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.

Grade and sex differences in psychological distress among Chinese vocational college students in routine digital screening: examining response engagement and measurement invariance

Shuai Yang

bs241011013@sust.edu.cn

Shaanxi University of Science and Technology

Ping Fan

Pingliang Vocational and Technical College

Ming Ren

Pingliang Vocational and Technical College

Juan Chao

Pingliang Vocational and Technical College

Xinyan Hu

Pingliang Vocational and Technical College

Pei Mu

Pingliang Vocational and Technical College

Research Article

Keywords: Psychological distress, Symptom Checklist-90, Vocational college students, First-year adjustment, Sex differences, Response engagement, Measurement invariance, Cross-sectional study

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11419807/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

Vocational college students are a large, under-studied group in China. Group comparisons from routine digital screening with the Symptom Checklist-90 (SCL-90) may reflect response engagement, floor effects and measurement non-invariance besides true differences in distress. We estimated adjusted group differences in distress, described response engagement, and tested robustness to these threats.

Methods

Cross-sectional analysis of 5,602 SCL-90 records (participation 95.4%) from the 2026 routine screening at a higher vocational college in China. The Global Severity Index (GSI) was regressed on sex, grade, ethnicity and college with class-clustered standard errors, checked by cluster bootstrap and mixed models. Measurement invariance was examined with multi-group confirmatory factor analysis and differential item functioning; floor effects with Tobit, rank-based and two-part models; and response quality with five prespecified and two supplementary sensitivity samples. The sex × grade interaction was exploratory.

Results

Mean GSI (1.261) was below Chinese adult (d = − 0.42) and undergraduate (d = − 0.48) norms; 25.5% of records answered “not at all” to all items. Scalar invariance held across sex and between grades 2025 and 2026. Compared with grade 2025, new entrants (grade 2026) had higher GSI (B = 0.236, 95% CI [0.195, 0.277]; d = 0.54) and grade 2024 lower GSI (B = − 0.142, 95% CI [− 0.198, − 0.086]). The female–male difference (B = 0.061, 95% CI [0.026, 0.095]) depended on grade ($F(2, 202) = 18.02$, $p < .001$) and was concentrated among new entrants (B = 0.178, 95% CI [0.102, 0.254]); the interaction persisted in a Tobit model ($p = .006$) but weakened on the rank scale ($p = .030$). Invariant responding fell from 50.2% (grade 2024) to 12.9% (grade 2026) and clustered by class (intraclass correlation 0.092). Grade and sex findings were stable across sensitivity analyses; the ethnicity contrast was not.

Conclusions

Distress was highest, and the sex gap largest, among students screened weeks after entry, and neither appeared attributable to measurement non-invariance or floor effects. Lower, class-clustered engagement in senior cohorts suggests that their lower distress may partly reflect repeated screening and administration. Screening reports should include response-quality and invariance checks and distinguish screening flags from diagnoses.

Introduction

University study is a period of high risk for the onset and worsening of mental disorders. The World Health Organization World Mental Health International College Student project found that about one third of first-year students screened positive for at least one common mental disorder in the previous 12 months [1], and the detected prevalence of depression among Chinese university students is also high [2, 3]. Higher vocational education combines academic demands with vocational skills training, placement internships and early employment decisions, and vocational college students therefore form a large but under-studied group. A recent longitudinal study found that vocational college students' SCL-90 total scores declined markedly from entry to two years later; female students scored higher than male students at entry, but the difference had disappeared two years later [4]. Large studies including vocational college samples have also linked psychological symptoms to sex, grade and urban–rural background [5]. These findings suggest that distress among vocational college students may be concentrated in the period of adjustment after entry and that sex differences may change with year of study, but this needs to be tested in larger samples.

The 90-item Symptom Checklist (SCL-90) [6, 7] is widely used in campus screening because it yields both an overall severity score and nine symptom-dimension scores. The apparent precision of dimension scores, however, does not imply corresponding discriminant validity. In Chinese undergraduates, the SCL-90 shows very high internal consistency but limited discriminant validity between dimensions [8]. A bifactor analysis of 14,371 first-year students found that a general severity factor explained about 81% of the common variance (explained common variance, ECV = 0.813), whereas the specific reliabilities of the dimensions were generally low (ωHS = 0.02–0.29) [9]. Studies in other populations likewise compare one-factor, correlated-factor, higher-order and bifactor structures rather than assuming that the nine dimension means represent distinct constructs [10]. Treating the Global Severity Index (GSI) as the primary outcome, and dimension results as secondary and exploratory evidence, is therefore more consistent with the psychometric evidence.

The choice of thresholds also affects conclusions. Chinese studies have long used historical rules such as a total score ≥ 160, ≥ 43 positive items or any dimension mean ≥ 2, but the inequality operator is applied inconsistently. Studies updating the norms show that applying old norms [11] or a dimension-mean cut-off of 2 to contemporary samples can produce widely differing detection rates [12]. Undergraduate norms can serve as a descriptive reference, but a self-report screening instrument cannot replace diagnostic interviews [13]. Such proportions are interpretable only when the rules and their operators are reported in full and the proportions are described as "screening flags" rather than "prevalence".

Demographic differences reported in previous studies are inconsistent. Some studies found higher symptom levels among female students, students from rural areas or students in particular grades [5, 13, 14], whereas longitudinal and long-term trend studies show that the size and direction of sex differences may change over time and with year of study [15–17]. From a developmental and adjustment

perspective, the transition to college concentrates stressors such as leaving home, reorganizing daily life, new academic demands and rebuilding social networks; psychological functioning changes markedly during the first year, and its trajectories may differ by sex [18]. Mental health problems in the first year are associated with later academic performance and help-seeking [19, 20], so knowing whether distress is concentrated at a particular stage after entry and in one sex is directly relevant to the timing of campus screening and follow-up. Cross-sectional grade differences, however, combine stage of study, entry cohort, age, assessment timing, previous screening experience and selective participation, and cannot be read directly as developmental trajectories.

Routine digital screening also faces problems of response quality. Very short completion times and fully invariant responding (for example, answering “not at all” to all 90 items) may indicate low effort, but neither is a perfect indicator: answering every item with the lowest option may reflect a genuinely symptom-free state as well as low effort or socially desirable responding, and time thresholds (a common reference is about 2 s per item) depend on reading speed, devices and testing conditions [21]. The methodological literature recommends combining several indicators and transparently reporting how substantive conclusions change across cleaning rules, rather than choosing post hoc the rule most favourable to a hypothesis [22–24]. In large campus screenings, if the proportion of invariant responding differs between groups, the cleaning rule itself can create or mask group differences, a point that has received little attention.

Response engagement can also be understood through survey response theory. Satisficing theory holds that when respondents’ motivation or ability is low or the task is burdensome, they are more likely to give merely acceptable answers, for example by selecting the same option without discrimination [25]. For students, routine mental health screening is a low-stakes task in which low-effort responding is common [26], and frequent surveys lead to survey fatigue [27]. This perspective is particularly relevant to grade differences. Senior students completed the same scale at entry and may invest less effort when it is administered again, or may under-report symptoms if they know that a screening flag triggers follow-up; conversely, students screened for the first time may show “initial elevation bias”, that is, higher reports of subjective symptoms at the first measurement and lower reports subsequently [28]. As campus screening has moved online, item presentation, response burden, progress indicators, trust in confidentiality and expectations about feedback may also affect respondents’ persistence and response quality [29, 30]. Completion time and invariant responding are therefore not only criteria for data cleaning but also indirect signals of response engagement; if engagement differs systematically between grades, part of a grade difference may reflect the assessment situation rather than distress itself.

Group comparisons also presuppose measurement invariance: GSI differences between groups can be interpreted as differences in distress only if the scale measures the same construct with the same metric and origin across sexes and grades [31]. Recent symptom-network studies have offered fine-grained views of first-year students and of students with different symptom profiles [32–34], but such analyses depend on a stable scale version, adequate item-level psychometric evidence and out-of-

sample validation. Without such evidence, it is more prudent to give priority to overall severity, adjusted group differences, measurement invariance and response-quality sensitivity.

Using a large routine screening at one vocational college covering three grades with no missing item responses, we addressed the following questions: (1) how overall distress, the nine dimensions and historical screening flags are distributed, and how they compare with Chinese adult and undergraduate norms; (2) whether SCL-90 overall severity is measurement invariant across sex and grade; (3) how sex, grade, ethnicity and college, adjusted for one another, are associated with GSI; (4) how response-engagement indicators are distributed across groups, and whether the main estimates are stable under progressively stricter response-quality rules and in models that account for the floor effect; and (5) whether, after removing each record's own mean symptom level, the demographic contrasts show interpretable relative symptom profiles. Because a previous longitudinal study suggested that sex differences may be concentrated at entry [4], we also tested the sex × grade interaction as an exploratory analysis. The response-engagement analyses are descriptive, and their implications for the organization of campus digital screening are presented as hypotheses that require experimental testing.

Methods

Study design and setting

This cross-sectional study used records from the routine student mental health screening of a higher vocational college in China; it is the first formal analysis of these data. Of the 5,872 students on the screening roster (enrolled students expected to be on campus during the assessment period), 5,602 were assessed (participation 95.40%; approximately 94%, 94% and 98% in grades 2024, 2025 and 2026, respectively); all non-participants were absent from campus during the assessment period, including students on off-campus placement internships. Third-year programmes whose students were on full-time off-campus placement internships were not on the roster (see Records and data processing), so participation rates refer to students expected on campus rather than to complete entry cohorts. Time stamps ranged from 17 to 22 September 2026 (17 September, 70.0%; 18 September, 18.9%; 19–22 September, 11.1%). The platform exported 12 workbooks of identical structure containing 5,602 records, each with responses to all 90 items. The screening covered all students on campus and took place in the college's computer laboratories on Windows 10 computers; students completed and submitted the questionnaire voluntarily under the supervision of laboratory staff. All available records were used, and no a priori sample-size calculation was performed. Reporting follows the STROBE guidelines for observational studies [35] (Additional file 2).

Ethics and consent

This study was approved by the Human Research Ethics Committee of Pingliang Vocational and Technical College (approval number: 34502501; September 2026) and was conducted in accordance

with the 1964 Declaration of Helsinki and its subsequent revisions, as well as relevant institutional guidelines. The routine screening was organized by the college's Mental Health Education Center as part of its regular work. Before answering, students read an electronic informed-consent form explaining the purpose of the screening, that results would be kept strictly confidential, used only by the center's professional staff and not provided to any organization or individual without the student's consent, and how to access free counselling; students began after clicking "I accept". The research team received only de-identified records; institutional authorization was obtained before the records were used for research, and the ethics committee waived individual informed consent for this study. Of the records, 213 belonged to students younger than 18 years (57 younger than 16 years), and informed consent was obtained from the parents or legal guardians of all of these students.

Records and data processing

All 90 items in the 5,602 records had complete integer responses from 1 to 5, with no missing or out-of-range values. In the analysis file, direct identifiers—names, platform accounts, platform user identifiers, assessment identifiers, exact dates of birth, home addresses and exact response times—had been deleted and replaced by random study numbers. Platform identifiers were all unique and no fully duplicated rows were found; however, 13 pairs of records (26 records) shared the same name, sex and date of birth but had different platform identifiers, and the 90 responses were identical in three of these pairs. These records were flagged as candidate repeat participants, retained in the main analysis and examined in a sensitivity analysis; accordingly, the number of unique students was 5,589 (participation at most 95.18%). The unit of analysis is the screening record.

Age was calculated from date of birth and assessment date; 11 values that were missing, invalid or outside 14–31 years were treated as missing (age was used only for description and a supplementary analysis). Sex was recorded as male or female. The grade field took the values 2024, 2025 and 2026, with median ages of 20.21, 19.37 and 18.79 years, consistent with the year of entry; grade was therefore interpreted as entry cohort, and grade 2026 comprised new entrants assessed a few weeks after enrolment. Grade 2024 was small (n = 279, with no records in two colleges) because students in some third-year programmes were on off-campus placement internships; grade 2026 had fewer records than grade 2025 because enrolment in 2026 was smaller after a reorganization of the college's schools. Ethnicity was collapsed into Han and ethnic minority to reduce sparse cells and protect privacy. College had seven cleaned categories; to protect confidentiality, college names are not reported, and colleges are coded College A–G in descending order of record count. One record whose original college label differed from its cleaned label (a student who had transferred programmes) was merged into College D, and excluding this record did not change any result (Additional file 1: Table S8).

Measures and scoring

We used the Chinese version of the SCL-90 [6, 7]. Respondents rate 90 symptoms for the past week from 1 (not at all), 2 (a little bit), 3 (moderately), 4 (quite a bit) to 5 (extremely); all items are positively keyed. The item text, instructions and scoring rules of the platform version were checked item by item against the scale document used for administration, and the assignment of items to the nine dimensions matched classical scoring (Additional file 1: Table S1): somatization (12 items), obsessive–compulsive (10), interpersonal sensitivity (9), depression (13), anxiety (10), hostility (6), phobic anxiety (7), paranoid ideation (6) and psychoticism (10); the remaining seven additional items (mainly on sleep and appetite) contribute to the total score and GSI but were not analysed as a dimension. The total score is the sum of the 90 items (90–450); GSI is the mean of the 90 items; the positive symptom total (PST) is the number of items scored ≥ 2; and the positive symptom distress index (PSDI) is the mean of the items scored ≥ 2 (missing when no item is positive). Because item wording differs between Chinese versions [36], results refer to the platform version used. For descriptive comparison we used the Chinese adult norms [11] and, in addition, the Chinese undergraduate norms [13]; because the latter are reported separately for urban and rural students, we pooled them by group size (2,236 and 2,123 students) to obtain an overall mean and standard deviation. The scale is intended for respondents aged 16 years or older; 57 records were from younger students, and excluding them did not change the main results (see Results).

The historical screening rules were a total score ≥ 160, PST ≥ 43 and any dimension mean ≥ 2; we also recomputed each rule with a strict inequality (>). A union flag indicates that a record met at least one rule. All thresholds were used only to show the sensitivity of historical scoring conventions; they were not used to diagnose mental disorders or to judge individual clinical need.

Response-quality indicators

Completion time was recorded in seconds, and records completed in less than 60, 90 and 120 s were flagged; we also flagged records with identical responses to all 90 items, with all 90 items answered 1, and with a within-person item standard deviation ≤ 0.20. These indicators were regarded as imperfect heuristics and were not used to declare any single record invalid. The main analysis retained all structurally complete records, because uniformly minimal responses may reflect a genuine symptom “floor” as well as low-effort responding. Five sensitivity samples were specified by the analysts before results were inspected (not preregistered), excluding in turn: (1) < 60 s; (2) < 120 s; (3) invariant responding; (4) < 60 s or invariant responding; and (5) < 120 s or invariant responding (Fig. 1). In addition, two supplementary samples were defined using the common reference of about 2 s per item [21]: (6) excluding < 180 s; and (7) excluding < 180 s or invariant responding. These two samples were defined after preliminary results had been inspected, serve only as supplementary references and are reported separately from the five prespecified samples. To describe the context of response engagement, we also computed the proportion of invariant responding in each completion-time band and the intraclass correlation coefficients (ICCs) of invariant responding and log completion time across classes.

Statistical analysis

Continuous variables are reported as mean, standard deviation (SD), median and interquartile range (IQR). Internal consistency was estimated with Cronbach's α; because a high α does not imply unidimensionality or discriminant validity, we also report the correlations between dimension means and the eigenvalues of the 90-item correlation matrix as descriptive evidence of the strength of the general severity component. Sample scores were compared with the norms using one-sample t tests and the standardized difference d = (sample mean − norm mean)/norm SD.

Measurement invariance was examined at two levels. First, a one-factor model with the nine dimension means as indicators was estimated by maximum likelihood in multi-group confirmatory factor analyses, testing in turn configural, metric (equal loadings) and scalar (equal loadings and intercepts) invariance across sex (male/female), grade (2025/2026/2024) and grade (2025/2026), in all records and again after excluding invariant responding. Invariance was supported when ΔCFI was not below − 0.010 and ΔRMSEA not above 0.015, with ΔSRMR not above 0.030 for metric and not above 0.010 for scalar invariance [31, 37]. Because dimension means are highly reliable indicators, RMSEA may be inflated in such models [38], so decisions relied mainly on the changes between nested models. Because dimension means act as item parcels, which can mask non-invariance at the item level [39], these models were complemented by item-level analyses. Second, item-level invariance was examined with logistic-regression differential item functioning (DIF) analysis [40]. With item endorsement (score ≥ 2) as the outcome and the total score excluding the studied item as the matching variable, we compared Nagelkerke R^2 before and after adding the group main effect and the group × matching-variable interaction (ΔR^2); following Jodoin and Gierl [41], $\Delta R^2 < 0.035$ was regarded as negligible, 0.035–0.070 as moderate and > 0.070 as large. The DIF analysis excluded records with invariant responding, which carry no item-level discriminating information.

The main analysis used ordinary least squares regression of GSI on sex, grade, ethnicity and college simultaneously. Reference groups were male, grade 2025, Han and College A (the largest college). A composite key of college, major, grade and class defined the clusters, and cluster-robust standard errors with the CR1 small-sample correction were computed, with inference based on t and F distributions with G − 1 degrees of freedom [42]. The main model included 5,599 records (3 had missing ethnicity) and 203 clusters. We report unstandardized coefficients (B), CR1 standard errors, 95% confidence intervals (CIs) and two-sided p values; effect sizes were expressed as B divided by the full-sample SD of GSI (a Cohen's d-type index) and, for predictor blocks, as partial η^2. The overall association of college was tested with a 6-df cluster-robust Wald test, and individual college coefficients were not used for ranking. Clustering by class was described with the one-way analysis-of-variance ICC. To check the dependence of inference on the variance estimator, the main coefficients were re-estimated with a cluster bootstrap (1,999 resamples of classes) and with HC3 heteroskedasticity-consistent standard errors.

The same model was then fitted to GSI and to each of the nine dimension means, and the false discovery rate across the 40 global tests (10 outcomes × 4 predictor blocks) was controlled with the Benjamini–Hochberg (BH) procedure [43]. To examine relative profiles after removing each person's overall symptom level, we subtracted each record's mean of the nine dimension means from each

dimension score, fitted the same model to each centred dimension, and applied BH correction to the nine p values within each demographic contrast. A positive centred coefficient means only that the dimension is relatively more prominent within the record's own profile, not that its absolute score is higher.

The main model was re-estimated in the five response-quality sensitivity samples, and all results are reported regardless of statistical significance. Additional robustness and exploratory analyses were: (1) the sex × grade interaction (2-df Wald test), with within-grade sex differences estimated from grade-stratified models and repeated in the six analysis samples; (2) ln(GSI) as the outcome; (3) a two-part model, in which part 1 was a CR1 cluster-robust logistic regression of the probability that any item was > 1 (i.e., not all items at the minimum) and part 2 estimated distress intensity among records with GSI > 1; (4) adjustment for assessment date (17, 18 and 19–22 September); (5) exclusion of the second record of each candidate repeat pair; (6) exclusion of records of students younger than 16 years; and (7) adjustment for log completion time. Completion time was moderately correlated with GSI (Spearman $\rho = 0.57$), probably partly as a consequence of symptom endorsement, so the time-adjusted model may over-adjust and serves only as the most conservative reference. The study was not preregistered; the main GSI model is the primary analysis, and all other analyses are secondary or exploratory. All scoring and models of the main and prespecified sensitivity analyses were implemented with matrix formulas in Python 3.13 (NumPy, SciPy, pandas) and checked with an independently written second code base, with identical results; the code is provided in Additional file 3.

After preliminary results had been inspected, the following supplementary analyses were added and are interpreted as post hoc: (8) Tobit regression [44] of ln(GSI), left-censored at 0 (corresponding to all 90 items answered 1), to estimate differences in latent distress that may exist below the "floor", with inference based on a cluster-robust sandwich variance; (9) a rank-transformed model with the percentile rank of GSI as the outcome; (10) a linear mixed model with a random intercept for class, estimated by restricted maximum likelihood (REML); (11) re-estimation of the main model and the sex × grade interaction after excluding grade 2024; (12) recomputation of GSI after excluding the four items with the most pronounced sex DIF, followed by re-estimation of the models; (13) ordinal (proportional-odds) logistic-regression DIF of the five-category item responses, with the same matching variable, nested models and analysis sample as the dichotomous analysis, flagging items with a change in McFadden pseudo-$R^2 \geq 0.02$, the criterion used in the lordif procedure [45]; and (14) SRMR including standardized mean residuals at the scalar step of the invariance models, because SRMR based on covariance residuals alone cannot reflect intercept constraints. Because whether an interaction is present may depend on the scale of the outcome [46], we also report within-grade sex differences standardized by the pooled within-grade SD, and the probability of superiority (the probability that a randomly drawn female record has a higher GSI than a randomly drawn male record), which is unaffected by monotonic transformations of the scale. The measurement-invariance, DIF and supplementary analyses were also implemented in Python 3.13, with the Tobit, mixed and multi-group confirmatory factor models programmed directly from their likelihoods; code and summary outputs are provided in Additional file 3.

Analyses (13) and (14) use only variables included in the anonymised dataset (Additional file 4) and can therefore be reproduced exactly from it.

Results

Sample characteristics and response engagement

Sample characteristics are shown in Table 1. Female students contributed 3,124 records (55.77%) and male students 2,478 (44.23%). For the 5,591 records with valid age, the mean age was 19.27 years (SD 1.05; median 19.14, IQR 18.67–19.82); 213 records (3.81%) were from students younger than 18 years and 57 from students younger than 16. Grades 2025, 2026 and 2024 accounted for 64.96%, 30.06% and 4.98% of records; 82.63% were Han, 17.32% ethnic minority, and ethnicity was missing for 3 records. The median size of the 203 class clusters was 33 records (range 1–79).

Table 1
Characteristics of screening records and response-quality indicators (N = 5,602)

Characteristic	n	%
Sex: female	3,124	55.77
Sex: male	2,478	44.23
Grade: 2024	279	4.98
Grade: 2025	3,639	64.96
Grade: 2026 (new entrants)	1,684	30.06
Ethnicity: Han	4,629	82.63
Ethnicity: ethnic minority	970	17.32
Ethnicity: missing	3	0.05
College A	1,631	29.11
College B	1,053	18.80
College C	969	17.30
College D	626	11.17
College E	488	8.71
College F	420	7.50
College G	415	7.41
Age < 18 years	213	3.81
Completion time < 60 s	92	1.64
Completion time < 90 s	533	9.51
Completion time < 120 s	1,353	24.15
Identical responses to all 90 items	1,429	25.51
All 90 items answered 1	1,428	25.49
Within-person item SD ≤ 0.20	2,255	40.25

Percentages use N = 5,602 except , which uses the 5,591 records with valid age. Age M = 19.27 years, SD = 1.05; median age by grade: 2024, 20.21; 2025, 19.37; 2026, 18.79 years. Colleges are coded A–G in descending order of record count; one record whose original college label differed from its cleaned label (a transfer student) was merged into College D.

The median completion time was 195.8 s (IQR 122.0–320.2). Ninety-two records (1.64%) were completed in less than 60 s and 1,353 (24.15%) in less than 120 s. A total of 1,429 records (25.51%) had identical responses to all 90 items, 1,428 (25.49%) of which were all 1. Response quality was unevenly distributed across groups (Additional file 1: Table S2): invariant responding accounted for 50.18%, 29.43% and 12.95% of records in grades 2024, 2025 and 2026 ($\chi^2(2) = 258.71$, $p < .001$), with median completion times of 138.0, 170.0 and 288.2 s; invariant responding was more common among male (29.02%) than female (22.73%) records and among Han (26.42%) than ethnic minority (21.13%) records. These differences do not show that any single record is invalid, but they indicate that cleaning rules may affect groups differently. Invariant responding coincided largely with fast completion: among records completed in < 60 s, 60–119 s, 120–179 s and ≥ 180 s, 59.8%, 50.6%, 35.7% and 10.0% were invariant, and within each grade the median completion time of invariant records was 119–132 s (about 1.3–1.5 s per item), much shorter than that of non-invariant records (175–311 s). Invariant responding also clustered markedly by class: among the 146 classes with ≥ 10 records, the class proportion ranged from 0% to 72.5% (median 24.5%); the class-level ICC of invariant responding was 0.092, and 0.036 after adjustment for sex, grade, ethnicity and college; the ICC of log completion time was 0.177 (Additional file 1: Table S15).

Distress distributions, score structure, measurement invariance and comparison with norms

Mean GSI was 1.261 (SD 0.436) and the median 1.067 (IQR 1.000–1.322); the distribution showed a marked floor effect and a long right tail (skewness 2.72, excess kurtosis 8.60). The mean total score was 113.51 (SD 39.28; median 96, IQR 90–119). Among the nine dimensions, obsessive–compulsive had the highest mean (1.401) and psychoticism the lowest (1.183); 39.2%–62.7% of records were at the minimum on each dimension (Table 2). GSI distributions differed clearly between grades, and grade 2026 had the highest scores on all nine dimensions (Fig. 2).

Compared with the Chinese adult norms, sample GSI was lower (1.261 vs 1.44; $d = -0.42$; $t(5601) = -30.66$, $p < .001$), as were PST (15.80 vs 24.92; $d = -0.50$) and PSDI (2.29 vs 2.60; $d = -0.52$); eight dimensions were below the norms ($d = -0.73$ to -0.25), and phobic anxiety did not differ significantly ($d = 0.01$, $p = .33$) (Table 2; Additional file 1: Table S11). After excluding invariant records, GSI remained below the norm ($d = -0.21$) but phobic anxiety was above it ($d = 0.21$), showing that the comparison with norms also depends on how invariant responding is handled. Compared with the Chinese undergraduate norms [13] (urban and rural pooled, n = 4,359), sample GSI was likewise lower (1.261 vs 1.494; $d = -0.48$), all nine dimensions were below these norms ($d = -0.58$ to -0.25), and after excluding invariant records the GSI difference narrowed to $d = -0.30$ (Additional file 1: Table S11).

Table 2
Distribution, internal consistency and comparison with Chinese adult norms of the SCL-90 dimensions and the Global Severity Index (N = 5,602)

Scale	Items	M	SD	Median	IQR	At minimum, %	α	Norm M ± SD	d
Somatization	12	1.235	0.443	1.000	1.00–1.25	52.5	.924	1.37 ± 0.48	−0.28
Obsessive–compulsive	10	1.401	0.561	1.200	1.00–1.60	39.2	.909	1.62 ± 0.58	−0.38
Interpersonal sensitivity	9	1.278	0.496	1.000	1.00–1.33	53.3	.902	1.65 ± 0.51	−0.73
Depression	13	1.282	0.511	1.077	1.00–1.31	48.8	.939	1.50 ± 0.59	−0.37
Anxiety	10	1.246	0.466	1.000	1.00–1.30	52.5	.919	1.39 ± 0.43	−0.34
Hostility	6	1.232	0.442	1.000	1.00–1.33	58.0	.838	1.48 ± 0.56	−0.44
Phobic anxiety	7	1.236	0.473	1.000	1.00–1.29	62.4	.875	1.23 ± 0.41	0.01
Paranoid ideation	6	1.218	0.429	1.000	1.00–1.33	62.7	.839	1.43 ± 0.57	−0.37
Psychoticism	10	1.183	0.397	1.000	1.00–1.20	62.3	.897	1.29 ± 0.42	−0.25
GSI (mean of 90 items)	90	1.261	0.436	1.067	1.00–1.32	25.5	.987	1.44 ± 0.43	−0.42

“At minimum, %” is the percentage of records whose dimension mean equals 1. Norms are the Chinese adult norms [11]; d = (sample mean − norm mean)/norm SD. All differences from the norms have $p < .001$ (one-sample t test; the same conclusions with class-cluster-robust tests) except phobic anxiety ($p = .33$). PST: 15.80 ± 21.85 (norm 24.92 ± 18.41, d = − 0.50); PSDI: 2.29 ± 0.40 (n = 4,174; norm 2.60 ± 0.59, d = − 0.52). α is not evidence of unidimensionality or discriminant validity. Comparisons with the Chinese undergraduate norms are given in Additional file 1: Table S11. GSI, Global Severity Index; IQR, interquartile range; PSDI, Positive Symptom Distress Index; PST, Positive Symptom Total.

Cronbach's α was .987 for the 90 items and .838–.939 for the nine dimensions. Correlations between dimension means ranged from 0.75 to 0.91 (mean 0.84). The first eigenvalue of the 90-item correlation matrix was 43.89, explaining 48.8% of the total variance and 15.0 times the second eigenvalue (2.92); after excluding invariant records, the proportion was still 46.0% (ratio of the first to the second eigenvalue 14.3) (Additional file 1: Table S10). These results indicate a dominant general severity component and support GSI as the primary outcome.

Measurement-invariance results are given in Additional file 1: Tables S12 and S13. The configural one-factor model of the nine dimension means had CFI 0.970–0.975 and SRMR 0.014–0.029 across groupings, but RMSEA 0.108–0.117, suggesting some local residual correlation. Across sex, metric and scalar invariance were supported (ΔCFI − 0.003 and − 0.002; ΔRMSEA negative; metric ΔSRMR 0.024); in the scalar model the latent mean of female students was 0.22 higher than that of male students (in units of the male latent SD). The same held for grades 2025 and 2026 (ΔCFI − 0.001 at both steps; metric ΔSRMR 0.011), with the latent mean of grade 2026 0.62 higher than that of grade 2025. When all three grades were compared, ΔCFI and ΔRMSEA met the criteria but the metric ΔSRMR was 0.061, above 0.030, mainly because of the very small latent variance in grade 2024 (about 13% of that in grade 2025); comparisons involving grade 2024 should therefore be interpreted more cautiously. Results were the same after excluding invariant records, and when standardized mean residuals were included in SRMR, its change at the scalar step was ≤ 0.003 in every comparison. At item level, only item 20 (crying easily) showed large sex DIF ($\Delta R^2 = 0.073$; more often endorsed by female students at the same overall level), and the remaining 89 items showed negligible DIF ($\Delta R^2 \leq 0.025$); in the comparisons of grades 2026 and 2024 with grade 2025, ΔR^2 was < 0.035 for all 90 items (maximum 0.029 and 0.006, respectively). At the same overall distress level, grade 2026 students more often endorsed lower back pain (item 27) and muscle soreness (item 42), although these effects were negligible. The ordinal DIF analysis gave the same picture: only item 20 in the sex comparison (change in McFadden $R^2 = 0.043$) and item 42 in the grade 2026 vs 2025 comparison (0.022) reached the 0.02 criterion, all other items had changes below 0.016, and no item approached it in the grade 2024 vs 2025 comparison (maximum 0.005) (Additional file 1: Table S13). Recomputing GSI without items 27 and 42 changed the grade 2026 vs 2025 coefficient by less than 0.01 (point estimates in the anonymised dataset).

Historical screening flags

Single-rule flag rates ranged from 9.91% for a total score > 160 to 18.69% for any dimension mean ≥ 2 (Table 3). With the three rules combined, 19.58% of records were flagged when all operators included equality and 17.67% with strict inequalities; changing the operator alone changed the union flag rate by 1.91 percentage points (107 records). By group, union flag rates (≥) were 3.23%, 14.70% and 32.84% in grades 2024, 2025 and 2026, and 22.86% and 15.46% among female and male records (Additional file 1: Table S3). None of these proportions is a prevalence of mental disorder.

Table 3
Sensitivity of historical SCL-90 screening flags to the threshold operator (N = 5,602)

Rule	Operator	n	%
Total score	≥ 160	567	10.12
Total score	> 160	555	9.91
Positive Symptom Total	≥ 43	724	12.92
Positive Symptom Total	> 43	690	12.32
Any dimension mean	≥ 2	1,047	18.69
Any dimension mean	> 2	911	16.26
Union of the three rules	All ≥	1,097	19.58
Union of the three rules	All >	990	17.67

Positive items are items scored ≥ 2. The values describe historical screening flags only and should not be called diagnoses, detection rates, prevalence or the proportion of students needing treatment.

Adjusted group differences in overall distress

The main model included 5,599 records and 203 class clusters (R^2 = 0.085; Table 4). Grade was the strongest group signal ($F(2, 202) = 84.59$, $p < .001$, partial $\eta^2 = 0.070$): compared with grade 2025, grade 2026 had higher GSI (B = 0.236, SE = 0.021, 95% CI [0.195, 0.277], $p < .001$, d = 0.54) and grade 2024 lower GSI (B = − 0.142, SE = 0.028, 95% CI [− 0.198, − 0.086], $p < .001$, d = − 0.33). Female records had slightly higher GSI than male records (B = 0.061, SE = 0.018, 95% CI [0.026, 0.095], $p < .001$, d = 0.14, partial $\eta^2 = 0.003$). The ethnicity difference was very small and non-significant (B = − 0.007, SE = 0.014, 95% CI [− 0.035, 0.021], p = .604). College was associated with GSI overall ($F(6, 202) = 5.99$, $p < .001$, BH $q < .001$, partial $\eta^2 = 0.010$), but college membership combines differences in programme composition, sex ratio, grade distribution and assessment timing; the college coefficients in Table 4 are reported for transparency only and are not used for ranking or causal attribution. The ICC of GSI between classes was 0.109, and the residual ICC after adjustment was 0.029, supporting cluster-robust inference. Cluster-bootstrap and HC3 standard errors did not change any of these conclusions (Additional file 1: Table S8).

Table 4
Cluster-robust adjusted model of the Global Severity Index (n = 5,599; 203 clusters)

Contrast	B	CR1 SE	95% CI	p	d
Female vs male	0.061	0.018	[0.026, 0.095]	< .001	0.14
Grade 2024 vs 2025	−0.142	0.028	[− 0.198, − 0.086]	< .001	−0.33
Grade 2026 vs 2025	0.236	0.021	[0.195, 0.277]	< .001	0.54
Ethnic minority vs Han	−0.007	0.014	[− 0.035, 0.021]	.604	−0.02
College B vs College A	−0.027	0.025	[− 0.077, 0.024]	.298	−0.06
College C vs College A	−0.032	0.025	[− 0.082, 0.018]	.205	−0.07
College D vs College A	0.064	0.027	[0.011, 0.118]	.019	0.15
College E vs College A	−0.041	0.034	[− 0.108, 0.027]	.236	−0.09
College F vs College A	0.005	0.027	[− 0.048, 0.058]	.855	0.01
College G vs College A	−0.128	0.032	[− 0.192, − 0.064]	< .001	−0.29

The model includes sex, grade, ethnicity and college simultaneously; reference groups are male, grade 2025, Han and College A. Intercept = 1.181. Inference uses t distributions with 202 df. d = B/SD(GSI), SD = 0.436. Global tests of predictor blocks: sex $F(1, 202) = 11.95$; grade $F(2, 202) = 84.59$; ethnicity $F(1, 202) = 0.27$, $p = .604$; college $F(6, 202) = 5.99$, $p = 8.8 \times 10^{-6}$ (BH q across 40 global tests = 2.2×10^{-5}). p values for individual colleges are not further corrected for pairwise comparisons. BH, Benjamini–Hochberg; CR1, cluster-robust standard error with small-sample correction.

Sex differences moderated by grade (exploratory analysis)

When the sex × grade interaction was added to the main model, it was significant overall ($F(2, 202) = 18.02$, $p < .001$; Fig. 3). In grade-stratified models, the adjusted sex difference was B = 0.178 (95% CI [0.102, 0.254], $p < .001$) in grade 2026, B = 0.015 ([− 0.024, 0.054], $p = .44$) in grade 2025 and B = − 0.002 ([− 0.035, 0.030], $p = .88$) in grade 2024. Raw cell means showed the same pattern: female and male GSI were 1.512 and 1.324 in grade 2026, 1.220 and 1.173 in grade 2025, and 1.063 and 1.064 in grade 2024. In other words, the female excess in the main model was driven mainly by new entrants. The within-grade sex differences implied by the pooled interaction model were in the same direction (grade 2026, 0.161; grade 2025, 0.027; grade 2024, − 0.054); the negative value for grade 2024 had a CR1 95% CI of [− 0.094, − 0.013], but its cluster-bootstrap interval ([− 0.101, 0.001]) included 0 and it was close to 0 in the grade-stratified model, so it cannot be regarded as a reliable reversed difference.

In the other five response-quality sensitivity samples, the interaction was significant in every case ($F = 11.73–17.70$, all $p < .001$); the sex difference within grade 2026 was 0.162–0.175, whereas the intervals for grade 2025 (estimates 0.001–0.012) and grade 2024 (− 0.056 to − 0.003) all included 0 (Additional file 1: Table S9). After adding log completion time, a potentially over-adjusting covariate, the interaction

remained significant (F(2, 202) = 12.57, p < .001), with a sex difference within grade 2026 of B = 0.106 (95% CI [0.040, 0.172]). In part 1 of the two-part model (endorsement of any symptom), the sex × grade interaction was not significant (F(2, 202) = 0.50, p = .61), indicating that the interaction lies mainly in distress intensity. Grade 2024 had only 22 class clusters, which limits the precision of its estimates.

To test whether the interaction was merely a scale artefact of the floor effect, we ran several supplementary checks (Additional file 1: Table S14). The interaction was significant with ln(GSI) as the outcome (F(2, 202) = 16.03, p < .001) and when ln(GSI) was analysed only among records endorsing at least one symptom (GSI > 1; F(2, 199) = 12.60, p < .001). In the Tobit model, which treats all-1 responding as censored, the interaction remained significant (F(2, 202) = 5.23, p = .006); the sex difference in latent ln(GSI) was 0.108 (95% CI [0.061, 0.156]) in grade 2026, 0.035 ([0.000, 0.071]) in grade 2025 and − 0.009 ([− 0.072, 0.055]) in grade 2024. With the percentile rank of GSI as the outcome, the interaction was weaker but still significant (F(2, 202) = 3.55, p = .030). Standardized by the pooled within-grade SD, the sex differences were d = 0.35, 0.13 and − 0.005 in grades 2026, 2025 and 2024; the probability that a female record had a higher GSI than a male record was 0.61 (p < .001), 0.55 (p < .001) and 0.54 (p = .22), respectively. The random-intercept mixed model (interaction Wald $\chi^2(2) = 29.00$, p < .001; within grade 2026, B = 0.168, 95% CI [0.118, 0.217]), exclusion of grade 2024 (interaction B = 0.134, 95% CI [0.063, 0.206]) and recomputation of GSI without the sex-DIF items (F(2, 202) = 17.13, p < .001) gave consistent results. Overall, the sex difference among new entrants was not simply produced by the floor effect, although the extent of its amplification relative to senior grades depended on the scale used.

Dimension-level and relative symptom profiles

After BH correction across the 40 global tests, the overall associations of grade and college with all nine absolute dimension means were supported; sex was associated with seven dimensions (not paranoid ideation or psychoticism) and ethnicity only with psychoticism (Additional file 1: Tables S4 and S5). Group patterns for the absolute dimensions closely matched those for GSI, consistent with the dominant general severity component, and should not be regarded as nine independent findings.

The within-person centred analysis narrowed the question to relative symptom profiles (Fig. 4; Additional file 1: Table S6). Compared with male records, female records showed relatively more prominent phobic anxiety (B = 0.035), depression (B = 0.023), obsessive–compulsive symptoms (B = 0.021) and anxiety (B = 0.012), and relatively less prominent psychoticism (B = − 0.035), paranoid ideation (B = − 0.032) and hostility (B = − 0.021). Compared with grade 2025, grade 2026 showed relatively more prominent obsessive–compulsive symptoms (B = 0.077), interpersonal sensitivity (B = 0.022), depression (B = 0.022) and anxiety (B = 0.010), and relatively less prominent psychoticism (B = − 0.057), paranoid ideation (B = − 0.044) and hostility (B = − 0.041); grade 2024 showed broadly the opposite pattern (obsessive–compulsive B = − 0.053, interpersonal sensitivity B = − 0.015, paranoid ideation B = 0.018, psychoticism B = 0.025). Ethnic minority records showed relatively more prominent somatization (B = 0.034) and relatively less prominent psychoticism (B = − 0.019) and hostility (B = − 0.015). All centred coefficients were small ($|B| \leq 0.077$ on the 1–5 scale) and indicate only deviations from each record's own mean level. Because the nine centred scores sum to zero, the coefficients are not independent;

together with the low specific reliability of the dimensions reported in a previous bifactor analysis, this means that the relative profiles can serve only as descriptive clues.

Response-quality sensitivity and other robustness checks

Table 5 and Fig. 5 show the estimates of the main contrasts in the six prespecified analysis samples (full results in Additional file 1: Table S7). The sex contrast and both grade contrasts had the same direction in every sample, with 95% CIs excluding 0: 0.054 to 0.067 for female vs male, − 0.168 to − 0.142 for grade 2024 and 0.215 to 0.237 for grade 2026. As the quality rules became stricter, the ethnicity coefficient was always negative and its absolute value generally increased (− 0.007 to − 0.050), with a CI excluding 0 only in the strictest and smallest analysis set, showing that the size and significance of the ethnicity estimate are sensitive to sample selection. In the two supplementary samples based on 2 s per item (Table 5), the main contrasts kept their direction: after excluding records completed in < 180 s (n = 3,043), female B = 0.057 (95% CI [0.008, 0.105]), grade 2024 B = − 0.151 and grade 2026 B = 0.221, with a sex × grade interaction $F(2, 187) = 8.06$ ($p < .001$); after also excluding invariant responding (n = 2,738), the corresponding estimates were 0.053, − 0.156 and 0.207, with $F(2, 184) = 6.71$ ($p = .002$). The ethnicity coefficients in these two samples were − 0.047 and − 0.067, with CIs excluding 0, further showing that this estimate depends on sample selection.

Table 5
Main GSI contrasts in the response-quality sensitivity samples: B [95% CI]

Analysis sample	n/clusters	Female vs male	2024 vs 2025	2026 vs 2025	Ethnic minority vs Han
Primary (all records)	5,599/203	0.061 [0.026, 0.095]	−0.142 [−0.198, −0.086]	0.236 [0.195, 0.277]	−0.007 [−0.035, 0.021]
Excluding < 60 s	5,507/203	0.057 [0.022, 0.092]	−0.144 [−0.201, −0.088]	0.236 [0.195, 0.277]	−0.011 [−0.039, 0.018]
Excluding < 120 s	4,247/199	0.059 [0.020, 0.098]	−0.153 [−0.215, −0.090]	0.237 [0.194, 0.280]	−0.031 [−0.065, 0.003]
Excluding invariant responding	4,171/200	0.056 [0.014, 0.097]	−0.167 [−0.227, −0.107]	0.216 [0.174, 0.257]	−0.028 [−0.062, 0.005]
≥ 60 s and non-invariant	4,134/200	0.054 [0.013, 0.095]	−0.168 [−0.228, −0.108]	0.215 [0.174, 0.256]	−0.032 [−0.066, 0.002]
≥ 120 s and non-invariant	3,512/198	0.067 [0.024, 0.110]	−0.167 [−0.235, −0.100]	0.220 [0.177, 0.263]	−0.050 [−0.088, −0.012]
Supplementary: excluding < 180 s	3,043/188	0.057 [0.008, 0.105]	−0.151 [−0.224, −0.077]	0.221 [0.175, 0.268]	−0.047 [−0.091, −0.003]
Supplementary: ≥ 180 s and non-invariant	2,738/185	0.053 [0.002, 0.105]	−0.156 [−0.233, −0.078]	0.207 [0.159, 0.254]	−0.067 [−0.112, −0.021]

Each model adjusts for sex, grade, ethnicity and college, with CR1 cluster-robust standard errors; n excludes the 3 records with missing ethnicity. The last two rows are supplementary samples based on 2 s per item that were defined after preliminary results had been inspected; they are not shown in Fig. 5. GSI, Global Severity Index.

The two-part model explains this selection sensitivity (Additional file 1: Table S8). Ethnic minority records had higher odds of endorsing any symptom (i.e., not all items at the minimum; OR = 1.31, 95% CI [1.11, 1.54]), but among records endorsing symptoms the two groups did not differ significantly in distress intensity (B = − 0.028, 95% CI [− 0.061, 0.006]). Excluding invariant records therefore removed more low-scoring records from the Han group, shifting the ethnicity contrast in the negative direction. By contrast, the grade and sex contrasts had the same direction in both parts: grade 2026 had about three times the odds of endorsing any symptom compared with grade 2025 (OR = 3.00) and higher intensity among records endorsing symptoms (B = 0.216).

Other checks were consistent. With ln(GSI) as the outcome, after adjustment for assessment date, and after excluding candidate repeat records, the one record with a merged college label or the 57 records of students younger than 16 years (female B = 0.060, grade 2024 B = − 0.143, grade 2026 B = 0.238; interaction F(2, 202) = 17.78), the direction and significance of the sex and grade estimates were unchanged (Additional file 1: Table S8). After adding log completion time, the differences for grade 2026 (B = 0.140) and grade 2024 (B = − 0.108) persisted but were smaller; the overall sex difference became non-significant (B = 0.017, 95% CI [− 0.015, 0.049]), but the sex difference within grade 2026 persisted (see above). The random-intercept mixed model gave almost identical estimates (female B = 0.062, grade 2024 B = − 0.133, grade 2026 B = 0.239, ethnic minority B = − 0.008; conditional ICC = 0.034). In the Tobit model, the sex and grade main effects had the same direction as in the main model (log scale: female 0.055, grade 2024 − 0.175, grade 2026 0.199), and ethnicity was not significant (0.012, 95% CI [− 0.009, 0.034]). With the percentile rank of GSI as the outcome, ethnic minority records had slightly higher ranks (B = 0.026, 95% CI [0.008, 0.044]), consistent with the finding of the two-part model that ethnic minority records more often endorsed any symptom.

Discussion

Based on 5,602 routine screening records from one vocational college, this study yielded five main findings. First, overall distress showed a marked floor effect and right skew, its level was below both Chinese adult and undergraduate norms, about one quarter of records gave the lowest response to every item, and historical screening flag rates changed by almost 2 percentage points with the inequality operator. Second, measurement-invariance tests supported comparing GSI across sex and between grades 2025 and 2026, whereas the invariance evidence for comparisons involving grade 2024 was weaker. Third, grade was the strongest adjusted group signal: new entrants (grade 2026), assessed a few weeks after enrolment, had GSI about 0.24 points higher than grade 2025 (d ≈ 0.54), and grade 2024 about 0.14 points lower. Fourth, the slightly higher overall distress of female students was concentrated in grade 2026 and was close to 0 in the senior grades, a pattern that persisted in censored and mixed models. Fifth, response engagement declined markedly with grade and clustered by class; the grade differences and the sex difference among new entrants were stable across response-quality rules, whereas the ethnicity difference was unstable and its changes could be explained by between-group differences in invariant responding.

Highest distress among new entrants

New entrants in grade 2026 had the highest distress, higher than other grades on all nine dimensions. This pattern is consistent with an entry-adjustment explanation: in the weeks before the assessment, new students had left home, moved into collective accommodation, completed military training and faced new academic demands and the rebuilding of their social networks, all of which may temporarily increase psychological load. Longitudinal evidence supports this explanation: a follow-up study of vocational college students found that the SCL-90 total score fell from 134.7 at entry to 122.1 two years later and the proportion with a total score > 160 from 23% to 17% [4], and a Chinese university freshman

cohort showed that mental health problems at entry were associated with later help-seeking [19]. The relative-profile analysis showed that obsessive–compulsive symptoms, interpersonal sensitivity, depression and anxiety were more prominent among new entrants, consistent with the worry, self-doubt and interpersonal tension typical of the adjustment period. Item-level results also fit this context: at the same overall distress level, new entrants more often endorsed lower back pain and muscle soreness, which may reflect the physical load of military training at entry, although these effects were very small.

Cross-sectional data, however, cannot attribute this difference to the stage of study itself. The three grades were observed in the same assessment window but differed in entry cohort, age, programme composition, internship arrangements and screening experience. At least three competing explanations related to the assessment situation deserve attention. First, senior students had completed the same scale at entry, and repeated administration may have reduced their motivation and made satisficing more likely [25, 27]; consistent with this, senior grades had more invariant responding and faster completion (about half of grade 2024 records were all 1), and invariant records took only about 1.3–1.5 s per item. Second, if senior students knew that a screening flag would trigger a follow-up assessment or interview, they may have deliberately under-reported symptoms, and such strategic responding cannot be detected with time or invariance indicators. Third, students screened for the first time may show initial elevation bias [28], which would inflate their scores. In addition, most grade 2024 students were away on placement internships, so the participants on campus may be a selected group, and the metric-invariance evidence for this grade was weaker. Although the difference between grades 2026 and 2025 persisted after excluding fast and invariant records, in censored models and after excluding grade 2024, and was only partly reduced by adjustment for completion time, these mechanisms may still have amplified the observed grade gradient. The more cautious conclusion is therefore that, in this screening, new entrants were the group with the highest overall distress and the greatest need for timely re-assessment; whether distress declines with the stage of study needs to be tested with repeated measurements of the same students, together with external indicators that do not depend on engagement with self-report.

Sex differences concentrated among new entrants

The most notable finding of this study is that the sex difference was moderated by grade. In the additive model, female GSI was only about 0.06 points higher ($d \approx 0.14$), a result that might be summarized as a “small, stable female excess”; the interaction analysis showed that this average masks marked heterogeneity: female new entrants scored about 0.18 points higher than male new entrants, whereas there was almost no difference in grades 2025 and 2024. This pattern closely matches the vocational college follow-up study, in which female students had higher distress at entry and the sex difference had disappeared two years later [4]. Longitudinal studies of university students also suggest that sex differences in depression and anxiety change over time [15]. Possible explanations include greater sensitivity of female new entrants to interpersonal and safety pressures in a new environment and a greater willingness to report emotional symptoms, and distress from entry adjustment that is more pronounced among female students but subsides as adjustment is completed. Another possibility is a

difference in engagement: among new entrants, invariant responding was less common among female than male students (9.8% vs 16.7%) and the median completion time was longer (303.3 vs 265.6 s). However, the sex difference among new entrants persisted after controlling for completion time (B ≈ 0.11), so it cannot be attributed entirely to response engagement.

The supplementary analyses showed that the interaction was not simply produced by the floor effect: it remained significant in the Tobit model treating all-1 responding as censored, in the intensity model restricted to records endorsing symptoms and in the mixed model. In the rank-transformed model, however, the interaction was clearly weaker, indicating that the extent to which the sex difference grows or shrinks across grades depends partly on the scale used [46]. At item level, crying easily was endorsed more often by female students at the same overall level in both the dichotomous and the ordinal DIF analyses; excluding the items with the most pronounced sex DIF did not change the interaction, showing that the sex difference among new entrants is not driven by a few items.

The relative profiles showed more prominent phobic anxiety, depression, obsessive–compulsive symptoms and anxiety among female records, and more prominent hostility, paranoid ideation and psychoticism among male records. These coefficients amount to only a few hundredths of a point on the 1–5 scale and should be regarded as configural clues requiring replication, not reified as fixed “sex-specific symptom types”. Because the sex × grade interaction was an exploratory analysis that was not preregistered, its robustness needs to be tested in independent samples.

Selection sensitivity of the ethnicity estimate

The main model did not support an overall difference between ethnic groups, but the size and significance of the estimate changed with the cleaning rule. The two-part model showed that this was because the proportion of all-minimum responding differed between the groups: ethnic minority records less often showed invariant responding, whereas among records endorsing symptoms the two groups had similar distress intensity. Excluding invariant records therefore removed a higher proportion of low-scoring records from the Han group, making the ethnicity difference appear “more negative”. This result has general methodological relevance: when response style or engagement is unevenly distributed across groups, seemingly neutral data cleaning can create group differences [24]. In the rank-transformed model, ethnic minority records even had slightly higher ranks, again showing that the direction of this contrast depends on how the many all-minimum records are handled. Without evidence of measurement invariance across ethnic groups, we do not interpret the ethnicity difference substantively. Future studies should retain richer sociocultural background variables while protecting privacy, ensure adequate subgroup sizes and test measurement equivalence across groups.

Overall distress, comparison with norms and symptom profiles

The score structure supports placing overall distress first. Correlations between the nine dimensions were 0.75–0.91, and the first eigenvalue of the 90 items was about 15 times the second; this agrees with

the bifactor analysis of Chinese first-year students, in which a general factor explained most of the common variance and most dimensions had very low specific reliability [8, 9]. It is therefore unsurprising that the global tests of the nine absolute dimensions largely repeated the group pattern of GSI. Within-person centring answers a narrower question: after removing a person's mean symptom level, which dimensions are relatively more prominent in a group? This approach cannot establish distinct disorders or causal pathways, but it avoids treating nine highly correlated "copies" of overall severity as nine independent findings.

The sample as a whole scored below the Chinese adult norms [11] and below the more recent Chinese undergraduate norms [13]. These comparisons have two limitations. The adult norms come from an adult sample collected decades ago that differs from contemporary vocational college students in age, education and historical context, and studies updating the norms also suggest that old norms have limited applicability to contemporary samples [12]; the undergraduate norms come from regular universities and a different mode of administration. Moreover, about one quarter of the present records gave the lowest response to every item; after excluding them, the gap between GSI and the adult norm halved, and phobic anxiety even exceeded the adult norm. The comparison with norms is therefore suitable only as descriptive background and does not support the conclusion that the college's students are "mentally healthier than the general population".

Screening thresholds and response quality

Historical SCL-90 thresholds should remain secondary. In the same data, changing only the inequality operator moved the union flag rate between 17.67% and 19.58%, and normative studies show that old rules may perform differently in contemporary samples [12]. Without structured diagnostic interviews or functional outcomes, the present data can estimate neither the prevalence of mental disorders nor the number of students needing treatment. Screening flags can serve only as a trigger for further assessment within campus services that have clear ethical rules, confidential re-assessment and crisis-referral procedures.

That about one quarter of records gave the lowest response to every item is an important finding for data governance. Automatically deleting such records mathematically raises symptom means and may remove genuinely symptom-free students; retaining them all may dilute group differences if some of them reflect low engagement. More importantly, because invariant responding was unevenly distributed across grade, sex and ethnicity, cleaning decisions affect group comparisons. Our results show that reporting estimates under several prespecified rules is more informative than reporting any single "cleaned sample". Routine screening programmes should retain process metadata, such as completion time and response sequences, and agree on a quality-analysis plan in advance.

Implications for the organization and design of routine digital screening

The response-engagement findings have some implications for the organization and design of campus digital mental health screening, although these are hypotheses based on observational differences. First, engagement differed markedly by grade: about half of grade 2024 records (50.18%) answered “not at all” to all 90 items, with a median completion time of 138.0 s, compared with 12.95% and 288.2 s in grade 2026. The median completion time of invariant records was about 1.3–1.5 s per item, below the common reference of 2 s per item [21]. That reference was proposed for longer questionnaire items, whereas the Chinese SCL-90 items are short symptom phrases that a symptom-free respondent can answer quickly, so this pattern is compatible both with rapid but attentive responding and with cursory reading and cannot by itself show that items were not read. Second, engagement clustered markedly by class: the proportion of invariant responding ranged from 0% to 72.5%, its ICC remained 0.036 after adjustment for group variables, and the ICC of completion time was 0.177, suggesting that class-level context, such as the testing session, instructions and on-site organization, may matter as much as individual motivation. Third, testable directions for improvement include presenting the 90 items across pages with a progress indicator to reduce perceived burden; explaining the purpose, confidentiality rules and use of results in the instructions to build trust; providing brief, non-diagnostic feedback and help-seeking information after completion to increase the perceived value of the screening; and adjusting the timing and organization of screening for senior students who have already been screened several times, or using a shorter screening instrument. Web-survey research shows that questionnaire length and progress indicators affect respondents’ persistence and response quality [29, 30], and preventing low-effort responding is often more effective than removing it afterwards [24]. Fourth, screening platforms should retain process data such as completion time and item-by-item response sequences by default, so that engagement can be monitored continuously. Because all students used the same interface, these suggestions are not a test of interface design; a feasible approach is to embed randomized comparisons of interfaces or instructions in routine screening and to evaluate them with both engagement indicators and external criteria.

Practical implications

For mental health work in vocational colleges, the results support the following practices. First, the entry screening is a key window for identifying distress, and re-assessment, education and support resources should be weighted towards the first weeks after entry, with particular attention to female new entrants. Because an assessment in the first weeks may also capture transient adjustment stress or initial elevation [28], a brief confirmatory re-assessment of flagged entrants several weeks later would help to separate persistent from transient distress before individual follow-up decisions are made. Second, group-level monitoring should report uncertainty, response-quality sensitivity and the exact threshold rules, and should not present screening flag rates as prevalence. Third, for senior students, attention should be paid to engagement with routine screening, for example by improving its presentation, organization and feedback (see the previous section); otherwise the large number of all-minimum records will reduce the information value of the screening. Fourth, individual follow-up should be carried out by qualified professionals through confidential re-assessment that incorporates information on functioning and safety. Cross-sectional group differences do not in themselves show that any

intervention is effective; intervention effects need to be evaluated with validated psychological, academic or service-use outcomes.

Strengths and limitations

Strengths of this study include the large number of records, the high participation rate (95.4%) and complete responses to all 90 items; simultaneous adjustment for the main recorded variables; cluster-robust inference checked by cluster bootstrap, together with false-discovery-rate control; multiple prespecified response-quality sensitivity samples, all reported; separation of overall severity, absolute dimensions and relative profiles; an explanation of the selection sensitivity of the ethnicity estimate through a two-part model; tests of measurement invariance across sex and grade; and checks of the floor effect and model specification with censored, rank-transformed and mixed models. All scores and models were recomputed from raw item scores, the main analyses were checked with independent code, and the analysis code and an anonymised dataset are publicly available.

The study also has limitations. First, the data come from one college's 6-day assessment; despite the high participation rate, the 270 non-participants and the third-year students on placement internships, who were not on the screening roster, may differ from participants, and the results cannot be generalized to vocational college students nationally. Second, the cross-sectional design cannot separate stage of study, entry cohort, age and period effects. Third, the available covariates were limited administrative information; family socioeconomic status, urban–rural background, left-behind experience, social support, academic stress, physical health and help-seeking were not included, so residual confounding is possible. Fourth, measurement invariance was tested with the nine dimension means as indicators, which act as item parcels, and with maximum likelihood estimation rather than with ordinal item-level factor models; the RMSEA of the one-factor model was high, and item-level evidence came from logistic DIF analyses (dichotomous and ordinal) rather than from multi-group item response models, so the invariance evidence remains preliminary. Fifth, self-report scores and historical thresholds were not validated against diagnostic interviews or functional criteria, and the comparisons with norms relied on old adult norms and on undergraduate norms reported separately by urban–rural residence, both of which differ from this study in sample and mode of administration. Sixth, the sensitivity and specificity of the completion-time and invariance rules are unknown, and the 2 s-per-item reference may not transfer to the short Chinese SCL-90 items; the sensitivity analyses reveal how the conclusions depend on the rules but cannot determine which records are truly valid. Seventh, the sex × grade interaction was an exploratory analysis that was not preregistered, and grade 2024 had only 22 class clusters, which limits precision. Eighth, the 13 candidate repeat pairs have not been verified against the platform's records, so the number of records should not be equated with the number of unique students; 57 records came from students younger than the recommended minimum age of 16 years, but excluding them did not change the conclusions. Ninth, we could not obtain the number of previous screenings or students' perceptions of the consequences of screening, so the contribution of repeated administration and strategic under-reporting to the grade differences could not be tested directly; all students used the same platform and interface, so the effect of interface design on engagement could not be tested; and engagement

indicators cannot distinguish low-effort responding from a genuinely low symptom level. Tenth, the within-person centred analysis is subject to an ipsative constraint, so the relative coefficients of the dimensions are not independent.

Conclusions

In this cross-sectional study at a single vocational college, overall distress was highest among new entrants assessed a few weeks after enrolment, and the female excess was concentrated among new entrants; senior students had lower distress and sex differences close to 0. These differences were stable across response-quality rules and in censored and mixed models, and measurement-invariance tests supported comparisons across sex and between grades 2025 and 2026; the ethnicity difference depended on the cleaning rules and on the scale and should not be interpreted substantively. Different formulations of the historical SCL-90 thresholds flag different proportions of records, and these proportions should not be reported as clinical prevalence. Engagement was clearly lower, and clustered by class, in the senior grades, suggesting that part of the grade difference may relate to screening experience and test administration; the organization of campus digital screening should be improved on the basis of experimental research. The results support making the first weeks after entry a priority window for psychological support at vocational colleges and call for multicentre longitudinal studies that include previous screening experience, social and academic background, functional outcomes and external clinical criteria.

Abbreviations

BH: Benjamini–Hochberg; CFI: Comparative fit index; CI: Confidence interval; CR1: Cluster-robust variance estimator with first-order small-sample correction; DIF: Differential item functioning; ECV: Explained common variance; GSI: Global Severity Index; HC3: Heteroskedasticity-consistent standard error (type 3); ICC: Intraclass correlation coefficient; IQR: Interquartile range; OR: Odds ratio; PSDI: Positive Symptom Distress Index; PST: Positive Symptom Total; REML: Restricted maximum likelihood; RMSEA: Root mean square error of approximation; SCL-90: Symptom Checklist-90; SD: Standard deviation; SE: Standard error; SRMR: Standardized root mean square residual; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Declarations

Ethics approval and consent to participate

This study was approved by the Human Research Ethics Committee of Pingliang Vocational and Technical College (approval number: 34502501; September 2026) and was conducted in accordance with the 1964 Declaration of Helsinki and its subsequent revisions, as well as relevant institutional guidelines. Before the routine screening, all students read an electronic informed-consent form and gave consent by clicking “I accept”; informed consent was also obtained from the parents or legal guardians

of the 213 students younger than 18 years. The research team received only de-identified records, and the ethics committee waived individual informed consent for this study.

Consent for publication

Not applicable.

Availability of data and materials

The anonymised record-level dataset supporting the conclusions of this article is included as Additional file 4, together with a codebook. Consent for public release of the dataset was not sought at the time of screening; to protect confidentiality, names, identifiers, dates of birth and age, assessment dates, detailed ethnic group, and major and class identifiers were removed, and local suppression ensured that every combination of college, grade, sex and ethnicity contains at least five records. The release of the anonymised dataset was approved by the Human Research Ethics Committee of Pingliang Vocational and Technical College. Class identifiers, which are needed to reproduce the class-clustered standard errors, are not publicly available for confidentiality reasons but are available from the corresponding authors on reasonable request with approval from the same committee. Analysis code and summary outputs are provided in Additional file 3.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Gansu Provincial Youth Talent Program (Individual Project) (grant number 2026QNGR074) awarded to Ping Fan. The funder had no role in the study design, data collection, analysis and interpretation, or writing of the manuscript.

Authors' contributions

SY: conceptualization, supervision, writing – review and editing. PF: formal analysis, visualization, writing – original draft, funding acquisition. MR, JC and XH: data curation. PM: data curation, writing – original draft. All authors read and approved the final manuscript.

Acknowledgements

We thank the Mental Health Education Center of the participating institution for assistance with data preparation.

Clinical trial number

Not applicable.

References

1. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. J Abnorm Psychol. 2018;127(7):623–38. https://doi.org/10.1037/abn0000362. WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders.
2. Han SS, Zhang YS, Zhu W, Ye YP, Li YX, Meng SQ, et al. Status and epidemiological characteristics of depression and anxiety among Chinese university students in 2023. BMC Public Health. 2025;25:1189. https://doi.org/10.1186/s12889-025-22443-7.
3. Lin ZZ, Cai HW, Huang YF, Zhou LL, Yuan ZY, He LP, et al. Prevalence of depression among university students in China: a systematic review and meta-analysis. BMC Psychol. 2025;13:373. https://doi.org/10.1186/s40359-025-02688-y.
4. Ping Z, Zhou W. Changes in vocational college students' mental health over two academic years: associations with family factors and co-curricular activities. Front Public Health. 2026;14:1901932. https://doi.org/10.3389/fpubh.2026.1901932.
5. Hou H, Feng X, Li Y, Meng Z, Guo D, Wang F, et al. Suboptimal health status and psychological symptoms among Chinese college students: a perspective of predictive, preventive and personalised health. EPMA J. 2018;9(4):367–77. https://doi.org/10.1007/s13167-018-0148-4.
6. Derogatis LR, Lipman RS, Covi L. SCL-90: an outpatient psychiatric rating scale—preliminary report. Psychopharmacol Bull. 1973;9(1):13–28.
7. Wang ZY. Symptom Checklist 90 (SCL-90). Shanghai Arch Psychiatry. 1984;2(2):68–70. (In Chinese).
8. Yu Y, Wan C, Huebner ES, Zhao X, Zeng W, Shang L. Psychometric properties of the Symptom Check List 90 (SCL-90) for Chinese undergraduate students. J Ment Health. 2019;28(2):213–9. https://doi.org/10.1080/09638237.2018.1521939.
9. Wen J, Deng Y. Psychometric structure and score interpretation of the SCL-90 among Chinese first-year undergraduates: an analysis of an open dataset. Front Psychol. 2026;17:1913286. https://doi.org/10.3389/fpsyg.2026.1913286.
10. Preti A, Carta MG, Petretto DR. Factor structure models of the SCL-90-R: replicability across community samples of adolescents. Psychiatry Res. 2019;272:491–8. https://doi.org/10.1016/j.psychres.2018.12.146.
11. Jin H, Wu WY, Zhang MY. Preliminary analysis of SCL-90 assessment results in the Chinese normal population. Chin J Nerv Ment Dis. 1986;12(5):260–3. (In Chinese).
12. Dang W, Xu Y, Ji J, Wang K, Zhao S, Yu B, et al. Study of the SCL-90 scale and changes in the Chinese norms. Front Psychiatry. 2021;11:524395. https://doi.org/10.3389/fpsyt.2020.524395.
13. Yu Y, Wan C, Zhao X, Huebner ES, Tan J, Xu C, et al. Undergraduate students' norms for the Chinese version of the Symptom Check-List-90-R (SCL-90-R). BMC Public Health. 2020;20:1588. https://doi.org/10.1186/s12889-020-09689-z.
14. Lei X, Liu C, Jiang H. Mental health of college students and associated factors in Hubei of China. PLoS ONE. 2021;16(7):e0254183. https://doi.org/10.1371/journal.pone.0254183.

15. Gao W, Ping S, Liu X. Gender differences in depression, anxiety, and stress among college students: a longitudinal study from China. J Affect Disord. 2020;263:292–300. https://doi.org/10.1016/j.jad.2019.11.121.
16. Hang Y, Feng Y, Zou H, Zhao H, Liu F, Qin X, et al. Tracing two decades: non-linear temporal trends in mental health among Chinese university students (2005–2024). J Affect Disord. 2026;393:120341. https://doi.org/10.1016/j.jad.2025.120341.
17. Li F, Liu J, Qiu F, Liu H, Xin S, Yang Q. Changes in mental health levels among Chinese physical education college students from 1995 to 2019. Front Psychol. 2022;13:1034221. https://doi.org/10.3389/fpsyg.2022.1034221.
18. Conley CS, Kirsch AC, Dickson DA, Bryant FB. Negotiating the transition to college: developmental trajectories and gender differences in psychological functioning, cognitive-affective strategies, and social well-being. Emerg Adulthood. 2014;2(3):195–210. https://doi.org/10.1177/2167696814521808.
19. Liu F, Zhou N, Cao H, Fang X, Deng L, Chen W, et al. Chinese college freshmen's mental health problems and their subsequent help-seeking behaviors: a cohort design (2005–2011). PLoS ONE. 2017;12(10):e0185531. https://doi.org/10.1371/journal.pone.0185531.
20. Zhang J, Zhang X. Chinese college students' SCL-90 scores and their relations to college performance. Asian J Psychiatr. 2013;6(2):134–40. https://doi.org/10.1016/j.ajp.2012.09.009.
21. Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. Detecting and deterring insufficient effort responding to surveys. J Bus Psychol. 2012;27(1):99–114. https://doi.org/10.1007/s10869-011-9231-8.
22. Curran PG. Methods for the detection of carelessly invalid responses in survey data. J Exp Soc Psychol. 2016;66:4–19. https://doi.org/10.1016/j.jesp.2015.07.006.
23. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods. 2012;17(3):437–55. https://doi.org/10.1037/a0028085.
24. Ward MK, Meade AW. Dealing with careless responding in survey data: prevention, identification, and recommended best practices. Annu Rev Psychol. 2023;74:577–96. https://doi.org/10.1146/annurev-psych-040422-045007.
25. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol. 1991;5(3):213–36. https://doi.org/10.1002/acp.2350050305.
26. Wise SL, DeMars CE. Low examinee effort in low-stakes assessment: problems and potential solutions. Educ Assess. 2005;10(1):1–17. https://doi.org/10.1207/s15326977ea1001_1.
27. Porter SR, Whitcomb ME, Weitzer WH. Multiple surveys of students and survey fatigue. New Dir Inst Res. 2004;2004(121):63–73. https://doi.org/10.1002/ir.101.
28. Shrout PE, Stadler G, Lane SP, McClure MJ, Jackson GL, Clavél FD, et al. Initial elevation bias in subjective reports. Proc Natl Acad Sci U S A. 2018;115(1):E15–23. https://doi.org/10.1073/pnas.1712277115.

29. Galesic M, Bosnjak M. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opin Q. 2009;73(2):349–60. https://doi.org/10.1093/poq/nfp031.
30. Villar A, Callegaro M, Yang Y. Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. Soc Sci Comput Rev. 2013;31(6):744–62. https://doi.org/10.1177/0894439313497468.
31. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev Rev. 2016;41:71–90. https://doi.org/10.1016/j.dr.2016.06.004.
32. Chen X, Bailey RP, Xie H, Fang Q, Ashok AI, Samsudin N. Profile-specific SCL-90-R symptom networks in Chinese university students: out-of-sample validation in three annual cohorts (2022–2024). Curr Psychol. 2026;45:733. https://doi.org/10.1007/s12144-026-09294-0.
33. Teng X, Wang X, Liu Y, Wu D, Dong Z, Lyu S. Longitudinal network analysis of mental health trends in Chinese university freshmen: a decadal study (2014–2023). Front Psychol. 2025;16:1611264. https://doi.org/10.3389/fpsyg.2025.1611264.
34. Wang S, Chong ZY, Yi M, Zhang Z, Xu W. Symptom network for psychological distress in college freshmen: a large sample Bayesian network analysis. BMC Psychol. 2026;14:737. https://doi.org/10.1186/s40359-026-04501-w.
35. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. https://doi.org/10.1371/journal.pmed.0040296.
36. Zhou J, Yu J, Zhou Y, Qiu J. Study of item text in the Chinese Symptom Checklist-90. Med (Baltim). 2021;100(11):e24841. https://doi.org/10.1097/MD.0000000000024841.
37. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Model. 2007;14(3):464–504. https://doi.org/10.1080/10705510701301834.
38. McNeish D, An J, Hancock GR. The thorny relation between measurement quality and fit index cutoffs in latent variable models. J Pers Assess. 2018;100(1):43–52. https://doi.org/10.1080/00223891.2017.1281286.
39. Meade AW, Kroustalis CM. Problems with item parceling for confirmatory factor analytic tests of measurement invariance. Organ Res Methods. 2006;9(3):369–403. https://doi.org/10.1177/1094428105283384.
40. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. J Educ Meas. 1990;27(4):361–70. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.
41. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Meas Educ. 2001;14(4):329–49. https://doi.org/10.1207/S15324818AME1404_2.
42. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. J Hum Resour. 2015;50(2):317–72. https://doi.org/10.3368/jhr.50.2.317.

43. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
44. Tobin J. Estimation of relationships for limited dependent variables. Econometrica. 1958;26(1):24–36. https://doi.org/10.2307/1907382.
45. Choi SW, Gibbons LE, Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw. 2011;39(8):1–30. https://doi.org/10.18637/jss.v039.i08.
46. Wagenmakers EJ, Krypotos AM, Criss AH, Iverson G. On the interpretation of removable interactions: a survey of the field 33 years after Loftus. Mem Cognit. 2012;40(2):145–60. https://doi.org/10.3758/s13421-011-0158-0.

Figures

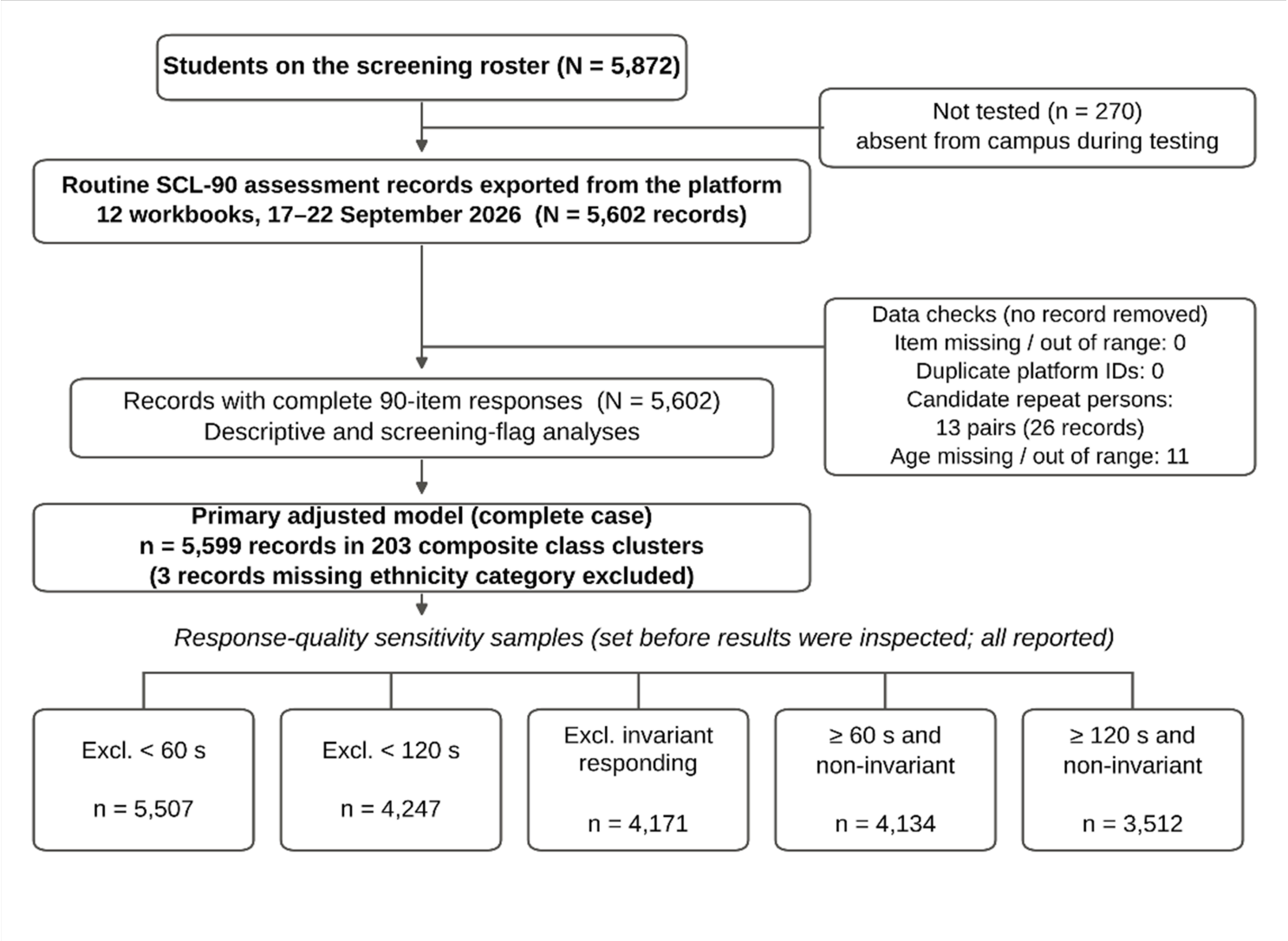


Figure 1

Flow of screening records and analysis samples. Of the 5,872 students on the screening roster, 270 were absent from campus and were not assessed. No records were deleted in the data checks; the five response-quality sensitivity samples were specified before results were inspected and are all reported. In the full sample, 92 records (1.6%) were completed in < 60 s and 1,353 (24.2%) in < 120 s, and 1,429 (25.5%) had identical responses to all 90 items (1,428 with all items answered 1). n for the sensitivity samples refers to complete cases in the adjusted models.

Supplementary Files

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

- Additionalfile1Supplementarytables.docx
- Additionalfile3codeandresults.zip
- Additionalfile4anonymiseddataset.xlsx