ARTÍCULO ORIGINAL

Gac Méd Caracas 2026;134(Supl 2):S514-S527
DOI: 10.47307/GMC.2026.134.S2.23

Psychometric performance of the alcohol use disorders identification test (AUDIT) in medical students: A pilot study

Rendimiento psicométrico de la prueba de identificación de trastornos por consumo de alcohol (AUDIT) en estudiantes de medicina: un estudio piloto

Ángel Ortega[1a], María P. Díaz[2a], José L. Pérez[3a], Raquel Santeliz[4a], Pablo Duran[5a], Valeria Gallo[6a], Rubén Carrasquero[7a], Carla Navarro[8a], Juan Salazar[9a], Juan Hernández-Lalinde[10b], Valmore Bermúdez[11c], Alfonso R. Romero-Conrado[12d], Anthony Lombana Jiménez[13d], Diego Rivera-Porras[14d]

SUMMARY

Background: *Hazardous alcohol use is a major risk factor for morbidity and mortality and contributes to a broad spectrum of conditions, including infectious diseases, cancers, neuropsychiatric disorders, and cardiovascular, hepatic, pancreatic, and metabolic disease. It is also linked to adverse social circumstances and behaviours that may harm others (e.g., accidents and violence). Robust screening instruments are needed to detect and characterise risky consumption early, particularly in high-demand training contexts such as medical education.*

Methods: *A descriptive cross-sectional study was conducted in a non-probabilistic sample of 229 medical students. Data were collected via online self-administered questionnaires. The AUDIT's measurement properties were examined in RStudio, including confirmatory factor analysis (CFA), convergent and discriminant validity (AVE, Fornell–Larcker, HTMT), internal consistency (Cronbach's α; McDonald's ω), and factorial invariance by sex (configural, metric, scalar, and strict).*

Results: *CFA provided provisional support for a coherent three-factor structure. Model fit was acceptable by the chi-square test ($p = 0.101$) and the χ^2/df ratio (1.33). RMSEA, CFI, TLI, GFI, and AGFI were within favourable ranges, whereas SRMR*

DOI: https://doi.org/10.47307/GMC.2026.134.S2.23

ORCID: 0000-0002-8030-1780[1]
ORCID: 0000-0001-6295-0229[2]
ORCID: 0000-0003-0914-5858[3]
ORCID: 0000-0003-0761-5879[4]
ORCID: 0000-0002-8030-1780[5]
ORCID: 0000-0002-2988-7309[6]
ORCID: 0000-0002-3401-2267[7]
ORCID: 0000-0002-1321-9568[8]
ORCID: 0000-0003-4211-528X[9]
ORCID: 0000-0001-6768-1873[10]
ORCID: 0000-0003-1880-8887[11]
ORCID: 0000-0003-4603-0785[12]
ORCID: 0000-0001-7290-8462[13]
ORCID: 0000-0003-2169-3208[14]

Recibido: 20 de diciembre 2025
Aceptado: 1 de febrero 2026

[a]Endocrine-Metabolic Research Center "Dr. Félix Gómez". Faculty of Medicine, University of Zulia, Maracaibo 4001, Venezuela.
[b]Instituto de Estadística Aplicada y Computación, IEAC. Universidad de Los Andes. Mérida 5115, Venezuela.
[c]Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Centro de Investigaciones en Ciencias de la Vida, Barranquilla 080001, Colombia.
[d]Universidad de la Costa, Departamento de Productividad e Innovación, Barranquilla 080001, Atlántico, Colombia.

*Corresponding author: Valmore Bermúdez, Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Centro de Investigaciones en Ciencias de la Vida, Barranquilla 080001, Colombia; E-mail: valmore.bermudez@unisimon.edu.co y Diego Rivera-Porras, Universidad de la Costa, Departamento de Productividad e Innovación, Barranquilla 080001, Atlántico, Colombia; E-mail: drivera23@cuc.edu.co.

was high (0.166), indicating residual misfit at the item level. Convergent validity was acceptable; discriminant validity showed mixed evidence, with overlap suggested by Fornell–Larcker for the dependence–harm boundary while HTMT values remained < 0.85. Internal consistency was adequate (α = 0.85; ω = 0.83). For sex-based measurement invariance, the configural, metric, and scalar models showed acceptable performance, whereas the strict model showed borderline evidence (p = 0.015).
Conclusions: *In this pilot, non-probabilistic sample, the AUDIT showed measurement properties broadly compatible with early screening and initial assessment of hazardous alcohol consumption among medical students in this setting, with latent-level comparisons by sex supported up to the scalar level of invariance. Replication in larger, more diverse samples is needed to confirm generalisability and to clarify the sources of residual misfit indicated by SRMR.*

Keywords: *Alcohol Use Disorders Identification Test; AUDIT; medical students; psychometric validation; confirmatory factor analysis; measurement invariance; internal consistency; hazardous alcohol use.*

RESUMEN

Antecedentes: *El consumo peligroso de alcohol constituye un factor de riesgo relevante para la morbilidad y la mortalidad y se asocia con un amplio conjunto de condiciones, incluyendo enfermedades infecciosas, cáncer, trastornos neuropsiquiátricos y enfermedades cardiovasculares, hepáticas, pancreáticas y metabólicas. Además, se vincula con circunstancias sociales adversas y conductas que pueden afectar a terceros (p. ej., accidentes y violencia). En este contexto, se requieren instrumentos de tamizaje robustos que permitan detectar y caracterizar tempranamente el consumo de riesgo, especialmente en entornos de alta exigencia como la formación médica.*
Métodos: *Se realizó un estudio descriptivo de corte transversal, con muestreo no probabilístico, en 229 estudiantes de medicina. La información se recolectó mediante cuestionarios autoadministrados en línea. Las propiedades de medida del AUDIT se evaluaron en RStudio mediante análisis factorial confirmatorio (AFC), validez convergente y discriminante (AVE, criterio de Fornell–Larcker y HTMT), consistencia interna (α de Cronbach; ω de McDonald) e invariancia factorial por sexo (configural, métrica, escalar y estricta).*
Resultados: *El AFC brindó apoyo provisional a una estructura trifactorial coherente. El ajuste fue aceptable según la prueba de Chi-Cuadrado (p = 0,101) y la razón χ^2/gl (1,33). RMSEA, CFI, TLI, GFI y AGFI se ubicaron en rangos favorables, mientras que el SRMR fue elevado (0.166), lo que indica un desajuste residual a nivel de ítems. La validez convergente fue adecuada; la validez discriminante mostró evidencia mixta, con solapamiento sugerido por Fornell–Larcker en la frontera dependencia–uso dañino, mientras que HTMT se mantuvo < 0,85. La consistencia interna fue adecuada (α = 0,85; ω = 0,83). En invariancia por sexo, los modelos configural, métrico y escalar mostraron un desempeño aceptable; el modelo estricto mostró evidencia limítrofe (p = 0,015), manteniéndose las comparaciones a nivel latente sustentadas hasta el nivel escalar.*
Conclusiones: *En esta muestra piloto con muestreo no probabilístico, el AUDIT mostró propiedades de medida compatibles con su uso como herramienta de tamizaje temprano y de evaluación inicial del consumo peligroso de alcohol en estudiantes de medicina en este contexto, con comparaciones a nivel latente por sexo sustentadas hasta el nivel escalar de invariancia. Se recomienda replicar en muestras mayores y más diversas para consolidar la generalización y aclarar el origen del desajuste residual sugerido por el SRMR.*

Palabras clave: *Prueba de Identificación de los Trastornos por Consumo de Alcohol; AUDIT; estudiantes de medicina; validación psicométrica; análisis factorial confirmatorio; invarianza factorial; consistencia interna; consumo peligroso de alcohol.*

INTRODUCTION

The conceptualisation of harmful alcohol use has shifted from a purely prospective risk framing to a definition centred on demonstrable physical or psychological harm, as codified by the World Health Organization (WHO) in ICD-10 (1). Despite sustained prevention and control efforts, alcohol-related harm remains a major global public-health challenge. Current international reporting attributes a substantial mortality burden to alcohol use, alongside broad contributions to disease, injury, and social harm (2,3). A persistent difficulty is that alcohol-related problems often become visible only after social, emotional, or health consequences have consolidated, which limits the window for timely intervention (4). Earlier identification of hazardous patterns is therefore a pragmatic prevention strategy, enabling risk reduction before entrenched dependence or severe harm emerges (4).

Within university populations, medical students constitute a group of particular interest because hazardous or harmful drinking has been documented across diverse settings, frequently with higher levels among men (5). Medical training typically entails sustained academic pressure, heavy workloads, and repeated exposure to emotionally demanding situations. These conditions co-occur with elevated stress and burnout and with symptoms of anxiety and depression in many cohorts (6,7). In that context, alcohol may be adopted as a maladaptive coping strategy, even when average consumption in the wider student body is moderate. Robust screening tools are consequently central to the detection of early risk among trainees in high-demand educational environments.

The Alcohol Use Disorders Identification Test (AUDIT) is one of the most widely implemented instruments for this purpose, in part because it offers a brief, self-administered format suitable for both clinical and educational settings (8). The questionnaire comprises ten items aligned with ICD-10-oriented domains capturing hazardous use, dependence symptoms, and harmful use, supporting a multidimensional characterisation of alcohol-related risk (8). Evidence from student and young-adult samples across multiple countries indicates that the AUDIT performs adequately in university settings, including medical student cohorts (9,10). Its practical scoring and interpretability have supported its adoption in primary care, population surveillance, and institutional screening programmes (8).

Spanish-language adaptations of the AUDIT have been developed and evaluated over several decades. Early validations were oriented to primary care and clinical use (11), followed by translations and cultural adaptations into Catalan and Spanish (12). Subsequent studies responded to shifting consumption patterns among young adults and assessed measurement properties in university populations (10,13). Further psychometric evaluations focusing specifically on medical students have also been reported (14–16). Nevertheless, measurement performance is not guaranteed to transport across contexts, and the validity of latent comparisons depends on demonstrating that the instrument operates similarly across key groups, particularly sex.

Against this background, the present work evaluates the measurement properties of a locally adapted Spanish-language AUDIT in medical students at a Latin American university, with specific attention to sex-based measurement invariance. The research question is whether the instrument supports a coherent factorial structure and equivalent measurement across male and female students, permitting defensible latent-level comparisons in this training context.

MATERIALS AND METHODS

Study Design and Participant Selection

A pilot cross-sectional, instrumented design was implemented to evaluate the measurement performance of the AUDIT among medical students enrolled in the first through sixth academic years at a major public university in Venezuela. Institutional enrolment records for the 2018 academic period indicated a target population of 3,480 students. A non-probabilistic quota sampling strategy was applied to ensure representation by academic year. Six cohorts were defined (first to sixth year) and quotas were set as follows: 33 (first year), 38 (second), 38 (third), 41 (fourth), 45 (fifth), and 34 (sixth), yielding a total sample of 229 participants; this corresponded to approximately 30 % of each year's quota. Eligibility criteria were age ≥18 years, any sex, and formal registration in the medical programme during the specified academic period. Participation was voluntary, and only records meeting all eligibility criteria were retained for analysis.

Procedure

Data were collected online via Google Forms, accessed via individual links distributed via institutional email. Participants completed the questionnaire independently and submitted responses electronically. Informed consent was obtained before any assessment content was displayed. Ethical approval was granted by the Bioethics Committee of the Endocrine–Metabolic Research Centre "Dr Félix Gómez", Faculty of Medicine, at the same university (committee minutes/approval reference: Acta No. (INSERT), dated (INSERT DATE), issued in (INSERT CITY), Venezuela). Data were

handled in a de-identified format for analysis, and participation involved no academic or administrative consequences.

Instrument

A Spanish-language adaptation of the Alcohol Use Disorders Identification Test (AUDIT) was administered. The AUDIT is a brief self-administered screening tool designed to identify hazardous and harmful alcohol use and related problems in clinical and community settings. The instrument comprises 10 items: items 1–8 are scored from 0 to 4, and items 9–10 use a 0–2–4 scoring format, producing total scores from 0 to 40. Consistent with commonly used thresholds, scores ≥8 for men (17) and ≥6 for women (8) were treated as indicators of hazardous consumption, while scores ≥13 were considered suggestive of greater severity and probable dependence in both sexes (18).

Statistical Analysis

Construct Validity (Confirmatory Factor Analysis)

Construct validity was evaluated through confirmatory factor analysis (CFA). Multivariate normality and atypical multivariate observations were examined using Mardia's test and robust Mahalanobis distances; these checks indicated departures from normality and the presence of atypical observations. Given the ordinal response format and non-normality, the weighted least squares estimator with mean and variance adjustment (WLSMV) was selected. Model fit was evaluated using the χ^2 statistic, the χ^2/df ratio, RMSEA, and SRMR. Ratios of χ^2/df below 2–3 were interpreted as acceptable. RMSEA values <0.06 and SRMR values <0.08 were treated as indicative of adequate fit, while additional indices (GFI, AGFI, CFI, and TLI) were also inspected, with values ≥0.95 considered optimal (19). To support interpretation when global fit indices diverged (e.g., acceptable RMSEA/CFI with elevated SRMR), residual-based diagnostics (e.g., standardised residuals and modification information) were used to characterise potential sources of local misfit; these diagnostics were reported descriptively alongside the retained model in the Results section.

Convergent and Discriminant Validity

Convergent validity was assessed using average variance extracted (AVE), adopting 0.50 as the minimum acceptable value (20). Discriminant validity was evaluated using the Fornell–Larcker criterion (square root of AVE exceeding shared variance with other constructs) and the heterotrait–monotrait ratio (HTMT), using <0.85 as evidence of adequate discriminant validity (21).

Internal Consistency

Reliability was estimated using Cronbach's alpha and McDonald's omega. Values <0.50 were treated as inadequate, ≥0.70 as acceptable, ≥0.80 as good, and 0.90–0.95 as excellent (22).

Factorial Invariance by Sex

Measurement invariance across sex was examined via a sequence of increasingly constrained models: configural, metric, scalar, and strict invariance. Model comparisons relied on a combination of ΔCFI and ΔRMSEA, supplemented by $\Delta\chi^2$, recognising the sensitivity of χ^2 to sample size and distributional features. Invariance was considered supported when changes satisfied ΔCFI ≤ 0.01 and ΔRMSEA ≤ 0.015 across nested models; larger deteriorations were interpreted as evidence against invariance (23). All analyses were conducted in SPSS version 24 and RStudio version 1.1.463, and statistical significance was set at $p < 0.05$.

RESULTS

General characteristics of the sample

Table 1 summarises the demographic profile of the cohort and the distribution of total AUDIT scores. This descriptive context is necessary for interpreting the subsequent psychometric results, particularly because the observed total-score range did not extend to the upper half of the instrument's theoretical scale.

Table 1. Sample characteristics and AUDIT distribution (N = 229)

Variable	Statistic
Age (years), mean (SD)	21 (2)
Age (years), 95 % CI for mean	20.74 to 21.26
Sex: Women, n (%)	125 (54.6)
Sex: Women, 95 % CI for proportion	48.1 to 60.9
Sex: Men, n (%)104 (45.4)	
Sex: Men, 95% CI for proportion	39.1 to 51.9
AUDIT total score, observed range	0–20
AUDIT total score, mean (SD)	4 (3)
Participants scoring above 8 points (AUDIT > 8), n (%)	28 (12.2)
Participants scoring above 8 points (AUDIT > 8), 95 % CI for proportion	8.6 to 17.1

Source: Own elaboration. Note. Values are mean (SD) or n (%), as indicated. Proportion confidence intervals are 95% Wilson intervals. The AUDIT theoretical total-score range is 0–40; the observed range in this sample was 0–20. "Participants scoring above 8 points" denotes those with total AUDIT scores strictly greater than 8, reported here for descriptive screening purposes. Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; CI, confidence interval; SD, standard deviation.

As shown in Table 1, the cohort was young, with a narrow age distribution (mean age 21 years; 95 % CI 20.74–21.26). Women constituted a modest majority (54.6 %; 95 % CI 48.1–60.9), indicating near balance by sex rather than a highly skewed composition. Total AUDIT scores were concentrated at low levels (mean 4; SD 3), and the observed maximum (20 points) reached only half of the instrument's theoretical range (0–40), indicating that higher-severity response options were infrequently endorsed in this pilot cohort. Despite the generally low central tendency, a distinct subgroup screened above 8 points (12.2 %; 95 % CI 8.6–17.1), corresponding to approximately one in eight participants. This distributional pattern (low average with a non-trivial right-tail subgroup) is relevant for interpreting subsequent construct validation results, because restricted score ranges can reduce item covariances and affect factor separation and reliability indicators.

Factorial validity and model fit (confirmatory factor analysis)

Table 2 reports the global fit indices for the competing CFA specifications evaluated in this study: the prespecified three-factor structure (Model 1), a reduced-item three-factor sensitivity model (Model 2; items with standardised loadings < 0.70 removed), and an alternative unidimensional specification (Model 3). This block aims to establish which specification offers the most defensible balance between empirical fit and theoretical interpretability before proceeding to validity and reliability evidence.

As shown in Table 2, the prespecified three-factor model (Model 1) demonstrated favourable global fit by χ^2 (p = 0.101), χ^2/df (1.33), and incremental indices (CFI = 0.989; TLI = 0.985), alongside a low RMSEA (0.038; 90 % CI 0.000–0.066) and high absolute fit indices (GFI = 0.992; AGFI = 0.982). However, SRMR was markedly elevated (0.166), indicating substantial residual misfit at the item level despite otherwise favourable global indices. This divergence implies that, in this pilot cohort, the three-factor structure is *globally coherent* but exhibits *local discrepancies* that should temper any overly definitive claims about item-level fit.

Model 2 reduced SRMR to .0.066, bringing residual fit within conventional thresholds. That improvement came with trade-offs: χ^2 became statistically significant (p=0.017), χ^2/df increased to 2.10, and RMSEA rose to 0.069 (90 % CI 0.028–0.109), a range consistent with only moderate approximation and greater uncertainty. Importantly, Model 2 is also structurally narrower by design (item removal driven by a loading threshold), which can improve fit mechanically while reducing content coverage of the AUDIT construct domain. As a result, Model 2 is best

Table 2. CFA model fit indices across competing specifications

Model	χ^2	df	p	χ^2/df	RMSEA	90% CI RMSEA	SRMR	CFI	TLI	GFI	AGFI
Model 1 (original)	42.54	32	0.101	1.33	0.038	0.000–0.066	0.166	0.989	0.985	0.992	0.982
Model 2 (items < .70 removed)	23.09	11	0.017	2.10	0.069	0.028–0.109	0.066	0.989	0.978	0.995	0.982
Model 3 (unidimensional)	3.90	2	0.143	1.95	0.064	0.000–0.160	0.045	0.997	0.991	0.999	0.987

Source: Own elaboration

Note. χ^2/df values closer to 1 indicate better relative fit. RMSEA values < 0.06 and SRMR values < .08 are typically interpreted as favourable approximation and residual fit, respectively. Incremental indices (CFI/TLI) ≥ .95 are commonly interpreted as favourable. Because χ^2 is sensitive to sample size and model constraints, interpretation prioritises convergence across multiple indices rather than χ^2 alone.

treated as a sensitivity check rather than a preferred measurement model for substantive interpretation.

Model 3 (unidimensional) achieved the most favourable residual fit (SRMR = 0.045) and strong incremental indices (CFI = 0.997; TLI = 0.991), with RMSEA = 0.064 and a wide confidence interval reflecting the very small df. While this specification fits the data well empirically, it collapses the theoretically meaningful AUDIT domains into a single dimension, thereby weakening interpretability when the research objective requires domain-specific validity evidence and sex-based invariance across the established AUDIT structure.

Taken together, these results support retaining the original three-factor specification (Model 1) as the primary model for subsequent validity, reliability, and invariance reporting, while explicitly acknowledging the elevated SRMR as evidence of residual item-level misfit in this pilot sample. This positioning is also consistent with the study's purpose (psychometric evaluation of the established AUDIT structure) and avoids overreliance on mechanically improved fit achieved via post hoc item removal.

Convergent validity, discriminant validity, and internal consistency

Table 3 summarises average variance extracted (AVE) and reliability estimates for each latent dimension in the retained three-factor solution (Model 1). Two dimensions met the conventional AVE ≥ 0.50 benchmark (F1 = 0.67; F2 = 0.58), whereas the harmful-use dimension fell slightly below that threshold (F3 = 0.46), consistent with a domain that often exhibits lower prevalence and greater heterogeneity in young, largely low-risk samples.

Table 3. Convergent validity (AVE) and reliability indices (Model 1)

Factor	AVE	Cronbach's α	McDonald's ω
F1 – Hazardous Use	0.67	0.85	0.80
F2 – Dependence Symptoms	0.58	0.80	0.59
F3 – Harmful Use	0.46	0.72	0.61
Total scale	0.56	0.85	0.83

Source: Own elaboration.

Note. AVE values ≥ 0.50 are typically interpreted as adequate convergent validity (20). Reliability coefficients ≥ 0.70 are commonly treated as acceptable for group-level inference (22).

From a reliability perspective, the full scale exhibited stable internal consistency ($\alpha = 0.85$; $\omega = 0.83$), supporting its use for screening-oriented applications. At the dimension level, alpha coefficients remained ≥ 0.70 across factors, while omega was notably lower for dependence symptoms ($\omega = 0.59$) and harmful use ($\omega = 0.61$). This divergence is compatible with (i) the smaller number of indicators per factor and (ii) the lower endorsement frequency expected for dependence and consequence items in a sample with low mean AUDIT scores, both of which can attenuate omega despite acceptable alpha.

Table 4 clarifies discriminant validity by presenting (a) the square roots of AVE on the diagonal and (b) latent correlations off-diagonal (derived from the squared correlations reported for Model 1). This format aligns directly with the Fornell–Larcker logic, where √AVE for each factor should exceed its correlations with other factors.

Table 4. Discriminant validity using √AVE (diagonal) and latent correlations (Model 1).

	F1	F2	F3
F1 (Hazardous use)	0.819	0.548	0.624
F2 (Dependence symptoms)	0.548	0.762	0.866
F3 (Harmful use)	0.624	0.866	0.678

Source: Own elaboration

Note. Diagonal values are √AVE (computed from Table 3). Off-diagonal values are latent correlations (computed as √ the squared correlations reported in the original Fornell–Larcker table). Under the Fornell–Larcker criterion (20), discriminant validity is supported when √AVE exceeds inter-factor correlations.

Using this representation, discriminant validity was clearly supported for F1 versus the other dimensions (√AVE_F1 = 0.819 > r_{F1–F2} = 0.548 and r_{F1–F3} = 0.624). In contrast, the dependence–harmful interface showed substantial overlap (r_{F2–F3} = 0.866), exceeding both √AVE_F2 (0.762) and √AVE_F3 (0.678). This pattern indicates that, in this sample, dependence symptoms and harmful consequences share a substantial proportion of variance, consistent with the empirical boundary between these domains being comparatively permeable in low-severity populations.

Because Fornell–Larcker can be conservative and sensitive to high factor correlations, Table 5 reports the Heterotrait–Monotrait ratio (HTMT) as a complementary discriminant validity check.

All HTMT values fell below 0.85, including the dependence–harmful contrast (HTMT = 0.83), which supports discriminant validity under this criterion while signalling a near-threshold relationship. Taken together, the discriminant validity evidence is best characterised as robust separation of hazardous use from the other domains, alongside a tight dependence–harmful coupling that remains acceptable by HTMT but should be interpreted cautiously when making fine-grained inferences that rely on strict domain separability.

Table 5. HTMT ratios (Model 1).

Factor pair	HTMT
F1 – F2	0.41
F1 – F3	0.55
F2 – F3	0.83

Source: Own elaboration.

Note. HTMT values < 0.85 are commonly interpreted as acceptable discriminant validity (21).

Measurement invariance across sex (multi-group CFA)

Table 6 summarises the sequence of increasingly constrained multi-group CFA models used to evaluate measurement invariance across sex (configural → thresholds equal → metric → scalar → strict). The interpretation below prioritises *approximate-fit stability* (ΔCFI and ΔRMSEA) alongside $\Delta\chi^2$, recognising the sensitivity of χ^2 to sample size and constraint accumulation.

Table 6. Measurement invariance across sex for the three-factor model

Model	χ^2 (df)	p	χ^2/df	$\Delta\chi^2$	p($\Delta\chi^2$)	CFI	|ΔCFI|	RMSEA	|ΔRMSEA|
1. Configural	66.29 (64)	0.398	1.04	—	—	0.998	—	0.018	—
2. Thresholds equal	70.06 (67)	0.375	1.05	4.59	0.203	0.998	0.000	0.020	0.002
3. Metric	81.70 (74)	0.252	1.10	13.64	0.057	0.994	0.004	0.030	0.010
4. Scalar	81.77 (77)	0.334	1.06	0.27	0.965	0.996	0.002	0.023	0.007
5. Strict	101.30 (87)	0.140	1.16	21.99	0.015	0.989	0.007	0.038	0.015

Source: Own elaboration.
Note. |ΔCFI| and |ΔRMSEA| denote absolute changes relative to the immediately preceding (less constrained) model. As a practical decision rule, invariance is typically considered supported when |ΔCFI| ≤ 0.010 and |ΔRMSEA| ≤ 0.015 across nested models; $\Delta\chi^2$ is reported for completeness.

As shown in Table 6, the configural model fit well (χ^2/df = 1.04; RMSEA = 0.018; CFI = 0.998), supporting a comparable baseline factor pattern across men and women. Constraining thresholds (thresholds-equal) produced negligible practical-fit change (|ΔCFI| = 0.000; |ΔRMSEA| = 0.002), indicating stable category functioning across sex. Moving to the metric model, fit deterioration remained modest (|ΔCFI| = 0.004; |ΔRMSEA| = 0.010), supporting the equivalence of factor loadings and, therefore, comparability of the measurement unit across groups.

Under the scalar model, practical-fit changes were again small (|ΔCFI| = 0.002; |ΔRMSEA| = 0.007), indicating stable intercepts/thresholds in addition to loadings. This level of invariance is the key requirement for unbiased latent mean comparisons by sex, and it is therefore the primary anchor for group comparisons in the present study.

For the strict model (residual constraints), $\Delta\chi^2$ reached conventional statistical significance (p = 0.015), while practical-fit changes remained within (CFI) or at the boundary (RMSEA) of common cut-offs (|ΔCFI| = 0.007; |ΔRMSEA| = 0.015). On that basis, strict invariance is best interpreted as borderline rather than unequivocally supported. Importantly, this does not undermine the main comparability claim in this manuscript, because scalar invariance was retained, and scalar (not strict) is the standard requirement for latent mean comparisons.

Across the pilot cohort, total AUDIT scores were concentrated at low values. They spanned only 0–20 despite the instrument's 0–40 theoretical range (Table 1), indicating limited endorsement of higher-severity response options in this setting. Within that distributional context, the retained three-factor CFA solution demonstrated favourable global fit (Table 2) with strong incremental indices and low RMSEA. At the same time, SRMR remained high, signalling meaningful item-level residual misfit, warranting cautious interpretation of fine-grained item functioning. Convergent validity was acceptable overall (Table 3), and reliability for the full scale was adequate (Table 3), supporting screening-oriented use. Discriminant validity evidence was strongest for the separation of hazardous use from the other domains. In contrast, dependence symptoms and harmful consequences were tightly coupled: Fornell–Larcker indicated substantial

overlap, while HTMT remained below the conventional threshold, supporting discriminant validity but at a near-boundary level (Tables 4-5). Finally, multi-group analyses supported measurement invariance across sex through the scalar level (Table 6), indicating that latent-level comparisons between male and female students are defensible within this sample. In contrast, strict invariance was borderline under approximate-fit criteria.

DISCUSSION

Medical students constitute a high-relevance population for alcohol-risk screening because training demands co-occur with sustained academic pressure, irregular sleep, and repeated exposure to clinical stressors. In several university settings, hazardous consumption has persisted or increased across cohorts, with psychosocial strain and coping dynamics shaping risk trajectories (24,25). Within this cohort, the descriptive profile was characterised by low average AUDIT scores alongside a non-trivially higher-risk subgroup (Table 1). That distributional configuration is clinically meaningful: a modest right tail can coexist with a low central tendency, particularly in early adulthood and in academically selective programmes (5,27).

The retained three-factor configuration aligns with the AUDIT's conceptual partitioning into hazardous use, dependence symptoms, and harmful use (8). Global fit for the original model was favourable, as indicated by χ^2/df, RMSEA, and incremental indices (Table 2), and the pattern is consistent with prior validations that have retained multidimensional AUDIT representations across diverse cultural and educational settings (9,10,28). Retaining the established structure, therefore, preserves interpretability at the domain level, which is relevant when the intent is not merely total-score screening but also a structured view of symptom clustering across consumption, dependence features, and consequences.

A key technical feature in this dataset is the divergence between favourable global indices and a markedly elevated SRMR in the original model (Table 2). SRMR is a residual-based index and, in this context, points to local item-level discrepancies that are not fully captured by global approximation measures (19). Two aspects of the observed data provide a parsimonious context for that pattern. First, the total-score range was restricted to 0–20 despite a theoretical range of 0–40 (Table 1), suggesting limited use of higher response categories. Range restriction can compress covariances and amplify the influence of small residual dependencies on residual-based diagnostics. Second, in low-severity samples, dependence and consequence indicators are often infrequently endorsed, which can increase local misfit even when the broader factor pattern remains stable.

Competing specifications reinforced this trade-off between empirical fit and construct coverage (Table 2). The reduced-item three-factor model achieved a materially lower SRMR, but at the cost of narrowing content representation through mechanical item removal. The unidimensional alternative fits well empirically, yet it collapses clinically meaningful domains into a single continuum, weakening domain-level interpretability when the analytic objective includes construct-level validity evidence and group comparability within an established AUDIT framework (8). For these reasons, the original three-factor model remains the most defensible primary specification for interpretation in this pilot cohort, provided that item-level fit is described cautiously and framed as an area for targeted diagnostic follow-up in larger samples.

The convergent validity profile was broadly supportive of the retained structure. Two dimensions met the conventional AVE benchmark (Table 3), indicating that, on average, indicators of hazardous use and dependence symptoms captured a substantial proportion of variance attributable to their intended constructs (20). The harmful-use domain showed a slightly lower AVE (Table 3). This pattern is plausible in a young, largely low-risk cohort where adverse consequences may be episodic, context-dependent, and less consistently endorsed. In such samples, consequence items can behave as low-prevalence indicators, thereby reducing common variance and depressing AVE, even when the factor remains substantively interpretable.

Discriminant validity was best characterised as robust separation of hazardous use from the other two domains, alongside a tight dependence–harm coupling (Tables 4-5). Under the Fornell–Larcker logic, the dependence and harmful-use factors exhibited substantial overlap (Table 4), suggesting limited separation by that criterion (20). However, HTMT values remained below the conventional 0.85 threshold, including for the dependence–harm contrast (Table 5), supporting discriminant validity under an alternative criterion that has been argued to provide improved sensitivity in simulation work (21). The coexistence of Fornell–Larcker "flags" and HTMT "passes" is not unusual when factors are strongly correlated but not redundant, and it is consistent with a partially permeable boundary between dependence symptoms and consequences in low-severity populations. Substantively, dependence features (e.g., impaired control) and early negative consequences can co-occur in tightly coupled patterns among young adults, even when overall consumption levels are modest, thereby increasing inter-factor associations and reducing apparent separation.

Reliability results also fit this interpretation. Internal consistency for the total AUDIT score was adequate, as indicated by both alpha and omega (Table 3), supporting its practical use as a screening-oriented index. At the domain level, alpha coefficients remained acceptable, whereas omega values were lower for the dependence and harmful-use dimensions (Table 3). This divergence is consistent with two technical features: (i) omega is more sensitive to the distribution of factor loadings and to the effective indicator strength, and (ii) small-item factors combined with restricted score variability can attenuate omega even when alpha remains acceptable. In practical terms, this suggests that the total score is the most stable reliability anchor in this pilot cohort. At the same time, domain-level interpretation should be maintained but framed with appropriate caution — particularly for contrasts that rely on fine-grained discrimination between dependence symptoms and harmful consequences.

Taken together, the validity and reliability evidence indicate that the instrument behaves in a manner compatible with its intended domain structure in this context, while also signalling where interpretative caution is warranted: domain separability at the dependence–harm boundary and the stability of domain-specific reliability in a low-severity sample with restricted score range.

A central contribution of this study is the evaluation of measurement invariance across sex. The multi-group CFA sequence indicated stable measurement performance through the scalar level (Table 6). Configural invariance supports a comparable baseline factor pattern across male and female students, and the negligible changes observed under threshold and metric constraints indicate that item functioning and factor loadings remained broadly comparable across groups. Scalar invariance — the combination of equivalent loadings and intercepts/thresholds — was also retained with small practical-fit changes (Table 6). This level is the key requirement for latent mean comparisons because it implies that group differences reflect variation in the underlying constructs rather than systematic measurement artefacts.

In contrast, strict invariance showed borderline evidence (Table 6): $\Delta\chi^2$ was statistically significant, ΔCFI remained within conventional tolerances, and ΔRMSEA fell within the typical decision boundary. This pattern is best interpreted as indicating that residual variances may not be fully equivalent across sexes, while the core measurement structure remains comparable. Importantly, strict invariance is not required for the primary interpretive use emphasised in this manuscript — namely, latent-level comparisons anchored to scalar invariance. Accordingly, the sex-comparability conclusion in this pilot cohort should be framed as robust through scalar invariance, with residual-level equivalence treated as uncertain and therefore not relied upon for fine-grained item-level or residual-based inference.

From an applied standpoint, these results support the use of the AUDIT as a feasible instrument for early screening and initial risk stratification among medical students in this setting. The presence of a discernible higher-risk subgroup despite a low mean score (Table 1) underscores the practical value of screening in academically demanding environments. However, the elevated SRMR in the retained model (Table 2) and the near-boundary discriminant-validity pattern at the dependence–harm interface (Tables

4–5) indicate that interpretation should focus on the instrument's global and construct-level properties rather than on overly granular item-level conclusions in this pilot sample.

Several limitations should be considered when interpreting these findings. First, the sampling strategy was non-probabilistic and quota-based, designed to ensure representation by academic year rather than population-level inference. Consequently, the results should be interpreted as evidence of measurement performance within this pilot cohort, not as definitive generalisability to all medical students or to broader university populations. Second, the study was conducted in a single institutional setting, which limits transportability across educational systems, regional contexts, and programme structures that may shape both drinking patterns and response behaviour.

Third, data were obtained via self-administered online questionnaires, which may be affected by social desirability bias, recall error, and differential willingness to disclose alcohol-related behaviours. These biases can be particularly salient in professional training contexts where participants may anticipate reputational consequences, even when confidentiality is emphasised. Fourth, the observed total-score range was restricted (0–20) relative to the AUDIT's theoretical range (0–40) (Table 1). Range restriction and low endorsement of high-severity categories can compress item covariance, reduce extracted common variance in consequence and dependence domains, and amplify the influence of local residual dependencies. This limitation provides a coherent context for three of the key empirical features observed here: the elevated SRMR despite otherwise favourable global fit (Table 2), the slightly weaker convergent metrics for the harmful-use factor (Table 3), and the strong coupling between dependence symptoms and harmful consequences reflected in discriminant validity checks (Tables 4–5).

Despite these limitations, the study has strengths relevant to applied measurement work in student populations. The evaluation was anchored in an established theoretical framework for the AUDIT (8), and the analysis integrated complementary evidence on global fit, convergent validity, discriminant validity, internal consistency, and multi-group invariance. Notably, demonstrating invariance through the scalar level across sex (Table 6) is a substantive contribution because it supports defensible latent-level comparisons in this cohort and reduces a common interpretive vulnerability in cross-group screening research.

Future work should prioritise replication in larger, multi-site, and probabilistic samples, with explicit attention to score-distribution properties and item-level diagnostics. In practical terms, this includes (i) evaluating whether local misfit is attributable to specific residual dependencies, threshold functioning, or context-bound item interpretation; (ii) testing alternative, theoretically justified model refinements when warranted; and (iii) re-examining domain-specific reliability and discriminant validity in cohorts with broader score ranges and higher endorsement of consequence items. Such studies would consolidate the measurement conclusions and refine guidance for both screening implementation and interpretation in medical-student settings.

CONCLUSIONS

In this pilot cohort of Venezuelan medical students, the AUDIT demonstrated evidence of measurement performance broadly compatible with early screening and initial risk stratification in this training context. Total-score reliability was adequate and supports use at the scale level for screening-oriented applications.

The retained three-factor structure showed favourable global fit. Still, residual-based misfit (elevated SRMR) indicates that item-level functioning and fine-grained domain separation should be interpreted cautiously in this sample. Convergent validity was acceptable overall, and discriminant validity was strongest for distinguishing hazardous use from the other domains, with a near-boundary dependence–harm interface.

Measurement invariance across sex was supported through the scalar level, indicating that latent-level comparisons between male and female students are defensible within this cohort, whereas strict invariance was borderline. Replication in larger, multi-site, and more diverse

samples is required to confirm generalisability, broaden coverage of the score range, and clarify the sources of local misfit identified in this pilot study.

DECLARATIONS

Author Contributions: Conceptualization, (A.O.), (M.P.D.), (J.H.-L.), (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); methodology, (J.L.P.), (J.H.-L.), (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); software, (J.H.-L.), (V.B.), and (D.R.-P.); validation, (J.L.P.), (R.S.), (R.C.), (J.H.-L.), (V.B.), and (D.R.-P.); formal analysis, (J.L.P.), (J.H.-L.), (V.B.), and (D.R.-P.); investigation, (A.O.), (M.P.D.), (J.L.P.), (R.S.), (P.D.), (V.G.), (R.C.), (C.N.), (J.S.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); resources, (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); data curation, (A.O.), (M.P.D.), (J.L.P.), (R.S.), (P.D.), (V.G.), (R.C.), (C.N.), (J.S.), and (J.H.-L.); writing—original draft preparation, (A.O.), (M.P.D.), (J.L.P.), (R.S.), (P.D.), (V.G.), (J.H.-L.), (V.B.), and (D.R.-P.); writing—review and editing, (A.O.), (M.P.D.), (J.L.P.), (R.S.), (P.D.), (V.G.), (R.C.), (C.N.), (J.S.), (J.H.-L.), (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); visualization, (C.N.), (J.S.), (J.H.-L.), (V.B.), and (D.R.-P.); supervision, (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.); project administration, (A.O.), (A.L.J.), and (D.R.-P.); funding acquisition, (V.B.), (A.R.R.-C.), (A.L.J.), and (D.R.-P.). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Ethical Considerations: This study was approved by the Endocrine and Metabolic Diseases Research Center's Bioethics Committee. All participants signed a written informed consent form before being interviewed and examined by a trained team.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. World Health Organization, editor. The ICD-10 classification of mental and behavioural disorders. Geneva: World Health Organization; 1992:320-324.
2. World Health Organization. Global status report on alcohol and health. 2018. Available from: https://www.who.int/publications/i/item/9789241565639
3. Rehm J, Mathers C, Popova S, Thavorn-charoensap M, Teerawattananon Y, Patra J. Global burden of disease and injury and eco-nomic cost attributable to alcohol use and alco-hol-use disorders. Lancet. 2009;373(9682):222-233.
4. Babor TF, Casswell S, Graham K, Huckle T, Livingston M, Österberg E, et al. Alcohol: No Ordinary Commodity: Research and Public Policy. 3rd edition, New to this Edition. 3rd edition, New to this Edition: Oxford, New York: Oxford University Press; 2023:384.
5. Karam E, Kypri K. Alcohol use among col-lege students: An international perspective. Curr Opin Psychiatry. 2007;20(3):213-221.
6. Dyrbye LN, Thomas MR, Shanafelt TD. Sys-tematic review of depression, anxiety, and other indicators of psychological distress among U.S. and Canadian medical students. Acad Med. 2006;81(4):354-373.
7. Jackson ER, Shanafelt TD, Hasan O, Satele DV, Dyrbye LN. Burnout and Alcohol Abuse/Dependence Among U.S. Medical Students. Acad Med. 2016;91(9):1251-1256.
8. Babor TF, Higgins-Biddle JC. The Alcohol Use Disorders Identification Test (AUDIT) manual Guidelines for use in primary care (In-ternet). 2nd ed. 2001. Available from: https://www.scirp.org/reference/referencespapers?referenceid=1104267
9. Gual A, Segura L, Contel M, Heather N, Colom J. Audit-3 and audit-4: Effectiveness of two short forms of the alcohol use disorders identi-fication test. Alcohol Alcohol. 2002;37(6):591-596.
10. Kokotailo PK, Egan J, Gangnon R, Brown D, Mundt M, Fleming M. Validity of the alcohol use disorders identification test in college stu-dents. Alcohol Clin Exp Res. 2004;28(6):914-920.
11. Delgado JMM. Validación de los cuestionarios breves AUDIT, CAGE y CBA para la detección precoz de los problemas relacionados con el consumo de bebidas alcohólicas en atención primaria. Comisionado para la Droga; 1999:147.
12. Guillamón MC, Solé AG, Farran JC. Test para la identificación de trastornos por uso de alcohol (AUDIT): Traducción y validación al catalán y al castellano. Adicciones. 1999;11(4):337-347.

13. Adewuya AO. Validation of the alcohol use disorders identification test (AUDIT) as a screening tool for alcohol-related problems among Nigerian university students. Alcohol Alcohol. 2005;40(6):575-577.
14. Gajda M, Sedlaczek K, Szemik S, Kowalska M. Determinants of Alcohol Consumption among Medical Students: Results from POLLEK Cohort Study. Int J Environ Res Public Health. 2021;18(11):5872.
15. Gaviria-Criollo CA, Martínez-Porras DA, Arboleda-Castillo AF, Mafla AC. Consumo de alcohol en estudiantes de medicina en Pasto (Colombia). Rev Salud Uninorte. 2015;31(3):458-466.
16. Zhang C, Yang G, Li Z, Li X, Li Y, Hu J, et al. Reliability and validity of the Chinese version on Alcohol Use Disorders Identification Test. Chinese J Endemiol. 2017;38:1064-1067.
17. Saunders JB, Aasland OG, Amundsen A, Grant M. Alcohol consumption and related problems among primary health care patients: WHO collaborative project on early detection of per-sons with harmful alcohol consumption--I. Addiction. 1993;88(3):349-362.
18. Rubio G, Bermejo J, Caballero M. Validación de la prueba para la identificación de trastornos por uso de alcohol (AUDIT) en atención primaria. Rev Clin Esp. 1998;198(1):11-14.
19. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1):1-55.
20. Fornell C, Larcker DF. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. J Marketing Reearch. 1981;18(1):39-50.
21. Henseler J, Ringle C, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. J Acad Marketing Science. 2015;43(1):115-135.
22. Nunnally J, Bernstein I. Psychometric theory. 3rd edition. New York: McGraw-Hill; 1994.
23. Chen F. Sensitivity of Goodness of Fit In-dexes to Lack of Measurement Invariance. Structural Equation Modeling. 2007;14:464-504.
24. Davoren MP, Shiely F, Byrne M, Perry IJ. Hazardous alcohol consumption among univer-sity students in Ireland: A cross-sectional study. BMJ Open. 2015;5(1):e006045.
25. Moreta-Herrera R, Rodas JA, Lara-Salazar M. Factor Validity of Alcohol Use Disorders Identification Test (AUDIT) Using Robust Estimations in Ecuadorian Adolescents. Alcohol Alcohol. 2021;56(4):482-489.
26. Tavernier R, Willoughby T. Bidirectional associations between sleep (quality and duration) and psychosocial functioning across the university years. Dev Psychol. 2014;50(3):674-482.
27. White A, Hingson R. The Burden of Alcohol Use. Alcohol Res. 2014;35(2):201-218.
28. Yeh MY, Wu CI, Shih YH, Chen YK. A Cognitive Model of Alcohol Use Among Taiwanese Adolescents: The Influence of Alcohol Expectancies and Drinking Refusal Self-Efficacy. Healthcare. 2025;13(22):2981.
29. Machado PMA, Campelo CL, Oliveira JVPD, Batista RFL, Simões VMF, Santos AMD. Análise da estrutura fatorial do Audit em adolescentes entre 18 e 19 anos. Rev Saúde Pública. 2021;55:27.
30. Carretero MÁG, Ruiz JPN, Delgado JMM, González CO. Validation of the Alcohol Use Disorders Identification Test in university students: AUDIT and AUDIT-C. Adicciones. 2016;28(4):194-204.
31. Tanudjaja SA, Chih H, Burns S, Crawford G, Hallett J, Jancey J. Alcohol consumption and associated harms among university students in Australia: Findings from a cross-sectional study. Health Promot J Austr. 2021;32(2):258–63.
32. Swahn MH, Bossarte RM, Choquet M, Hassler C, Falissard B, Chau N. Early substance use initiation and suicide ideation and attempts among students in France and the United States. Int J Public Health. 2012;57(1):95-105.
33. Cordero-Oropeza R, García-Méndez M, Cordero-Oropeza M, Corona-Maldonado JJ. Characterization of alcohol consumption and related problems in university students from Mexico City. Salud Mental. 2021;44(3):107-115.
34. Miller M, Borges G, Orozco R, Mukamal K, Rimm EB, Benjet C, et al. Exposure to alco-hol, drugs and tobacco and the risk of subsequent suicidality: Findings from the Mexican Adolescent Mental Health Survey. Drug Alcohol Depend. 2011;113(2-3):110-117.
35. Hair J, Hult G, Ringle C. A primer on partial least squares structural equation modeling. 2nd edition. Thousand Oaks, CA: Sage.; 2017.
36. Kugbey N, Manortey S, Dziwornu E, Kyei-Arthur F, Boateng MO, Kushitor SB, et al. Alcohol use among adolescents in eight sub-Saharan African countries: Evidence from the Global School-based student health survey (2012–2017) using the socio-ecological model. BMC Psychiatry. 2025;25:1080.
37. Andrade AG de, Duarte P do CAV, Barroso LP, Nishimura R, Alberghini DG, Oliveira LG de. Use of alcohol and other drugs among Brazilian college students: Effects of gender and age. Braz J Psychiatry. 2012;34(3):294-305.
38. Reinert DF, Allen JP. The alcohol use disorders identification test: An update of research findings. Alcohol Clin Exp Res. 2007;31(2):185-199.
39. Demartini KS, Carey KB. Optimizing the use of the AUDIT for alcohol screening in college students. Psychol Assess. 2012;24(4):954-963.

40. Kypri K, Vater T, Bowe SJ, Saunders JB, Cunningham JA, Horton NJ, et al. Web-based alcohol screening and brief intervention for university students: A randomized trial. JAMA. 2014;311(12):1218-1224.

41. Slade T, Chapman C, Swift W, Keyes K, Tonks Z, Teesson M. Birth cohort trends in the global epidemiology of alcohol use and alcohol-related harms in men and women: Systematic review and metaregression. BMJ Open. 2016;6(10):e011827.

42. Room R, Babor T, Rehm J. Alcohol and public health. Lancet. 2005;365(9458):519-530.