Counselling and Psychotherapy Research

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Comparing the GP-CORE and WEMWBS Scores as Predictors of Future Mental Health Service Utilisation and Help-Seeking Behaviour Among UK Students in Higher Education

Kehinde Akin-Akinyosoye[1,2] | Paul A. Tiffin[1,2] | Lewis W. Paton[1,2] | Paul Heron[1,2] | Lisa Edwards[3] | Bridgette M. Bewick[4] | James Tait[2] | Kristenn-Lenaig Leporcher[5] | Emma Broglia[5,6] | Dean McMillan[1,2] | Michael Barkham[5]

[1]Hull York Medical School, University of York, York, UK | [2]Department of Health Sciences, University of York, York, UK | [3]Faculty of Health Studies, University of Bradford, Bradford, UK | [4]School of Medicine, University of Leeds, Leeds, UK | [5]School of Psychology, University of Sheffield, Sheffield, UK | [6]Student Mental Health, Counselling and Therapies Service, The University of Sheffield, Sheffield, UK

Correspondence: Michael Barkham (m.barkham@sheffield.ac.uk)

Received: 21 November 2025 | **Revised:** 21 April 2026 | **Accepted:** 20 May 2026

Keywords: GP-CORE | helping behaviours | service utilisation | students | WEMWBS

ABSTRACT

Background: The General Population Clinical Outcomes in Routine Evaluation (GP-CORE) and the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) are two instruments often used to measure psychological distress or well-being in student samples. This study aimed to directly compare the ability of the two instruments to yield information on psychological distress in this population. Specifically, the ability of the scores to predict mental health service utilisation was compared.

Methods: GP-CORE and WEMWBS scores, sociodemographic and educational data were available for 1260 students at a UK University (Time 1) and for 265 students at 6-month follow-up (Time 2). Rasch calibration was performed on baseline data to evaluate test information, with item and person fit assessed using infit/outfit and item characteristic curves. Effect sizes (Cohen's d) and discriminative ability (area under the curve—AUC) were estimated for predicting future mental health service utilisation.

Results: Both measures were acceptably reliable ($\alpha > 0.8$), though the WEMWBS was better at discriminating between respondents, as indicated by the person separation indices (GP-CORE = 2.34; WEMWBS = 3.06). The GP-CORE scores were more sensitive to service utilisation status (Cohen's $d = 0.31$, $p = 0.02$) compared to those from the WEMWBS (Cohen's $d = 0.25$, $p = 0.06$). The GP-CORE scores were statistically significantly superior at discriminating between students reporting service utilisation vs. those that did not (AUC = 0.59 vs. 0.43, $p = 0.02$).

Conclusion: Although the WEMWBS scores may be superior at differentiating between different levels of 'psychological well-being', the GP-CORE scores are more predictive of service utilisation and may be more appropriate when screening students for potential mental health issues that warrant formal support.

1 | Introduction

Timely access to mental health services is essential for effective intervention and support, particularly among student populations (Lipson et al. 2022; Priestley et al. 2022; Zhao et al. 2025), who are at heightened risk of psychological distress compared to the general population of the same age (Baker and Kirk-Wade 2023; Lewis and Bolton 2023). Effective planning and resource allocation require tools that can indicate emerging service needs. Questionnaires measuring

Implications

Implications for Practice

1. Integrate the GP-CORE into triage and assessment processes to support resource planning: GP-CORE scores showed a stronger association with mental health service utilisation, but the WEMWBS scores did not. This suggests that embedding the GP-CORE into standard triage or assessment pathways within Higher Education settings could enhance decision-making around resource distribution and clinical prioritisation.
2. Review barriers to help-seeking and service engagement within university systems: GP-CORE scores alone predicted 'mental health service utilisation', whereas the GP-CORE and WEMWBS scores predicted 'help-seeking'. This highlights a potential gap between students' intentions and their actual engagement, indicating the need for institutions to review individual (e.g., stigma, literacy, symptom burden) and structural (e.g., accessibility, administrative processes) barriers within their support pathways.
3. Use the GP-CORE or WEMWBS for early identification of students whose mental health is worsening: The GP-CORE and WEMWBS are sensitive to identifying changes in distress and well-being levels in students within higher education settings, making them useful instruments for routine monitoring, enabling services to detect rising risk earlier and intervene in a timely manner.

Implications for Policy

1. Use GP-CORE and WEMWBS in ways aligned with the University Mental Health Charter (UMHC): The UMHC emphasises both promoting well-being across the whole university and providing accessible clinical support. The significant association between GP-CORE and mental health service use positions the GP-CORE as a more appropriate measure for informing policy decisions related to identifying students requiring clinical intervention and ensuring adequate service capacity, supporting evidence-informed approaches adopted within the UMHC framework. Although WEMWBS did not predict service utilisation, its precision in capturing population-level well-being makes it well suited for evaluating preventative and promotional initiatives encouraged by the UMHC.

distress or well-being can provide valuable data regarding mental health needs and help inform service use and help-seeking demands. Integrating such measures into service planning could enhance the understanding of help-seeking behaviours and inform strategies for improving access to timely and appropriate support.

The experience of well-being and psychological distress, while being on a continuum, is distinct. Well-being refers to positive feelings, resilience and the capacity to manage the pressures of student life, whereas psychological distress reflects difficulties that negatively affect functioning and exceed perceived coping capacity (Barkham et al. 2019). Whole-university frameworks, including the University Mental Health Charter and Universities UK Stepchange framework, aim to bridge the divide between preventative well-being promotion and clinical mental health services (Hughes and Spanner 2024; Universities UK 2020). Professional bodies such as the Association of Managers of Student Services in Higher Education (AMOSHE) have highlighted the need for evidence-based measures. Recommended tools for well-being include the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS; Tennant et al. 2007) and General Population-Clinical Outcomes in Routine Evaluation (GP-CORE; Sinclair et al. 2005), which assess broader well-being and distress (Dodd et al. 2021). Beyond well-being, popular tools for measuring psychological distress include the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al. 2001) and Generalised Anxiety Disorder-7 Assessment (GAD-7; Spitzer et al. 2006), which focus specifically on depression and anxiety symptoms, respectively.

GP-CORE captures psychological distress and general functioning without relying solely on severe clinical symptoms, making it a valuable tool for identifying individuals experiencing significant difficulties. Consequently, it has been recommended for tracking UK student cohorts (Dodd et al. 2021; Heron et al. 2023). In contrast, WEMWBS focuses on positive mental well-being and is widely used within university services. Despite this popularity, student feedback indicates a preference for assessment tools that capture a realistic mix of positive and negative experiences to better reflect student needs and help-seeking behaviours (Broglia et al. 2024; Priestley et al. 2022). Although both tools provide meaningful data on mental health status, their relative effectiveness in predicting help-seeking behaviour remains unclear. Comparing scales such as the GP-CORE and WEMWBS is crucial for student mental health and well-being services to identify the most responsive and effective tools for assessing the needs of service users who are more likely to require input from clinical, well-being or counselling services (hereby referred to as mental health services). By evaluating the performance of these scales, student mental health and well-being healthcare services can determine which measures are better at detecting meaningful differences in distress and well-being among this population. Utilising the most responsive scales at any time during the student journey better ensures that interventions are timely and tailored to the specific needs of individuals, ultimately enhancing the effectiveness of mental health services.

This study aimed to evaluate the utility of these two measures in a population of university students in relation to mental health service utilisation and help-seeking behaviours. Specifically, the study sought to

1. compare the psychometric properties of GP-CORE and WEMWBS, including measurement precision in this specific population;
2. assess the responsiveness of both measures to changes in distress and well-being in students over time; and
3. examine the association between scale scores at baseline and future mental health service use or help-seeking behaviours.

2 | Methods

2.1 | Study Population

Data were available from the ongoing 'Student Wellbeing At Northern England Universities' (SWANS) cohort study at the University of York. The SWANS study started in September 2019 and recruits new students every 6 months. Full details about eligibility and recruitment are available elsewhere (Paton et al. 2023). Potential participants were directed to an information sheet and informed that survey completion constituted informed consent for the study. At the end of the survey, participants were additionally invited to consent to: (i) being contacted every 6 months for follow-up surveys; (ii) receiving invitations to related research; and (iii) participating in future data-linkage studies. In subsequent waves, those who had consented to follow-up contact received two personalised email invitations in addition to the general study invitation.

We used data from two time points, drawn from ongoing waves of data collection: Time 1 (September/October 2023) and Time 2 (March/April 2024). For simplicity, we refer to these as Time 1 and Time 2 throughout this article. Data from the most recent waves were utilised to ensure that the analytic sample accurately represented the contemporary university student population at the time of the study. We had data available for $n = 1260$ students who participated in the survey at Time 1 as our baseline sample. Of these, $n = 265$ also completed the survey at Time 2. Data analysed for the current study are not publicly available due to privacy concerns but will be made available upon reasonable request.

The University's Department of Health Sciences Research Governance Committee granted ethical approval for this study (reference number HSRGC/2019/346/J).

2.2 | Measures

Students at each time point were asked to complete questions relating to their sociodemographic characteristics (e.g., age, gender identity, postcode, etc.), educational status (e.g., level of study) and a battery of mental health and wellbeing related measures (Paton et al. 2023), comprising the GP-CORE (Sinclair et al. 2005) and WEMWBS (Tennant et al. 2007).

Gender was recorded as male or female as recorded in the parent survey. Due to small numbers in the 'not known' category, gender was analysed as a binary variable in the current study. Age was categorised as under 21 or 21 and over, to differentiate between young and adult learners (HESA 2022). Ethnicity was presented in discrete categories as 'White', 'Mixed', 'Asian/Asian British' and 'Black/African/Caribbean'. Domicile was categorised as 'home' or 'overseas' and for home students, index of multiple deprivation (IMD) data were available and were grouped into three tertiles of relative deprivation in England as follows: IMD tertile 0 (most deprived) = IMD 1–3; IMD tertile 1 = IMD 4–7; and IMD tertile 2 (least deprived) = IMD 8–10. Distance or in-person learning was also captured, and level of study was dichotomised as undergraduate or postgraduate levels.

The GP-CORE is a 14-item self-report questionnaire designed to assess psychological distress and general functioning. To capture the range of student responses to their experiences, the measure comprises eight positively worded items (e.g., 'I have felt OK about myself') rated 4 = 'Not at all' to 0 = 'Most or all of the time' and six negatively worded (e.g., 'I have felt tense, anxious or nervous') rated 0 = 'Not at all' to 4 = 'Most or all of the time', with higher scores indicating greater levels of distress. Total scores range from 0 to 56. GP-CORE has demonstrated strong reliability (Cronbach's $\alpha \approx 0.82–0.90$) and validity in both clinical and non-clinical populations and is used for monitoring mental health outcomes (Sinclair et al. 2005).

The WEMWBS is a 14-item measure of positive mental wellbeing, assessing aspects such as optimism (e.g., 'I've been feeling optimistic about the future'), confidence (e.g., 'I've been feeling confident') and social functioning (e.g., 'I've been feeling close to other people'). Responses are recorded on a 5-point Likert scale (1 = 'None of the time' to 5 = 'All of the time'), with higher scores reflecting greater well-being. Total scores range from 14 to 70. WEMWBS has been validated for use in student populations and has been shown to be a reliable tool (Cronbach's $\alpha \approx 0.88–0.93$) and sensitive to changes in mental well-being over time (Tennant et al. 2007).

2.3 | Outcome Measures

We used the GP-CORE and WEMWBS scores at Time 2 as outcome measures.

Presence/absence of self-reported recent mental health service utilisation, as measured by a single-item measure ('have you *received* help for emotional distress, low wellbeing, or a mental health concern from a professional source of support (e.g., a talking therapy, medication, other professional advice to improve emotional wellbeing) within the past six months? "Yes," "No," or "Prefer not to say"'), served as the primary outcome measure.

Presence/absence of self-reported recent mental health help-seeking, as measured by a single-item measure ('have you *sought* help for emotional distress, low wellbeing, or a mental health concern from a professional source of support (e.g., from a GP, wellbeing, or counselling service) within the past six months? "Yes," "No" or "Prefer not to say"'), served as the secondary outcome measure.

2.4 | Statistical Analyses

The differences in categorical variables between baseline and follow-up groups were analysed using chi-squared test or Fisher's exact test when cell counts were small. Where appropriate, 95% confidence intervals (CIs) are presented. Due to low levels of missing data, all analyses used listwise deletion to account for missing data.

2.4.1 | Comparing the Responsiveness of GP-CORE and WEMWBS

Independent sample *t*-tests were used to assess whether GP-CORE and WEMWBS baseline scores differed between students who, at follow-up, reported utilising mental health services or having sought help in the last 6 months, and those who did not. Effect size was quantified using Cohen's *d*, calculated as the difference between baseline and follow-up mean scores on the GP-CORE and WEMWBS, divided by the standard deviation (SD) of baseline scores (Kazis et al. 1989). Cohen's (1985) guidelines were used to classify the magnitude of effect sizes: 0.2 (i.e., two-tenths of an SD) represents a small effect; 0.5 a moderate effect; and 0.8 a large effect. A mixed-design repeated-measures ANOVA was conducted to examine the interaction effect of time (baseline vs. follow-up) and outcome status (present vs. absent).

2.4.2 | Comparing the Psychometric Properties of the GP-CORE and WEMWBS

The Generalised Partial Credit Model (GPCM) (Muraki 1992) was applied to polytomous items from GP-CORE and WEMWBS at baseline. Item fit was explored using standardised residuals, chi-squared statistics and item characteristic curves (ICCs). Misfitting items or individuals were iteratively deleted and re-analysed until adequate fit statistics were achieved.

Reliability was evaluated using person reliability, item separation, item reliability and Person Separation Index (PSI). PSI values derived from Rasch of 0.7, 0.8 and 0.9 indicate the ability to differentiate between 2, 3 and 4 or more groups, respectively. We assessed internal consistency using Cronbach's alpha (Cronbach 1951). A coefficient over 0.70 was considered acceptable (Nunnally and Bernstein 1994). Associations between GP-CORE and WEMWBS total scores were examined using the Pearson correlation test.

To investigate construct validity and unidimensionality of the scales, confirmatory factor analysis (CFA) was conducted (McArdle 1996). Model fit was assessed using Tucker–Lewis index (TLI ($\geq$0.95 for excellent fit; $\geq$0.90 for acceptable fit)), comparative fit index (CFI ($\geq$0.95 for excellent fit; $\geq$0.90 for acceptable fit)) and root mean square error of approximation (RMSEA ($\leq$0.06 for excellent fit; $\leq$0.08 for acceptable fit)) (Hoyle 2023; Hu and Bentler 1998). Principal component analysis of standardised residuals (PCAR) verified unidimensionality (Boone and Staver 2020).

Measurement invariance across outcome groups at follow-up was evaluated. This was achieved by fitting a multigroup CFA for those with and without declared service utilisation. The relevant constraints across the groups were then applied (equal factor structure, loadings and thresholds) to test for configural, metric and scalar invariance. Items with sparse response categories were excluded to ensure model identification (Millsap and Yun-Tein 2004). Fit indices for metric and scalar models were compared with those for configural models. Although the chi-squared difference test traditionally reflects the presence of invariance (Chen 2007), Cheung and Rensvold (2002) recommend evaluating change in CFI values as chi-squared difference test may be sensitive to larger sample sizes. In this study, a reduction in CFI and TLI by >0.01, and a reduction in RMSEA by >0.015, indicate significantly worse model fit (Cheung and Rensvold 2002).

Sensitivity analyses evaluated the performance of a combined version of these questionnaires, using raw scores for psychometric analyses and Rasch-transformed scores for predictive analyses to determine whether combining them provided any added predictive or psychometric benefits.

2.4.3 | Comparing GP-CORE and WEMWBS as a Predictor of Outcomes at 6-Month Follow-Up

The predictive performance of both scale scores was assessed using univariate logistic regression models, examining the association between baseline scale scores and reported mental health service utilisation/help-seeking behaviours at follow-up. Discriminative ability was evaluated by estimating and comparing the area under the receiver operating characteristic (ROC) curve (AUC) for each scale, derived from the respective logistic regression models (Cleves 2002).

Descriptive statistics, logit models and ROC analyses were performed in Stata version 17 (StataCorp 2021). Factor analyses were performed in Mplus version 8.11 (Muthén and Muthén 2017) and GPCM analyses in R (Version 4.4.2; R Core Team 2023) and Winsteps (Version 5.2. 0.1) (Linacre 2023) was used to compute person separation reliability (PSI) and related Rasch-based item and scale fit statistics. Statistical significance was set at $p<0.05$.

3 | Results

3.1 | Baseline Characteristics

The study population comprised students who completed the study survey at baseline ($n=1260$) and at 6-month follow-up ($n=265$). Key participant characteristics at baseline were similar to those observed at follow-up (Table 1).

Compared to the 2023/24 University of York student population, the study sample at both baseline and follow-up showed consistent patterns in the relative distribution of key student characteristics. Undergraduates comprised a larger proportion of the student population than postgraduates (University: 75.1% vs. 24.9%; baseline: 69.9% vs. 30.1%; follow-up: 67.3% vs. 32.6%), females represented a greater proportion than males (University: 59.5% vs. 41.5%; baseline: 65.2% vs. 29.6%; follow-up: 68.2% vs. 26.9%), and students under 21 constituted a larger proportion than those aged 21 and over (University: 83.3% vs. 16.7%; baseline: 56.7% vs. 43.3%; follow-up: 58.5% vs. 41.5%). The overall proportion of younger students, male students and undergraduates was lower in the study sample compared to the wider University population. Ethnicity, deprivation and distance learning data were not available for the overall University population (University of York 2025).

TABLE 1 | Characteristics of the study group.

	N (%)		
Age group	**Baseline (*n* = 1260)**	**Follow-up (*n* = 265)**	***p***
Under 21	570 (56.7%)	138 (58.5%)	0.66
21 and over	436 (43.3%)	98 (41.5%)	
Sex			
Male	330 (31.2%)	66 (26.9%)	0.43
Female	727 (68.8%)	167 (68.2%)	
Ethnicity			
White	782 (71.2%)	179 (73.9%)	0.81
Mixed	56 (5.1%)	13 (5.4%)	
Asian/Asian British	215 (19.6%)	41 (16.9%)	
Black/African/Caribbean	22 (2.0%)	3 (1.3%)	
Other	23 (2.1%)	6 (2.5%)	
Distance Learner			
In-person	1047 (94.8%)	227 (92.7%)	0.22
Distance Learner	58 (5.2%)	18 (7.3%)	
Study level			
Undergraduate	754 (69.9%)	161 (67.4%)	0.44
Postgraduate	324 (30.1%)	78 (32.6%)	
Domicile			
Home	829 (74.6%)	194 (79.2%)	0.14
Overseas	283 (25.4%)	51 (20.8%)	
IMD deciles			
1–3 (highest deprivation)	99 (13.0%)	24 (13.4%)	0.98
4–7 (moderate deprivation)	277 (36.5%)	66 (36.9%)	
8–10 (lowest deprivation)	383 (50.5%)	89 (49.7%)	

Note: Analysis conducted using Fischer's exact test, excluding missing data.

3.2 | Comparing the Responsiveness of GP-CORE and WEMWBS

The mean change in GP-CORE scores between baseline and follow-up was 2.33 points (95% CI 1.39–3.26, $p < 0.001$), indicating a general increase in distress over time for those who completed measures at both time points. Change in WEMWBS scores between baseline and follow-up was −1.72 points (95% CI: −2.66 to −0.78, $p < 0.001$), indicating a general decrease in wellbeing over time.

Students who used mental health services had significantly higher distress scores at baseline, as measured by the GP-CORE, compared to those who had not used these services (2.91-point difference, $p = 0.02$). However, well-being scores at baseline, as measured by the WEMWBS, were not statistically significantly different between those students who reported service utilisation and those who did not (2.36, $p = 0.06$) (Table 2). Cohen's *d* for the GP-CORE was 0.31 and 0.25 for the WEMWBS.

Compared to students who had not sought help for mental health concerns, those who had sought help had significantly higher distress scores and significantly lower well-being scores at baseline (Table 2). Cohen's *d* for the GP-CORE was 0.33 and 0.35 for the WEMWBS.

A mixed effects model was conducted to examine the relationship between scale scores and service utilisation over time. Compared to those who had not utilised mental health services, service users reported significantly higher distress scores (GP-CORE, $\beta = 3.97$, $p < 0.001$) and significantly lower well-being scores (WEMWBS, $\beta = -3.16$, $p = 0.001$) across timepoints. However, the interaction between time and service utilisation was non-significant for both measures (GP-CORE, $\beta = -1.39$, $p = 0.151$; $\beta = 0.72$, $p = 0.465$), suggesting that distress or well-being scores did not change differently over time between service users and non-users. These findings remained consistent when analysing help-seeking behaviour as the outcome. Descriptive statistics for GP-CORE and WEMWBS scores at baseline and follow-up, as well as mean change scores, are presented by outcome status in Tables S1 and S2.

3.3 | Comparing the Psychometric Properties of GP-CORE and WEMWBS

GPCM of GP-CORE and WEMWBS scores at baseline revealed no item misfit after removing misfitting individuals, with item residuals within the −2.5 to +2.5 range. However, GP-CORE

TABLE 2 | Baseline scores for GP-CORE and WEMWBS scales, by mental health service utilisation and help-seeking in the past 6 months.

			GP-CORE		WEMWBS	
Outcome (yes/no)		***n* (%)**	**Mean (SD)**	***p***	**Mean (SD)**	***p***
Utilisation of mental health services in the past 6 months ($n = 259$)	Yes	90 (34.75)	25.46 (9.73)	0.02	42.17 (8.58)	0.06
	No	169 (65.25)	22.47 (9.04)		44.53 (10.01)	
Help-seeking in the past 6 months ($n = 259$)	Yes	93 (35.91)	25.55 (9.46)	0.01	41.60 (8.96)	0.004
	No	166 (64.09)	22.46 (9.16)		44.89 (9.74)	

Note: Analysis conducted using an independent sample *t*-test.

item 5 showed flat, overlapping ICC curves, suggesting poor discrimination and potential category redundancy. A residual correlation of 0.3 between GP-CORE items 2 and 8 indicated local dependency, whereas three WEMWBS items showed local dependency on GP-CORE items ($r = 0.33–0.46$). The PSI values were 2.34 for GP-CORE, 3.06 for WEMWBS and 3.50 for the combined version, indicating similar reliability. Item separation indices were 3.28 (GP-CORE), 3.64 (WEMWBS) and 4.25 (merged scale), with item reliabilities of 0.91, 0.93 and 0.95, respectively. The GP-CORE and WEMWBS total scores were strongly negatively correlated at both baseline ($r = -0.76$, $n = 1260$) and follow-up ($r = -0.86$, $n = 263$), indicating that higher levels of psychological distress were associated with lower levels of mental wellbeing.

A one factor model showed adequate fit for the GP-CORE (TLI = 0.94, CFI = 0.95, RMSEA = 0.09, SRMR = 0.04) and WEMWBS (TLI = 0.96, CFI = 0.96, RMSEA = 0.1, SRMR = 0.04) questionnaires. A method effect on negatively worded items significantly improved the fit for the GP-CORE. PCAR results supported unidimensionality for both questionnaires. For GP-CORE, the first contrast had an eigenvalue of 2.08, explaining 14.8% of the variance. WEMWBS had a first contrast eigenvalue of 1.81, explaining 12.9% of the variance. The combined questionnaire showed similar findings (eigenvalue of first contrast = 2.52, variance explained = 9.01%). For all measures, the size of the first contrast suggests that only a small amount of unexplained variance remained after accounting for the main factor. This indicates that each questionnaire largely measures a single underlying concept, with minimal evidence of additional hidden dimensions or factors influencing responses. Cronbach's α was 0.87 and 0.92 for the GP-CORE and WEMWBS, respectively, confirming that each assessed a homogeneous construct.

Five WEMWBS items (2, 3, 5, 8 and 10) showed sparse response distributions across mental health service use groups and were removed, leaving 9 items for invariance testing. No sparse distributions were observed for the GP-CORE, and all 14 items were retained.

Consistent with the one-group model fit indices reported above, configural invariance showed acceptable fit for both the GP-CORE (TLI = 0.94, CFI = 0.93, RMSEA = 0.10, SRMR = 0.08) and WEMWBS (TLI = 0.96, CFI = 0.95, RMSEA = 0.11), supporting a consistent factor structure across mental health service utilisation groups. Despite a significant chi-squared difference test ($p < 0.001$), metric invariance was supported for the GP-CORE as changes in approximate fit indices were minimal (ΔTLI = +0.001, ΔCFI = +0.005, ΔRMSEA = 0.000). For the WEMWBS, the chi-squared difference test was not statistically significant ($p > 0.05$); changes in TLI and CFI were acceptable (both Δ = +0.010), although there was a larger change in RMSEA (ΔRMSEA = −0.019). Given that changes in CFI and TLI were within commonly accepted criteria (≤0.01) and recognising the sensitivity and instability of RMSEA in invariance testing, particularly in small samples (Little 2013), these findings provide overall support for metric invariance of the WEMWBS across groups. Scalar invariance was also supported for the GP-CORE, as changes in fit indices were negligible (ΔTLI = +0.006, ΔCFI = +0.001, ΔRMSEA = −0.004), though the chi-squared difference test was significant ($p = 0.049$). For the WEMWBS, the chi-squared difference test was not significant ($p > 0.05$), and changes in fit indices were within accepted thresholds (ΔTLI = +0.012, ΔCFI = +0.004, ΔRMSEA = +0.009), supporting scalar invariance across groups.

The same invariance procedures were also applied to the help-seeking outcome, with configural, metric and scalar invariance supported for both measures (Tables S3 and S4).

3.4 | Comparing Predictive and Discriminative Performance of GP-CORE and WEMWBS

Table 3 displays the results from logistic regression models examining the association between baseline scores on the GP-CORE and/or WEMWBS, and reported mental health service use or help-seeking behaviour at 6-month follow-up.

As can be seen, baseline GP-CORE scores were a statistically significant predictor of reporting mental health service use in the past 6 months at follow-up (odds ratio [OR] = 1.03; 95% CI [1.01, 1.06], $p = 0.02$). In contrast, WEMWBS scores were not statistically significant predictors of reporting mental health service use (OR = 0.97; 95% CI [0.95, 1.00], $p = 0.06$). Baseline scores on both scales were statistically significant predictors for reporting help-seeking in the past 6 months at follow-up (GP-CORE; OR = 1.04, $p = 0.01$; WEMWBS OR = 0.96, $p = 0.01$). In the same model, neither the GP-CORE nor the WEMWBS questionnaire scores were independent predictors of service use or help-seeking behaviour.

The Rasch-derived combined score, based on responses from both questionnaires, was a statistically significant predictor of reporting mental health service use (OR = 1.44; 95% CI [1.09, 1.91], $p = 0.01$) and reporting help-seeking (OR = 1.55; 95% CI [1.17, 2.06], $p = 0.002$).

TABLE 3 | Prediction of outcomes at 6-month follow-up by baseline GP-CORE and WEMWBS scales.

		Utilisation of mental health services		Help-seeking	
		OR (95% CI)	*p*	OR (95% CI)	*p*
Unadjusted	GP-CORE	1.03 (1.01–1.06)	0.02	1.04 (1.01 to 1.07)	0.01
	WEMWBS	0.97 (0.95–1.00)	0.06	0.96 (0.94–0.99)	0.01
Adjusted[a]	GP-CORE	1.04 (0.96–1.05)	0.94	1.02 (0.97–1.07)	0.31
	WEMWBS	1.00 (0.99–1.09)	0.15	0.98 (0.93–1.02)	0.49

[a]GP-CORE and WEMWBS were adjusted for each other in the same model.

ROC analysis indicated that GP-CORE demonstrated significantly better discriminative ability than WEMWBS in discriminating between mental health service users and non-users (AUC = 0.59 vs. 0.43, $p = 0.02$) (Figure 1). The discriminative performance of the Rasch-derived combined score (AUC = 0.59) was similar to that of the GP-CORE ($p = 0.89$) but was significantly better than the WEMWBS discriminative performance ($p = 0.03$). However, the absolute discriminative performance of both measures individually or combined was limited, with GP-CORE showing only poor predictive value (AUC < 0.7) and WEMWBS performing worse than chance (AUC < 0.5). These findings remained consistent when analysing help-seeking behaviour as the outcome (Figure 2).

4 | Discussion

In this study, we assessed the psychometric properties of the GP-CORE and WEMWBS in a sample of UK Higher Education students. We observed that although both measures exhibited good psychometric fit in our sample, they showed different predictive and psychometric strengths when assessing mental health service utilisation and help-seeking behaviours. The GP-CORE demonstrated greater responsiveness to change over time, as evidenced by a larger shift in mean scores from baseline to follow-up. Additionally, it had a marginally higher effect size (Cohen's *d*) compared to the WEMWBS, indicating a slightly stronger ability to detect meaningful changes in distress or well-being over time. Importantly, the GP-CORE was significantly associated with service utilisation, whereas the WEMWBS showed no such relationship, suggesting that the GP-CORE may be more relevant in identifying individuals who do engage with mental health services.

In contrast, the WEMWBS had a slightly higher Person Separation Index (PSI) score, reflecting marginally stronger measurement precision in differentiating between individuals across the latent trait. This suggests that while the WEMWBS may provide a stable measure of well-being across a population, it may be less sensitive to changes in distress levels over time or to variations in help-seeking behaviours.

Consistent with COSMIN guidelines (Gagnier et al. 2021), the psychometric analyses provide evidence for the structural validity and internal consistency of both instruments in this population, based on acceptable factor structure and reliability estimates. Predictive validity was more modest, with both measures showing limited ability to predict future service use and help-seeking. Despite this, the strong internal consistency and significant association with service utilisation reinforce their utility in mental health research and practice. Configural, metric and scalar invariance were also supported for both instruments, although a reduced version of the WEMWBS was required because of sparse response distributions.

We showed high internal consistency of both measures (GP-CORE $\alpha = 0.87$; WEMWBS $\alpha = 0.92$), mirroring existing findings for the WEMWBS (α range = 0.99–0.93) (Clarke et al. 2011) and GP-CORE (α range = 0.82–0.90) (Sinclair et al. 2005) in similar student populations. This suggests effectiveness of both measures in distinguishing between individuals with higher and lower levels of distress or well-being. Consistent with the high internal consistency observed, the PSI values (GP-CORE = 2.34, WEMWBS = 3.06, combined = 3.50) further indicated adequate to good discrimination. However, the GP-CORE exhibited some local dependency and item redundancy, which may have inflated model fit and affected item

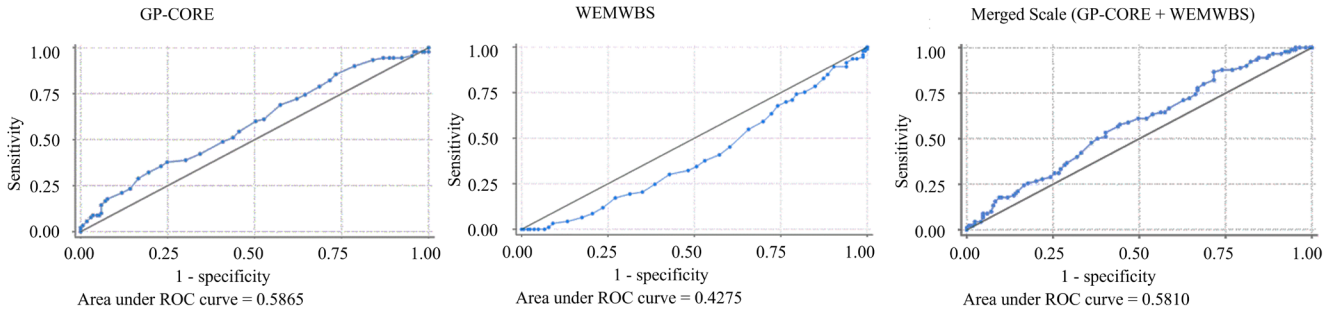

FIGURE 1 | Comparison of receiver-operating characteristic (ROC) curves for the GP-CORE, WEMWBS and merged scale in discriminating mental health service users from non-users. Unadjusted models presented.

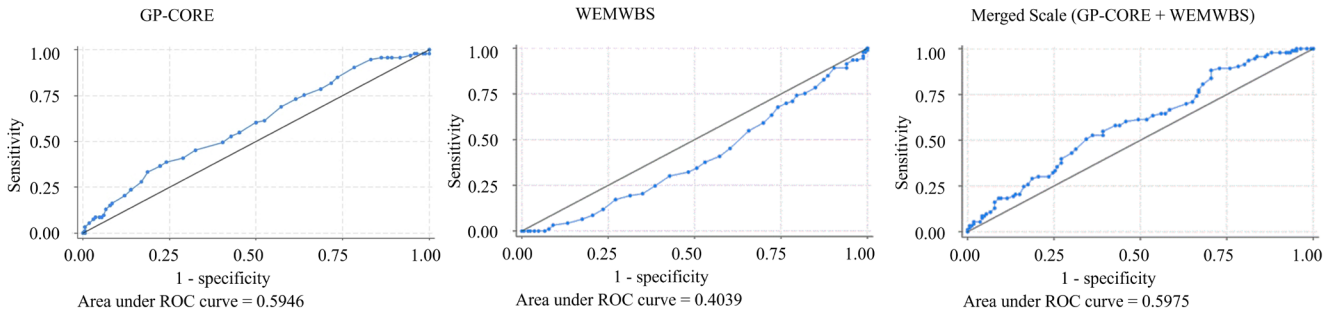

FIGURE 2 | Comparison of receiver-operating characteristic (ROC) curves for the GP-CORE, WEMWBS, and merged scale in discriminating mental help seekers from non-help seekers. Unadjusted models presented.

independence. Item 5 ('I have been troubled by aches, pains or other physical problems') showed limited discrimination, with flat, overlapping ICCs indicating reduced sensitivity across levels of psychological distress. In a university student sample, somatic complaints may reflect factors unrelated to mental health, such as fatigue, injury, or poor ergonomics, reducing the item's specificity to psychological distress. Although collapsing categories or removing the item may have improved model fit, the standard 14-item structure was retained to preserve comparability with existing norms. Item 5 should therefore be interpreted cautiously in isolation, as it may provide less precise information about mental health status than more emotionally focused items. Evidence from the current study supports the unidimensionality of both measures. This finding aligns with previous factor analytic studies of the WEMWBS, which have also demonstrated a unidimensional structure in general population and student samples (Clarke et al. 2011; Tennant et al. 2007). Although four components of positive and negative subjective well-being, physical problems and social functioning have previously been identified within the GP-CORE in general population samples (Sinclair et al. 2005), the respective subscale scores are typically highly correlated. Accordingly, Sinclair et al. (2005) recommended that the GP-CORE be treated as a unidimensional measure, supporting the findings in the current study. Notably, the 'negative subjective well-being' component identified in their PCA supports the presence of a method effect associated with negatively phrased items, as indicated in our current CFA findings. Residual correlations between WEMWBS and GP-CORE items suggested overlapping variance between well-being and distress beyond their shared latent traits. Although GP-CORE and WEMWBS scores were moderately negatively correlated, the measures remained distinguishable, supporting their assessment of related but distinct constructs. This overlap may explain why combining the measures did not meaningfully improve predictive performance, suggesting that measure selection should be driven by assessment purpose rather than an assumption that combining well-being and distress measures necessarily adds value.

Differences in distress and well-being levels were observed between service users and non-users, as well as between individuals who sought help and those who did not. These findings are consistent with previous research demonstrating associations between the GP-CORE and current help-seeking behaviour in the general population (Sinclair et al. 2005), and similar patterns have been reported for the WEMWBS in student samples (Moss et al. 2022). Importantly, the present study extends this literature by demonstrating that both measures are also associated with future mental health service use and help-seeking behaviour. This prospective association provides novel evidence for the predictive utility of these measures in identifying individuals who may engage with mental health support over time. The establishment of invariance confirms that the mean differences reported in this study reflect genuine variation in psychological distress and well-being rather than measurement bias.

Our interaction analysis found no significant differences in how these distress and well-being levels changed over time between help-seeking and non-help-seeking groups. This may be due to variability in the types and effectiveness of interventions received at follow-up, as well as individual differences in response to treatment (Berli et al. 2021). Future research should consider controlling for intervention type and intensity to better understand the impact on distress and well-being over time.

Although this study found differences in the AUC values of GP-CORE and WEMWBS, the GP-CORE demonstrated better predictive value (AUC = 0.59) compared with WEMWBS (AUC = 0.43). The lack of improved predictive performance when combining the measures may reflect that WEMWBS captures the lower-severity, positive end of the mental health continuum, whereas GP-CORE is more sensitive to distress and symptom burden and therefore already captures most of the variance relevant to service utilisation. Although the GP-CORE shows some utility, the overall limited AUC values highlight the need for future research aimed at identifying additional predictors to improve predictive performance. However, it is important to acknowledge the limitations of relying solely on AUC as an indicator of a measure's practical utility (De Hond et al. 2022). AUC thresholds for discrimination are inherently arbitrary, and high discriminatory ability alone is not sufficient to justify the application of a measure in practice. The primary value of AUC lies in comparing the relative discriminative performance of different models or measures rather than serving as a definitive measure of predictive validity.

4.1 | Implications for Practice

Both GP-CORE and WEMWBS showed acceptable psychometric properties in this student sample, but limited ability to distinguish individuals who later used mental health services, suggesting they may be better suited to assessing population mental health than predicting individual service use. Despite their limitations in predicting help-seeking behaviours, the GP-CORE's superior responsiveness, marginally higher effect size, and significant association with service utilisation suggest it may be particularly useful for identifying emerging service needs and informing resource allocation (Hughes et al. 2024). By capturing fluctuations in distress over time and aligning more closely with actual service engagement, its use may provide valuable data to guide mental health service planning and ensure that support is available when needed.

Our findings hold important implications when framed within UK national policies like the UMHC, which promotes a whole-university approach that balances preventative mental health promotion with the provision of clinical services (Hughes and Spanner 2024). While the WEMWBS was not a significant predictor of service utilisation, it may serve as a precise instrument for evaluating population-level well-being initiatives and institutional conditions that enable students to thrive, rather than forecasting demand for support services. In contrast, the GP-CORE is better suited to the domain of providing accessible support; its statistically significant association with service utilisation makes it a more appropriate tool for identifying students who may require clinical intervention and for informing resource allocation. Moreover, the discrepancy in predictors for 'help-seeking' versus 'service utilisation' may indicate a gap between a student's intention to seek help and their successful engagement with services, which should prompt institutions to examine

potential individual barriers (e.g., stigma, literacy, symptom burden) and structural barriers (e.g., access, processes) in their support pathways (Broglia et al. 2021; Cage et al. 2020).

Together, the findings highlight the importance of using psychometrically sound and responsive measures to inform mental health service provision. Incorporating such measures into routine student mental health assessments could improve service planning, enabling institutions to allocate resources more effectively and provide timely support. However, given the poor discrimination ability of both scales, future research should explore complementary approaches that integrate broader contextual factors influencing help-seeking behaviours.

4.2 | Strengths and Limitations

This study employs a longitudinal cohort design and provides a novel comparative analysis of both GP-CORE and WEMWBS to further enrich the understanding of their respective utility in assessing mental well-being. The study, however, has limitations. The low follow-up rate introduces potential attrition and non-response bias, which may limit the generalisability of longitudinal findings. Self-report measures for both predictors and outcomes introduce a risk of reporting bias, particularly concerning sensitive topics like mental health service use. The 6-month follow-up may have missed delayed service use due to barriers, waiting times or varied help-seeking pathways, while the binary sex categories used did not capture the full diversity of sex and gender identities. Furthermore, the reliance on single-item questions for help-seeking and service utilisation outcomes may lack sensitivity and fail to capture the complexity of mental health support experiences.

4.3 | Future Research

Future efforts should focus on integrating the use of the GP-CORE alongside other indicators—such as service access data, patterns of engagement and demographic or social determinants. This will provide a more complete model of mental health service demands (Graham et al. 2017). By combining psychological distress measures with contextual factors, service providers can better anticipate fluctuations in demand, allocate resources more effectively, and implement targeted interventions. This approach could facilitate earlier identification of individuals requiring support, improve service accessibility and ultimately enhance mental health outcomes within student populations.

5 | Conclusion

In conclusion, these findings suggest a modest predictive power of the GP-CORE over that of the WEMWBS in identifying students engaging with mental health services. WEMWBS remains a valuable measure of positive mental health and population wellbeing, despite its more limited utility for predicting future service use. Further research should seek to validate these findings using objective mental health service utilisation data to provide more definitive evidence of their real-world applicability.

Ethics Statement

The University's Department of Health Sciences Research Governance Committee granted ethical approval for this study (reference number HSRGC/2019/346/J).

Conflicts of Interest

M.B. was a co-developer of the GP-CORE but receives no financial gain from its use. The other authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Baker, C., and E. Kirk-Wade. 2023. *Mental Health Statistics: Prevalence, Services and Funding in England*. Vol. 13. House of Commons Library.

Barkham, M., E. Broglia, G. Dufour, et al. 2019. "Student Mental Health and Wellbeing: Definitions, Developmental Transitions, and Data Sets." *Counselling and Psychotherapy Research* 19, no. 4: 351–357. https://doi.org/10.1002/capr.12227.

Berli, C., J. Inauen, G. Stadler, U. Scholz, and P. E. Shrout. 2021. "Understanding Between-Person Interventions With Time-Intensive Longitudinal Outcome Data: Longitudinal Mediation Analyses." *Annals of Behavioral Medicine* 55, no. 5: 476–488. https://doi.org/10.1093/abm/kaaa066.

Boone, W. J., and J. R. Staver. 2020. "Principal Component Analysis of Residuals (PCAR)." In *Advances in Rasch Analyses in the Human Sciences*, 13–24. Springer International Publishing.

Broglia, E., K.-L. Leporcher, N. Deglurkar, et al. 2024. "Insights for Practitioners From Student-Led Research: Measuring Student Mental Health and Wellbeing." https://sites.google.com/sheffield.ac.uk/share/home.

Broglia, E., A. Millings, and M. Barkham. 2021. "Student Mental Health Profiles and Barriers to Help Seeking: When and Why Students Seek Help for a Mental Health Concern." *Counselling and Psychotherapy Research* 21, no. 4: 816–826. https://doi.org/10.1002/capr.12462.

Cage, E., M. Stock, A. Sharpington, E. Pitman, and R. Batchelor. 2020. "Barriers to Accessing Support for Mental Health Issues at University." *Studies in Higher Education* 45, no. 8: 1637–1649. https://doi.org/10.1080/03075079.2018.1544237.

Chen, F. F. 2007. "Sensitivity of Goodness of Fit Indexes to *Lack of Measurement Invariance*." *Structural Equation Modeling: A Multidisciplinary Journal* 14, no. 3: 464–504. https://doi.org/10.1080/10705510701301834.

Cheung, G. W., and R. B. Rensvold. 2002. "Evaluating Goodness-Of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9, no. 2: 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

Clarke, A., T. Friede, R. Putz, et al. 2011. "Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): Validated for Teenage School Students in England and Scotland. A Mixed Methods Assessment." *BMC Public Health* 11, no. 1: 487. https://doi.org/10.1186/1471-2458-11-487.

Cleves, M. A. 2002. "From the Help Desk: Comparing Areas Under Receiver Operating Characteristic Curves From Two or More Probit or Logit Models." *Stata Journal* 2, no. 3: 301–313.

Cohen, J. 1985. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16, no. 3: 297–334. https://doi.org/10.1007/BF02310555.

De Hond, A. A., E. W. Steyerberg, and B. Van Calster. 2022. "Interpreting Area Under the Receiver Operating Characteristic Curve." *Lancet Digital Health* 4, no. 12: e853–e855. https://doi.org/10.1016/S2589-7500(22)00188-1.

Dodd, A., M. Priestley, and N. C. Byrom. 2021. "University Student Well-Being in the United Kingdom: A Scoping Review of Its Conceptualisation and Measurement." *Journal of Mental Health* 30, no. 3: 375–387. https://doi.org/10.1080/09638237.2021.1875419.

Gagnier, J. J., J. Lai, L. B. Mokkink, and C. B. Terwee. 2021. "COSMIN Reporting Guideline for Studies on Measurement Properties of Patient-Reported Outcome Measures." *Quality of Life Research* 30, no. 8: 2197–2218. https://doi.org/10.1007/s11136-021-02822-4.

Graham, A., P. Hasking, J. Brooker, D. Clarke, and G. Meadows. 2017. "Mental Health Service Use Among Those With Depression: An Exploration Using Andersen's Behavioral Model of Health Service Use." *Journal of Affective Disorders* 208: 170–176. https://doi.org/10.1016/j.jad.2016.08.074.

Heron, P., K. Balloo, M. Barkham, et al. 2023. *Measuring Psychological Wellbeing and Mental Health in University Student Cohorts*. SMaRteN. https://kclpure.kcl.ac.uk/portal/en/publications/measuring-psychological-wellbeing-and-mental-health-in-university/.

Higher Education Statistics Agency. 2022. "Definitions: 2022/23." https://www.hesa.ac.uk/support/definitions/students.

Hoyle, R. H., ed. 2023. *Handbook of Structural Equation Modeling*. Guilford Publications.

Hu, L. T., and P. M. Bentler. 1998. "Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification." *Psychological Methods* 3, no. 4: 424–453. https://doi.org/10.1037/1082-989X.3.4.424.

Hughes, G., M. Priestley, and L. Spanner. 2024. "Stretched at Both Ends: Pressure on Student Services and the Impact on Academic Staff at UK Universities." *Education in Science* 15, no. 1: 13. https://doi.org/10.3390/educsci15010013.

Hughes, G., and L. Spanner. 2024. *The University Mental Health Charter*. 2nd ed. Student Minds. https://hub.studentminds.org.uk/university-mental-health-charter/.

Kazis, L. E., J. J. Anderson, and R. F. Meenan. 1989. "Effect Sizes for Interpreting Changes in Health Status." *Medical Care* 27: S178–S189. https://doi.org/10.1097/00005650-198903001-00015.

Kroenke, K., R. L. Spitzer, and J. B. Williams. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16, no. 9: 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Lewis, J., and P. Bolton. 2023. *Student Mental Health in England: Statistics, Policy, and Guidance*. Vol. 4. House of Commons Library.

Linacre, J. M. 2023. *WINSTEPS (Version 5.2. 0.1) [Computer Software]*. Winsteps.com.

Lipson, S. K., S. Zhou, S. Abelson, et al. 2022. "Trends in College Student Mental Health and Help-Seeking by Race/Ethnicity: Findings From the National Healthy Minds Study, 2013–2021." *Journal of Affective Disorders* 306: 138–147. https://doi.org/10.1016/j.jad.2022.03.038.

Little, T. D. 2013. *Longitudinal Structural Equation Modeling*. Guilford Press.

McArdle, J. J. 1996. "Current Directions in Structural Factor Analysis." *Current Directions in Psychological Science* 5, no. 1: 11–18. https://doi.org/10.1111/1467-8721.ep10772681.

Millsap, R. E., and J. Yun-Tein. 2004. "Assessing Factorial Invariance in Ordered-Categorical Measures." *Multivariate Behavioral Research* 39, no. 3: 479–515. https://doi.org/10.1207/S15327906MBR3903_4.

Moss, R. A., P. Gorczynski, W. Sims-Schouten, K. Heard-Laureote, and J. Creaton. 2022. "Mental Health and Wellbeing of Postgraduate Researchers: Exploring the Relationship Between Mental Health Literacy, Help-Seeking Behaviour, Psychological Distress, and Wellbeing." *Higher Education Research and Development* 41, no. 4: 1168–1183. https://doi.org/10.1080/07294360.2021.1906210.

Muraki, E. 1992. "A Generalized Partial Credit Model: Application of an EM Algorithm." *ETS Research Report Series* 1992, no. 1: 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

Muthén, B., and L. Muthén. 2017. "MPlus version 8." Los Angeles, CA.

Nunnally, J. C., and I. H. Bernstein. 1994. *Psychometric Theory*. 3rd ed. McGraw-Hill.

Paton, L. W., P. A. Tiffin, M. Barkham, et al. 2023. "Mental Health Trajectories in University Students Across the COVID-19 Pandemic: Findings From the Student Wellbeing at Northern England Universities Prospective Cohort Study." *Frontiers in Public Health* 11: 1188690. https://doi.org/10.3389/fpubh.2023.1188690.

Priestley, M., E. Broglia, G. Hughes, and L. Spanner. 2022. "Student Perspectives on Improving Mental Health Support Services at University." *Counselling and Psychotherapy Research* 22, no. 1: 197–206. https://doi.org/10.1002/capr.12391.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing (Version 4.3.1)." https://www.R-project.org/.

Sinclair, A., M. Barkham, C. Evans, J. Connell, and K. Audin. 2005. "Rationale and Development of a General Population Well-Being Measure: Psychometric Status of the GP-CORE in a Student Sample." *British Journal of Guidance and Counselling* 33, no. 2: 153–173. https://doi.org/10.1080/03069880500132581.

Spitzer, R. L., K. Kroenke, J. B. Williams, and B. Löwe. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7." *Archives of Internal Medicine* 166, no. 10: 1092–1097. https://doi.org/10.1001/archinte.166.10.1092.

StataCorp. 2021. "Stata Statistical Software: Release 17." College Station, TX.

Tennant, R., L. Hiller, R. Fishwick, et al. 2007. "The Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): Development and UK Validation." *Health and Quality of Life Outcomes* 5: 63. https://doi.org/10.1186/1477-7525-5-63.

Universities UK. 2020. *Stepchange: Mentally Healthy Universities*. Universities UK. https://www.universitiesuk.ac.uk/stepchange.

University of York. 2025. "Student Statistics." https://www.york.ac.uk/about/student-statistics/.

Zhao, R., Y. Amanvermez, J. Pei, et al. 2025. "Research Review: Help-Seeking Intentions, Behaviors, and Barriers in College Students–a Systematic Review and Meta-Analysis." *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 66, no. 10: 1593–1605. https://doi.org/10.1111/jcpp.14145.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Baseline, follow-up and change scores for the GP-CORE scale by mental health service utilisation and help-seeking in the past 6 months. **Table S2:** Baseline, follow-up and change scores for the WEMWBS by mental health service utilisation and help-seeking in the past 6 months. **Table S3:** Fit indices and comparisons for measurement invariance for the GP-CORE. **Table S4:** Fit indices and comparisons for measurement invariance for the WEMWBS.