions. ChatGPT was prompted to generate questions matching the items and scoring criteria of the PHQ-9 and GAD-7, framed within daily-life situations relevant to university students and covering academic study, social interaction, and future career concerns. The resulting instrument comprised four open-ended scenario questions administered in Arabic in version A and in English in version B. To distinguish them from the self-report instruments, all model-derived scores carry the prefix "GPT-". The model returned a score of 0 to 3 for each of the nine PHQ-9 items and each of the seven GAD-7 items, which were summed to give GPT-PHQ (range 0 to 27) and GPT-GAD (range 0 to 21) totals. Sixteen item scores were therefore derived from four free-text responses.

After students completed the survey, their free-text responses were entered manually into the ChatGPT web interface and scored with GPT-5.5. The scoring dates, number of raters, and verbatim scoring prompt were not recorded. ChatGPT also assisted with manuscript editing and formatting.

Reference standard

The PHQ-9 is a nine-item self-report instrument for depressive symptoms. Respondents rate the frequency of each symptom over the preceding two weeks on a four-point scale (0 = not at all, 1 = several days, 2 = more than half the days, 3 = nearly every day), giving a total of 0 to 27, with higher scores indicating greater severity [2]. The GAD-7 uses the same response scale across seven items, giving a total of 0 to 21 [3]. Both instruments have been validated in Arabic-speaking populations, with reported Cronbach's alpha coefficients of 0.857 for the PHQ-9 and 0.763 for the GAD-7 in a sample of Saudi university students [7].

A total score of 10 or above on each self-report instrument defined a positive screen. No clinical interview was performed. The reference standard was therefore a validated screening instrument rather than a diagnosis; all screening-performance estimates describe classification against the corresponding self-report threshold.

For categorical analyses, PHQ-9 totals were grouped as 0 to 4 minimal, 5 to 9 mild, 10 to 14 moderate, 15 to 19 moderately severe and 20 to 27 severe [2], and GAD-7 totals as 0 to 4 minimal, 5 to 9 mild, 10 to 14 moderate and 15 to 21 severe. Both self-report instruments and the scenario questionnaire were completed in the same sitting; model scoring was performed subsequently.

Data handling

Records with missing item-level responses, or with values outside the permitted range of 0 to 3, were excluded from analyses requiring a complete total score. Missing values were not imputed. Analyses were conducted on complete cases, and the number of participants contributing to each estimate is reported with that estimate. Item-level analyses used pairwise deletion, so the denominator varies between items.

Before analysis, we screened records for response patterns indicating a data-handling or transcription problem. Three specified indicators were applied, namely a model score vector identical to the self-report vector of the same participant, a model score vector identical to the self-report vector of an adjacent record, and a model score of zero on all 16 items. Records meeting any indicator were retained in the primary analysis and removed in a specified sensitivity analysis. Straight-line responding on the self-report instruments was also quantified.

Statistical analysis

Analyses were performed in Python 3.11 (pandas, SciPy, pingouin, scikit-learn) with a two-sided $\alpha = 0.05$. Due to non-normality (Shapiro-Wilk test), totals were summarized as medians with interquartile ranges, and categorical variables as frequencies with 95% Wilson confidence intervals. Score associations and internal consistency were evaluated using Spearman rank correlations (Fisher z CIs) and Cronbach's alpha (2000 bootstrap CIs). Agreement between total scores was quantified using two-way, single-

measure absolute and consistency intraclass correlation coefficients (ICC[A,1] and ICC[C,1]), while categorical severity agreement utilized linearly and quadratically weighted Cohen's kappa. Bland-Altman analysis assessed measurement agreement, utilizing regression-based limits if proportional bias was detected, alongside ordinary least squares regression to estimate limits following optimal linear recalibration. Diagnostic performance against self-report screens was evaluated via ROC-AUC (DeLong CIs) and Youden's index, with optimism corrected via 2000 bootstraps; sensitivity, specificity, PPV, and NPV were reported at both optimal and specified thresholds. Pre-specified secondary analyses examined PHQ-9 item 9 (thoughts of death or self-harm) individually, tested convergent and discriminant validity by comparing monotrait versus heterotrait correlations (Mann-Whitney U test), and assessed the independent effect of response character count via partial Spearman correlation. Finally, sensitivity analyses re-evaluated primary estimates after applying data-quality exclusions and stratifying by survey version, with no multiple comparison corrections applied to exploratory item-level tests.

Results

Participants

A total of 359 students completed the survey. Most were aged 18 to 23 years (337, 93.9%, 95% CI 90.9 to 95.9), resident in Egypt (349, 97.2%, 95% CI 94.9 to 98.7), and of Egyptian nationality (348, 96.9%, 95% CI 94.6 to 98.5). Participants reported attending 41 institutions; Damietta University (76, 21.2%), Zagazig University (60, 16.7%), Mansoura University (49, 13.6%), Delta University for Science and Technology (38, 10.6%), and Horus University in Egypt (17, 4.7%) were the five most frequently reported. By sector, 265 participants (73.8%, 95% CI 69.0 to 78.1) were enrolled at a public university and 84 (23.4%, 95% CI 19.3 to 28.0) at a private university. The distribution across institutions is given in Additional file 1 (Table S1). Participant characteristics are summarised in Table 1.

Table 1
Participant characteristics and total-score distributions (N = 359).

Variable	Category	n	%	95% CI
Age group				
	18 to 23 years	337	93.9	90.9–95.9
	Over 23 years	18	5.0	3.2–7.8
	Under 18 years	4	1.1	0.4–2.8
Country of residence				
	Egypt	349	97.2	94.9–98.7
	Other country	8	2.2	1.1–4.3
	Multiple	2	0.6	0.1–2.0
Nationality				
	Egyptian	348	96.9	94.6–98.5
	Other	11	3.1	1.7–5.4
Institution				
	Damietta University	76	21.2	17.3–25.7
	Zagazig University	60	16.7	13.2–20.9
	Mansoura University	49	13.6	10.5–17.6
	Delta University for Science and Technology	38	10.6	7.8–14.2
	Horus University in Egypt	17	4.7	3.0–7.5
Institution type				
	Public university	265	73.8	69.0–78.1
	Private university	84	23.4	19.3–28.0
	Private higher institute	5	1.4	0.6–3.2
	National (non-profit) university	4	1.1	0.4–2.8
	Regional academy (Arab League affiliated)	1	0.3	0.0–1.6
Survey version				
	Version A	214	59.6	54.5–64.6

Variable	Category	n	%	95% CI
Age group				
	Version B	145	40.4	35.4–45.5
Total score	**n**	**Median (IQR)**	**Range**	**Shapiro–Wilk W, P**
PHQ-9	358	11 (8–15)	0–27	0.970, < 0.001
GPT-PHQ	356	6 (2–9)	0–25	0.924, < 0.001
GAD-7	359	9 (6–14)	0–21	0.962, < 0.001
GPT-GAD	353	4 (2–7)	0–21	0.894, < 0.001
Positive screen, PHQ-9 ≥ 10	358	232 (64.8%)	95% CI 59.7–69.6	NA
Positive screen, GAD-7 ≥ 10	359	162 (45.1%)	95% CI 40.1–50.3	NA

[Place Table 1 here.]

Data completeness and quality

Complete total scores were available for 358 participants for the PHQ-9, 359 for the GAD-7, 356 for GPT-PHQ, and 353 for GPT-GAD. Paired complete-case analyses included 355 participants for depression comparisons and 353 for anxiety comparisons. Exclusions comprised one record with a missing PHQ-9 item, two records with missing GPT-PHQ items, one record with a GPT-PHQ item value outside the permitted range, and six records with missing GPT-GAD items.

The specified data-quality indicators identified 78 records (21.7%). Twenty-four records carried a model score vector identical to the participant's own self-report vector, 21 carried a vector identical to that of an adjacent record, and 39 carried a model score of zero on all 16 items, with overlap between categories. Straight-line responding was present on the GAD-7 in 50 records (13.9%) and on the PHQ-9 in 8 records (2.2%). All 78 records were retained in the primary analysis and removed in sensitivity analyses (Table S2).

Score distributions

All four total-score distributions departed significantly from normality (Shapiro–Wilk, all $P < 0.001$; Table 1). The median PHQ-9 total was 11 (IQR 8 to 15) against a median GPT-PHQ of 6 (IQR 2 to 9), and the median GAD-7 total was 9 (IQR 6 to 14) against a median GPT-GAD of 4 (IQR 2 to 7). A positive screen was present in 232 of 358 participants (64.8%, 95% CI 59.7 to 69.6) on the PHQ-9 and in 162 of 359 (45.1%, 95% CI 40.1 to 50.3) on the GAD-7.

Correlation and internal consistency

GPT-PHQ totals correlated positively with PHQ-9 totals (Spearman's ρ = 0.317, 95% CI 0.220 to 0.407, *n* = 355, *P* < 0.001), as did GPT-GAD with GAD-7 (ρ = 0.323, 95% CI 0.226 to 0.414, *n* = 353, *P* < 0.001).

Cronbach's alpha was 0.863 (95% CI 0.834 to 0.888) for GPT-PHQ, 0.811 (95% CI 0.775 to 0.843) for the PHQ-9, 0.862 (95% CI 0.825 to 0.890) for GPT-GAD and 0.893 (95% CI 0.873 to 0.910) for the GAD-7. Mean inter-item correlation was 0.413 for GPT-PHQ (range 0.181 to 0.794) against 0.317 for the PHQ-9 (range 0.148 to 0.438), and 0.448 for GPT-GAD (range 0.302 to 0.751) against 0.536 for the GAD-7 (range 0.389 to 0.714). Across the nine PHQ items, the model assigned two or fewer distinct values in 79.6% of records against 21.8% for self-report, and across the seven GAD items in 81.6% against 55.4%. Mean within-participant standard deviation across items was 0.499 for the model against 0.828 for self-report on the PHQ items, and 0.489 against 0.609 on the GAD items.

Agreement between total scores

Absolute agreement was poor. The two-way single-measures absolute-agreement coefficient was 0.169 (95% CI − 0.040 to 0.360) for depression totals and 0.176 (95% CI − 0.020 to 0.350) for anxiety totals. The corresponding consistency coefficients were 0.290 (95% CI 0.190 to 0.380) and 0.274 (95% CI 0.180 to 0.370). Coefficients are reported in Table 2.

Table 2
Agreement, internal consistency and screening performance of model scores.

Measure	Depression GPT-PHQ vs PHQ-9	Anxiety GPT-GAD vs GAD-7
Participants analysed, n	355	353
Association and internal consistency		
Spearman ρ	0.317 (0.220–0.407)	0.323 (0.226–0.414)
Cronbach α, model-derived scale	0.863 (0.834–0.888)	0.862 (0.825–0.890)
Cronbach α, self-report scale	0.811 (0.775–0.843)	0.893 (0.873–0.910)
Agreement between total scores		
ICC(A,1), absolute agreement	0.169 (− 0.040–0.360)	0.176 (− 0.020–0.350)
ICC(C,1), consistency	0.290 (0.190–0.380)	0.274 (0.180–0.370)
Mean bias, self-report minus model	+ 6.07 (+ 5.44 to + 6.70)	+ 5.04 (+ 4.44 to + 5.64)
Proportional bias slope	+ 0.279 (+ 0.127–+0.431)	+ 0.477 (+ 0.327–+0.627)
Agreement on severity category		
Quadratically weighted κ	0.133 (0.075–0.191)	0.175 (0.110–0.244)
Exact category agreement, %	26.2 (21.9–31.0)	32.3 (27.6–37.3)
Model assigned a lower category, %	69.3 (64.3–73.9)	61.5 (56.3–66.4)
Screening performance against a self-report total ≥ 10		
AUC	0.662 (0.604–0.720)	0.628 (0.570–0.686)
Threshold maximising Youden's J	≥ 7	≥ 5
Sensitivity at that threshold, %	53.5 (47.0–59.8)	59.6 (51.9–66.9)
Specificity at that threshold, %	72.0 (63.6–79.1)	59.9 (52.8–66.6)
PPV at that threshold, %	77.8 (70.8–83.6)	55.5 (48.0–62.7)
NPV at that threshold, %	45.7 (38.9–52.7)	63.9 (56.6–70.5)
Sensitivity at the pre-specified ≥ 10, %	19.6 (15.0–25.2)	15.5 (10.7–21.9)
Specificity at the pre-specified ≥ 10, %	95.2 (89.9–97.8)	93.2 (88.8–96.0)
Agreement on the suicidal ideation item (PHQ-9 item 9, n = 358)		

Values in parentheses are 95% confidence intervals. ICC, intraclass correlation coefficient; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; NA, not applicable.

Measure	Depression GPT-PHQ vs PHQ-9	Anxiety GPT-GAD vs GAD-7
Endorsed any ideation, n (%)	139 (38.8, 33.9–44.0)	NA
Of those, scored 0 by the model, n (%)	75 (54.0, 45.7–62.0)	NA
Endorsed ideation > half the days, n (%)	62 (17.3, 13.8–21.6)	NA
Of those, scored 0 by the model, n (%)	29 (46.8, 34.9–59.0)	NA

Values in parentheses are 95% confidence intervals. ICC, intraclass correlation coefficient; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; NA, not applicable.

[Place Table 2 here.]

Item-level agreement

Item-level correlations between model-derived and self-report items were weak, ranging from ρ = 0.069 to ρ = 0.299 for depression items and ρ = 0.072 to ρ = 0.283 for anxiety items (Figure S1, Table S3). The strongest item-level agreement was observed for PHQ-9 item 9 and GAD-7 item 7.

The mean monotrait correlation, being the same symptom measured by both methods, was 0.194 for depression items against a mean heterotrait correlation of 0.132 (difference 0.062, Mann–Whitney *P* = 0.007). For anxiety items, the corresponding values were 0.189 and 0.170 (difference 0.019, Mann–Whitney *P* = 0.644). The matching self-report item was the strongest correlate of its model-derived counterpart in 4 of 9 depression items and 1 of 7 anxiety items. These item-level analyses are exploratory and uncorrected for multiple comparisons.

Severity category agreement

Exact agreement on severity category was 26.2% (95% CI 21.9 to 31.0) for depression and 32.3% (95% CI 27.6 to 37.3) for anxiety. Quadratically weighted kappa was 0.133 (95% CI 0.075 to 0.191) and 0.175 (95% CI 0.110 to 0.244), respectively; linearly weighted kappa was 0.097 (95% CI 0.059 to 0.138) and 0.131 (95% CI 0.083 to 0.179). The model assigned a lower severity category than self-report in 69.3% (95% CI 64.3 to 73.9) of participants for depression and 61.5% (95% CI 56.3 to 66.4) for anxiety, and a higher category in 4.5% (95% CI 2.8 to 7.2) and 6.2% (95% CI 4.2 to 9.3). Twenty-eight participants in the two highest PHQ-9 categories were assigned the minimal category by the model. Cross-classifications are shown in Fig. 1.

Agreement on PHQ-9 item 9

PHQ-9 item 9 captures thoughts of being better off dead or self-harm and does not independently establish suicidal intent. Among 358 participants with complete data on PHQ-9 item 9, 139 (38.8%, 95% CI 33.9 to 44.0) endorsed ideation at any frequency and 62 (17.3%, 95% CI 13.8 to 21.6) endorsed it on

more than half the days. The model assigned a score of zero on this item to 75 of the 139 participants endorsing any ideation (54.0%, 95% CI 45.7 to 62.0) and to 29 of the 62 endorsing ideation on more than half the days (46.8%, 95% CI 34.9 to 59.0). Sensitivity of a model item 9 score of 1 or above for detecting any endorsed ideation was 46.0% (95% CI 38.0 to 54.3), with specificity 80.4% (95% CI 74.6 to 85.1). Results are shown in Fig. 2.

Screening performance

Against the self-report positive screen, GPT-PHQ achieved an AUC of 0.662 (DeLong 95% CI 0.604 to 0.720) and GPT-GAD an AUC of 0.628 (DeLong 95% CI 0.570 to 0.686; Fig. 3A and 3B). Youden's index identified an optimal GPT-PHQ threshold of 7 or above, with sensitivity 53.5% (95% CI 47.0 to 59.8) and specificity 72.0% (95% CI 63.6 to 79.1), and an optimal GPT-GAD threshold of 5 or above, with sensitivity 59.6% (95% CI 51.9 to 66.9) and specificity 59.9% (95% CI 52.8 to 66.6).

At these thresholds, positive predictive value was 77.8% (95% CI 70.8 to 83.6) and negative predictive value 45.7% (95% CI 38.9 to 52.7) for GPT-PHQ, and 55.5% (95% CI 48.0 to 62.7) and 63.9% (95% CI 56.6 to 70.5) for GPT-GAD. At the specified threshold of 10, GPT-PHQ sensitivity was 19.6% (95% CI 15.0 to 25.2) with specificity 95.2% (95% CI 89.9 to 97.8), and GPT-GAD sensitivity was 15.5% (95% CI 10.7 to 21.9) with specificity 93.2% (95% CI 88.8 to 96.0). Performance at selected thresholds is shown in Fig. 3C and 3D and in Table S4.

Bootstrap estimation gave an optimism of 0.032 in Youden's index for GPT-PHQ, reducing the apparent value of 0.255 to 0.223, and 0.027 for GPT-GAD, reducing 0.195 to 0.168. The threshold selected by Youden's index varied across bootstrap replications from 2 to 10 for GPT-PHQ (median 7, IQR 6 to 7) and from 1 to 12 for GPT-GAD (median 4, IQR 3 to 5).

Bland–Altman agreement

Mean bias, calculated as the self-report score minus the model score, was + 6.07 points (95% CI + 5.44 to + 6.70) for depression with a standard deviation of the differences of 6.03, and + 5.04 points (95% CI + 4.44 to + 5.64) for anxiety with a standard deviation of 5.73. Conventional fixed 95% limits of agreement were − 5.76 to + 17.90 points for depression and − 6.18 to + 16.26 points for anxiety (Table S5).

The difference was not constant across the measurement range. Regressing the difference on the mean gave a slope of + 0.279 (95% CI + 0.127 to + 0.431, $P < 0.001$) for depression and + 0.477 (95% CI + 0.327 to + 0.627, $P < 0.001$) for anxiety. The differences also departed from normality (Shapiro–Wilk $P < 0.001$ and $P = 0.004$). We therefore computed regression-based limits of agreement. For depression, these spanned − 2.8 to + 12.7 points at a mean score of 5 and − 9.0 to + 24.5 points at a mean score of 15; for anxiety,y they spanned − 3.9 to + 11.8 and − 9.3 to + 26.7 at the same mean scores (Fig. 1A and 1B, and Table S5).

Linear regression of the self-report total on the model total gave PHQ-9 = 9.93 + 0.35 × GPT-PHQ (R^2 = 0.087) and GAD-7 = 7.94 + 0.39 × GPT-GAD (R^2 = 0.083). After this recalibration removed the bias, the residual 95% limits of agreement spanned 20.6 points on the 0 to 27 depression scale and 20.4 points on the 0 to 21 anxiety scale.

Response length and model score

Total free-text length across the four scenario responses correlated with the model score (ρ = 0.223, P < 0.001 for depression; ρ = 0.330, P < 0.001 for anxiety) more strongly than with the self-report score (ρ = 0.121, P = 0.023; ρ = 0.113, P = 0.034). After adjustment for the self-report total, the partial correlation between response length and the model score remained 0.197 (95% CI 0.09 to 0.29, P < 0.001) for depression and 0.312 (95% CI 0.21 to 0.40, P < 0.001) for anxiety. Mean GPT-PHQ rose from 4.60 in the shortest response quartile to 6.92 in the longest, and mean GPT-GAD from 3.37 to 6.28, while mean self-report totals varied by less than 1.2 points across quartiles (Figure S2).

Sensitivity analyses

Estimates were stable across all exclusion scenarios (Table S2). Excluding all 78 records meeting a data-quality indicator, Spearman's ρ was 0.321 for depression and 0.293 for anxiety, AUC was 0.662 and 0.627, and mean bias was + 6.27 and + 5.36 points, none differing materially from the primary analysis. Analysed separately, the Arabic-language (A) and English-language (B) versions gave ρ of 0.312 and 0.332 for depression and AUC of 0.655 and 0.665. Median self-report totals did not differ between versions (Mann–Whitney P = 0.95 for PHQ-9 and P = 0.35 for GAD-7), whereas median model scores did (P = 0.005 for GPT-PHQ and P = 0.048 for GPT-GAD).

Discussion

This study evaluated the validity and screening performance of a scenario questionnaire scored with GPT-5.5 for depression and anxiety symptoms among university students, using the validated PHQ-9 and GAD-7 as reference instruments. The results showed weak positive correlations between the model-derived scores and the corresponding self-report measures (ρ = 0.317 for depression and ρ = 0.323 for anxiety). Although higher model-derived scores were generally associated with higher symptom scores on the validated instruments, the strength of these relationships was limited. The findings therefore suggest that the questionnaire captured some aspects of participants' reported symptoms but did not measure them with sufficient agreement to be considered equivalent to the established instruments.

The findings also differed from those reported by Liu et al. [12], who found stronger correlations between ChatGPT-generated scores and the PHQ-9 and GAD-7 among college students, with coefficients of 0.63 and 0.68, respectively. Several factors may help explain this difference, including variations in the study population, questionnaire design, language and cultural context, and scoring procedures. In the present study, participants described their experiences through four open-ended, scenario-based questions, and their responses could vary considerably in length and detail. This format may have introduced additional variability in how symptoms were expressed and interpreted. However, the present findings do not

establish which of these factors contributed to the weaker associations. The difference between the two studies highlights the importance of evaluating model-based assessment tools independently across different settings rather than assuming that performance in one population will be reproduced in another.

An important finding was the difference between internal consistency and agreement with the reference instruments. The model-derived depression and anxiety scales showed good internal consistency, with Cronbach's alpha values of 0.863 and 0.862, respectively. However, the absolute-agreement ICCs were only 0.169 for depression and 0.176 for anxiety, indicating poor agreement between the two approaches. This distinction is important because internal consistency reflects how closely items within a scale relate to one another; it does not establish that the scale measures symptom severity accurately or produces scores comparable to those of a validated instrument. In this study, the model-generated items appeared to produce internally consistent scores, but those scores did not reliably correspond to the participants' PHQ-9 and GAD-7 totals. The weak item-level correlations further support this interpretation, as the relationships between corresponding items were generally limited. Taken together, these findings suggest that good internal consistency alone is insufficient to establish the validity of a model-generated screening questionnaire.

The differences between the two scoring approaches were also apparent in the severity-category and Bland–Altman analyses. Exact agreement in severity classification was 26.2% for depression and 32.3% for anxiety. The model assigned a lower severity category than the corresponding self-report instrument in 69.3% of depression cases and 61.5% of anxiety cases. Similarly, Bland–Altman analysis showed that the validated questionnaire scores were higher on average than the model-derived scores, with mean differences of 6.07 points for depression and 5.04 points for anxiety. The limits of agreement were wide, indicating considerable variation in the difference between the two methods at the individual level. In addition, the presence of proportional bias suggested that this difference was not constant across the score range. These results indicate that the model tended to underestimate symptom severity, although the size of the discrepancy varied between participants. Consequently, the model-derived scores should not be treated as interchangeable with PHQ-9 or GAD-7 scores, even when the overall correlation is positive.

The ROC analysis provided further evidence of limited screening performance. The area under the curve was 0.662 for depression and 0.628 for anxiety, indicating only modest ability to distinguish participants who met the reference questionnaire thresholds from those who did not. At the cutoffs selected using Youden's index, sensitivity was 53.5% for depression and 59.6% for anxiety, while specificity was 72.0% and 59.9%, respectively. These values suggest that the model would fail to identify a considerable proportion of participants who screened positive on the validated instruments. When the prespecified threshold of 10 was applied to the model-derived scores, sensitivity decreased further, reaching 19.6% for depression and 15.5% for anxiety, despite higher specificity. Although the optimal cutoffs performed better in terms of sensitivity, they were selected and evaluated within the same dataset and therefore

require validation in an independent sample. Overall, the results do not support using the questionnaire as a stand-alone screening tool.

Another finding was the association between response length and model-derived scores. The total length of participants' free-text responses was positively correlated with the model score for both depression ($\rho = 0.223$) and anxiety ($\rho = 0.330$). These associations remained after adjustment for the corresponding self-report scores, with partial correlations of 0.197 and 0.312, respectively. This raises the possibility that the amount of text provided influenced scoring, in addition to the symptoms described. Participants who elaborated more on their experiences may have provided more information for the model to interpret, whereas brief responses may not have conveyed the same level of detail. However, the observed associations do not establish that response length directly caused higher scores. They do suggest that future studies should examine whether standardized prompts, clearer response requirements, and more explicit scoring criteria can reduce variability related to response length.

The findings for PHQ-9 item 9 are particularly relevant to the potential use of this approach in mental health screening. Among the 358 participants with complete data for this item, 139 endorsed it at some frequency, yet the model assigned a score of zero to 75 of these participants. Furthermore, 29 of the 62 participants who endorsed the item on more than half of the days received a model score of zero. The sensitivity of a model score of at least 1 for identifying any endorsement of item 9 was 46.0%. Although endorsement of this item does not independently establish suicidal intent, the discrepancy indicates that the model did not consistently identify responses related to thoughts of being better off dead or self-harm. In a screening context, such missed responses are an important concern and reinforce the need for direct, validated assessment of suicide-related symptoms rather than reliance on model-generated scores.

The analyses by survey version also require cautious interpretation. The model-derived scores showed some differences between the Arabic and English survey versions, whereas the self-report PHQ-9 and GAD-7 scores did not differ significantly between groups. However, participants were allowed to answer the free-text questions in either language, meaning that survey version did not necessarily represent the language used in every response. These findings therefore cannot establish that one language version was more culturally sensitive or accurate than the other. Instead, they indicate that the model-derived results may vary across survey versions and support the need for more controlled evaluations of language-related effects, using standardized procedures and clearly documented response languages.

Future studies should evaluate model-generated questionnaires in larger and more diverse populations and use independent validation samples. The model version, dates of scoring, prompts, and scoring procedures should be documented in detail and kept consistent throughout data collection and analysis. Comparisons with structured clinical interviews, alongside validated screening instruments, would help clarify whether these tools can contribute to clinical assessment rather than merely reproduce questionnaire scores. Further research should also investigate the effects of response length and language, and assess the model's ability to recognize safety-related symptoms reliably. Until such

evidence is available, model-generated scores should be interpreted cautiously and should not replace validated screening instruments or clinical evaluation.

Limitations

This cross-sectional convenience sample consisted predominantly of students aged 18 to 23 years and cannot establish generalizability to clinical populations or other settings. The reference instruments were self-report screening scales, not structured diagnostic interviews. The recruitment channel, procedures for consent among participants aged under 18, and timing of ethics approval relative to recruitment were not documented in the materials available for this report. Missing or invalid scores were handled using complete-case analyses, and 78 records met at least one data-quality indicator. Sensitivity analyses were broadly consistent with the primary analysis but do not exclude transcription or scoring errors. Thresholds were selected and evaluated within the same dataset; bootstrap correction does not replace external validation. Exploratory item-level comparisons were not adjusted for multiple testing. Scoring dates, number of raters, and verbatim scoring prompt were unavailable. Blinding and session-level scoring procedures could not be verified, limiting reproducibility. The analysis code and participant-level dataset were unavailable for independent verification of the reported estimates.

Conclusions

GPT-5.5 scores showed weak correlations and poor absolute agreement with PHQ-9 and GAD-7 totals. The model generally assigned lower severity and missed more than half of self-reported endorsements of PHQ-9 item 9. In this sample, the four-scenario questionnaire did not support independent screening for depression or anxiety.

Abbreviations

AUC, area under the receiver operating characteristic curve. CI, confidence interval. GAD-7, Generalized Anxiety Disorder-7. GPT-GAD and GPT-PHQ, model-derived anxiety and depression scores. ICC, intraclass correlation coefficient. IQR, interquartile range. NPV, negative predictive value. PHQ-9, Patient Health Questionnaire-9. PPV, positive predictive value. ROC, receiver operating characteristic. SD, standard deviation. STARD, Standards for Reporting Diagnostic Accuracy Studies.

Declarations

Ethics approval and consent to participate

The Research Ethics Committee of Mansoura University approved the study (reference 36265PR126/6/26). Participants provided electronic informed consent.

Consent for publication

Not applicable.

Funding

No funding

Author Contribution

MSSM, HFMA, OSEA, KE, KEH, NE, RHM, RMA and LE collected data and contributed to writing the manuscript. MSSM conceived the study and is responsible for submission to the journal. MMK and NMEE coordinated the ethics application and obtained approval from the research ethics committee. The original nine authors reviewed and approved the final manuscript and agreed to submission to Middle East Current Psychiatry.

Acknowledgement

Mostafa Mamdoh kamelLecturer of psychiatry Faculty of medicine - Tanta university, Ethical approvalNada Mahmoud Ebrahim El MeyaFaculty of Medicine , Tanta University, Ethical approval

Availability of data and materials

[Data can be shared, by first Author]

Competing interests

No Competing Interests

References

1. Paiva U, Cortese S, Flor M et al (2025) Prevalence of mental disorder symptoms among university students: an umbrella review. Neurosci Biobehav Rev 175:106244. https://doi.org/10.1016/j.neubiorev.2025.106244
2. Kroenke K, Spitzer RL, Williams JBW (2001) The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 16:606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x
3. Spitzer RL, Kroenke K, Williams JBW, Löwe B (2006) A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 166:1092–1097. https://doi.org/10.1001/archinte.166.10.1092
4. Fonseca-Pedrero E, Diez-Gomez A, Perez-Albeniz A, Al-Halabi S, Lucas-Molina B, Debbane M (2023) Youth screening depression: Validation of the Patient Health Questionnaire-9 (PHQ-9) in a representative sample of adolescents. Psychiat Res 328:115486. https://doi.org/10.1016/j.psychres.2023.115486

5. Phelan E, Williams B, Meeker K, Bonn K, Frederick J, Logerfo J, Snowden M (2010) A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. BMC Fam Pract 11:63. https://doi.org/10.1186/1471-2296-11-63
6. Woldetensay YK, Belachew T, Tesfaye M, Spielman K, Biesalski HK, Kantelhardt EJ, Scherbaum V (2018) Validation of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression in pregnant women: Afaan Oromo version. PLoS ONE 13:e0191782. https://doi.org/10.1371/journal.pone.0191782
7. AlHadi AN, AlAteeq DA, Al-Sharif E et al (2017) An Arabic translation, reliability, and validation of Patient Health Questionnaire in a Saudi sample. Ann Gen Psychiatry 16:32. https://doi.org/10.1186/s12991-017-0155-1
8. Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, prospects, and ethical considerations. Front Artif Intell 6:1169595. https://doi.org/10.3389/frai.2023.1169595
9. Abdaljaleel M, Barakat M, Alsanafi M et al (2024) A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. Sci Rep 14(1):1983. https://doi.org/10.1038/s41598-024-52549-8
10. Ravšelj D, Keržič D, Tomaževič N et al (2025) Higher education students' perceptions of ChatGPT: A global study of early reactions. PLoS ONE 20(2):e0315011. https://doi.org/10.1371/journal.pone.0315011
11. McBain RK, Bozick R, Diliberti M et al (2025) Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. JAMA Netw Open 8(11):e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281
12. Liu J, Gu J, Tong M et al (2025) Evaluating the agreement between ChatGPT-4 and validated questionnaires in screening for anxiety and depression in college students: a cross-sectional study. BMC Psychiatry 25:359. https://doi.org/10.1186/s12888-025-06798-0
13. Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem 61:1446–1452. https://doi.org/10.1373/clinchem.2015.246280

Figures

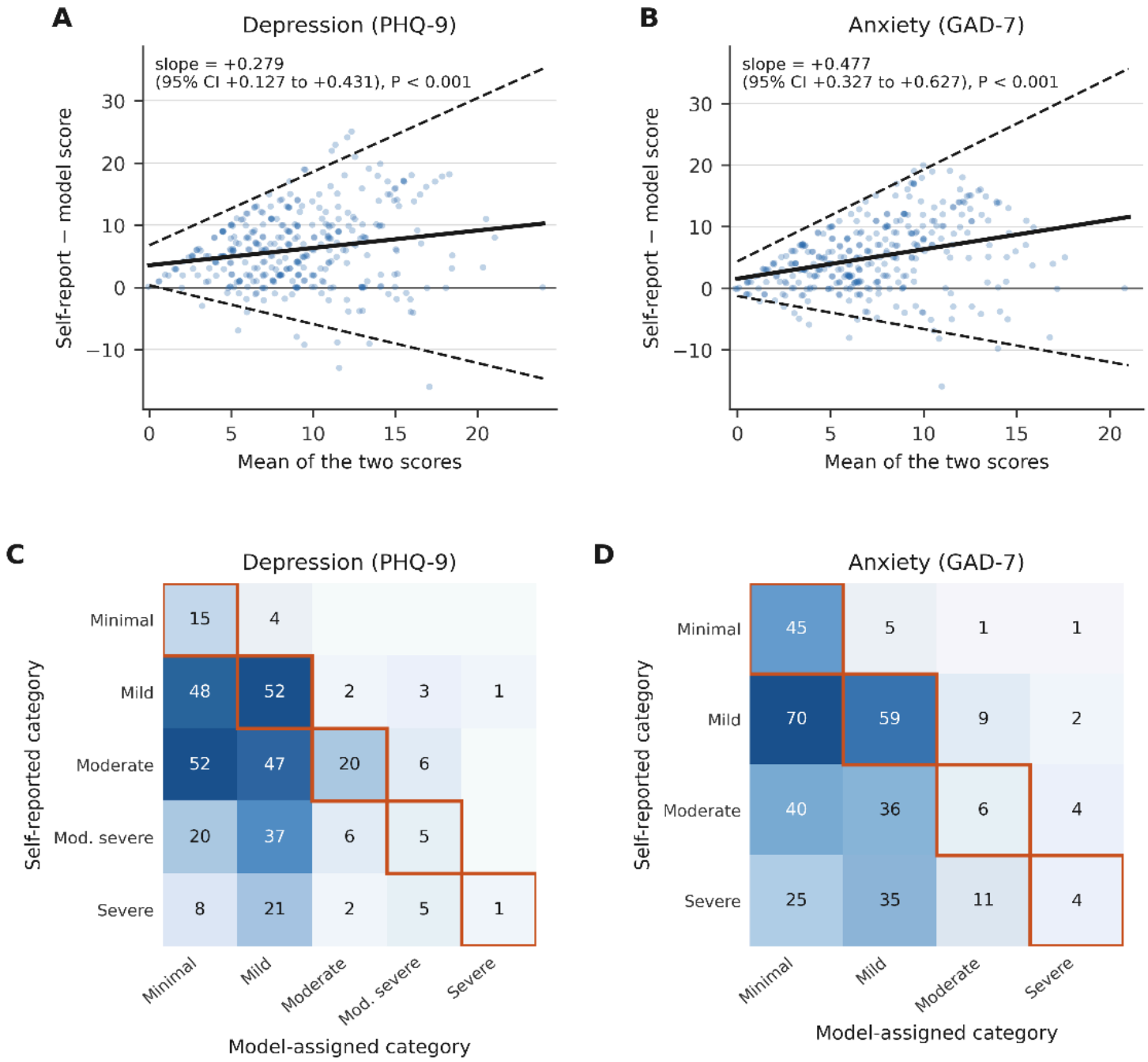

Figure 1

Agreement between self-report and model-derived scores. Panels A and B plot self-report minus model score against the mean of both scores for depression and anxiety. Panels C and D show counts by self-reported and model-assigned severity category.

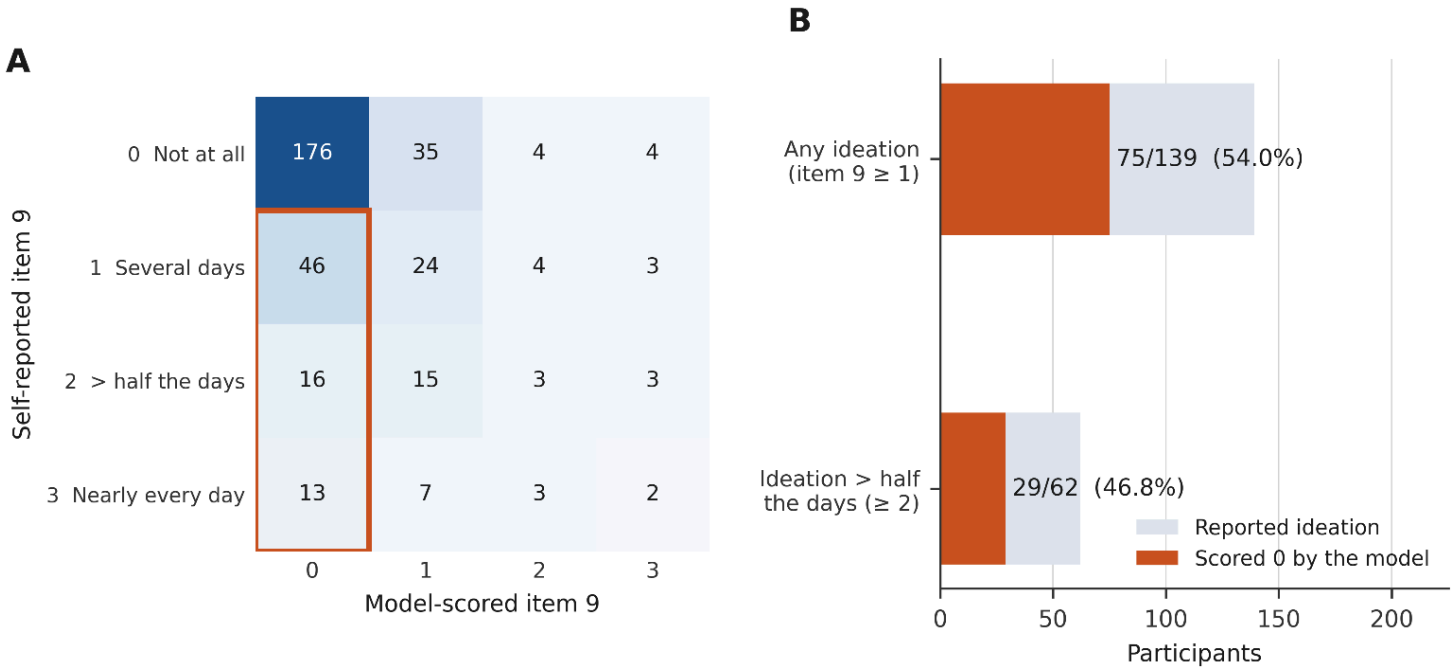

Figure 2

Agreement on PHQ-9 item 9. Panel A shows the paired item scores. Panel B shows how often the model assigned zero among students who endorsed the item at any frequency or on more than half the days.

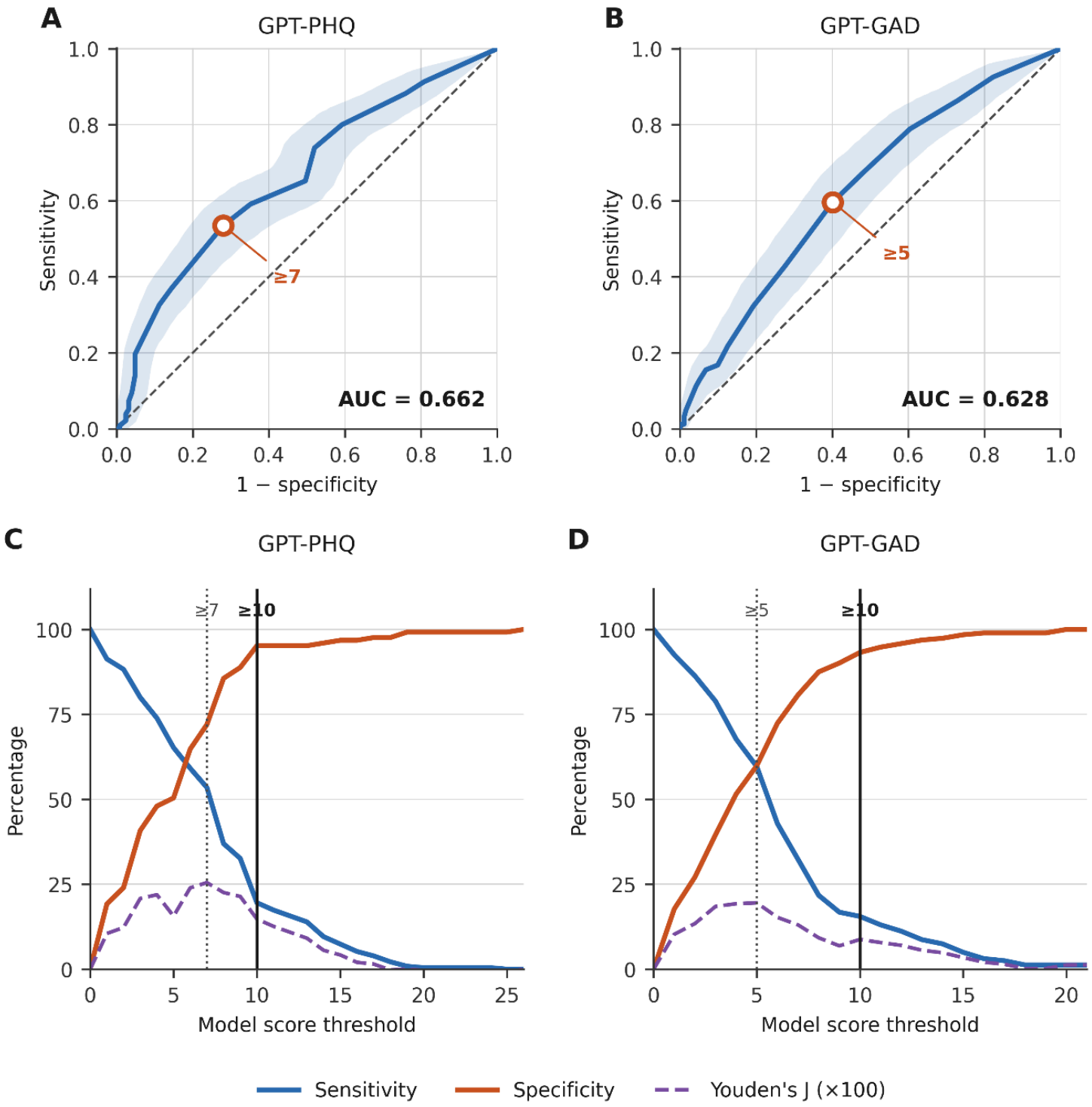

Figure 3

Screening performance against questionnaire-defined positive screens. Panels A and B show receiver operating characteristic curves for depression and anxiety. Panels C and D show sensitivity, specificity and Youden's index across model thresholds.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx