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.

Measuring What Future Physicians Know About Contraception: Development and Validation of the Clinical Contraceptive Knowledge Test (CCKT) in Vietnam Using Item Response Theory

Nguyen Dinh Ky
Hanoi Medical University
Nguyen Quynh Anh
Hanoi Medical University
Nguyen Ngoc Minh
Hanoi Medical University
Pham The Lam
Hanoi Medical University
Do Hoa Cham
Hanoi Medical University
To Khanh Linh
Hanoi Medical University
Nguyen Dang Linh Chi
Hanoi Medical University
Pham Hong Anh
Hai Phong University of Medicine and Pharmacy
Nguyen Thi Hong Hanh
Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen
Nguyen Tien Thang
The University of Texas Medical Branch
Nguyen Thi Huyen Anh
Hanoi Medical University
Nguyen Manh Thang
manhthang@hmu.edu.vn

Hanoi Medical University

Research Article

Keywords: Contraceptive knowledge, Medical students, Item Response Theory, Validation, Medical education

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10109841/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

Accurate measurement of contraceptive knowledge among medical students remains a critical challenge due to limitations of traditional assessment tools. This study aimed to develop and validate the Clinical Contraceptive Knowledge Test (CCKT) using a modern psychometric framework.

Methods

A cross-sectional multicenter study was conducted among undergraduate medical students in Vietnam (723 recruited; 636 retained for analysis). The instrument was developed through four sequential phases: (1) Item generation based on Bloom's taxonomy and WHO Medical Eligibility Criteria (MEC), (2) expert review for content validity, (3) pilot testing using Classical Test Theory, and (4) large-scale validation using Item Response Theory. A two-parameter logistic (2PL) IRT model was applied to evaluate item characteristics, dimensionality, and measurement precision.

Results

The final instrument, CCKT-18, demonstrated strong psychometric performance. A unidimensional 2PL model was retained based on comparable fit to multidimensional models and high inter-factor correlations ($r > 0.80$). Model fit indices indicated good fit (CFI = 0.983, TLI = 0.980, RMSEA = 0.036, SRMR = 0.039). The scale showed high measurement precision across a broad range of ability levels, with peak information around the average proficiency level (marginal reliability = 0.844). Item parameters were stable in sensitivity analyses (ΔCFI = 0.003), supporting model robustness. DIF analyses supported measurement invariance across sex, although several items demonstrated DIF across the training stage.

Conclusion

The CCKT-18 is a valid, reliable, and robust instrument for assessing contraceptive knowledge among medical students. Its IRT-based framework enables precise, competency-oriented evaluation and provides actionable insights for targeted curriculum improvement in reproductive health education.

1. Background

Contraception is a critical primary prevention strategy, and equitable access to contraceptive care is recognized as a fundamental human right.[1] However, the world currently faces severe reproductive health challenges, with approximately 73 million abortions occurring annually, accounting for 61% of all unintended pregnancies and 29% of total pregnancies globally.[2] Notably, Vietnam has the highest

abortion rate in Asia and ranks second globally, with 53.6% of these cases attributable to unintended pregnancies.[3] Against this backdrop, high-quality contraceptive counseling serves as a core intervention - a psychosocial support process that empowers patients to make evidence-based decisions. Empirical evidence confirms that effective counseling improves patient adherence and reduces the incidence of unintended pregnancies. To standardize this practice, the World Health Organization (WHO) introduced the Medical Eligibility Criteria for Contraceptive Use (MEC), guiding healthcare providers in selecting safe, individualized contraceptive methods.[4]

Medical students are the future nucleus of the reproductive healthcare workforce. Multinational studies from India and Malaysia to Egypt indicate that, despite excelling academically, medical students consistently lack practical knowledge of modern contraceptives and the clinical skills to apply WHO MEC standards.[5–7] In Vietnam, while family planning counseling is a mandatory graduation competency, students exhibit profound practical knowledge gaps. A recent Vietnamese survey indicated that only 52.5% of health science students possess adequate contraceptive knowledge.

Quantifying these educational gaps to inform curriculum reform requires precise measurement instruments. However, legacy tools, such as the Contraceptive Knowledge Assessment (CKA), exhibit significant methodological limitations.[8] Predicated solely on Classical Test Theory (CTT), these scales evaluate competency by aggregating raw scores, which assume equal item weighting and yield an imprecise measure of latent clinical competence. Furthermore, modern psychometric techniques such as Item Response Theory (IRT), which can independently estimate item difficulty and discrimination to provide highly precise assessments of student ability, remain notably underutilized in medical education research in this region. To the best of our knowledge, no validated IRT-based instrument for assessing contraceptive knowledge among medical students currently exists, either in Vietnam or internationally.

Therefore, our study aimed to develop and validate the Clinical Contraceptive Knowledge Test (CCKT) through a sequential psychometric framework, integrating CTT for preliminary screening with IRT for advanced validation. Both unidimensional and multidimensional Two-Parameter Logistic Item Response Theory (2PL-IRT) model specifications were systematically evaluated to empirically identify the most parsimonious and psychometrically robust structure. This approach enables precise, competency-oriented assessment of contraceptive knowledge and may help identify specific curricular deficits to support targeted medical education reforms.

2. Methods

2.1 Study Design

The present research was conducted as part of the broader V-COC Study (Readiness and determinants of contraceptive counseling behavior among Vietnamese medical students: A nationwide study). This study employed a cross-sectional, multicenter design to develop and validate the CCKT among undergraduate medical students in Vietnam, conducted from 31 August 2025 to 20 April 2026. The study

is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines. The validation process comprised four sequential phases: (1) item generation, (2) expert review for content validity, (3) pilot testing, and (4) large-scale psychometric validation.

2.2 Study Setting and Participants

Five medical universities in Vietnam participated in the study: Hanoi Medical University (main campus and Thanh Hoa campus), Thai Nguyen University of Medicine and Pharmacy, Hai Phong University of Medicine and Pharmacy, and the Faculty of Medicine at Phenikaa University. These institutions were selected to achieve geographic diversity across northern Vietnam and to represent variation in institutional size and student profiles.

At each participating institution, a designated on-site research contact - typically a faculty member or senior student researcher affiliated with the study team - coordinated in-person data collection. The research contact visited classrooms across all six academic years (Year 1 through Year 6) of the undergraduate Doctor of Medicine program, during which students completed a structured self-administered questionnaire hosted on the KoBoToolbox platform on their personal devices. Participants were recruited through convenience sampling. Each account was restricted to a single submission to prevent duplicate responses.

Eligibility criteria were: (i) current enrollment as an undergraduate student in a Doctor of Medicine program at a participating institution, (ii) provision of electronic informed consent, and (iii) complete questionnaire submission. Participants were excluded if they had incomplete responses, exhibited invalid response patterns (e.g., completion time under 3 minutes or failed attention-check items), or had missing key demographic data.

2.3 Instrument Development

Phase 1 - Item generation. An initial 27-item knowledge scale was developed through a structured literature review, with two researchers independently searching PubMed, Scopus, and Embase. Items were grounded in two WHO evidence-based guidelines[4,9] and mapped onto Bloom's Revised Taxonomy across three cognitive levels: Recall, Understanding, and Application. Items were formatted as 'True/False/I don't know' to minimize guessing-induced measurement error. The full English version of the developed questionnaire is provided in Supplementary Table S6.

Phase 2 - Expert review and content validity. Content validity was assessed by a panel of six specialists in reproductive health, medical education, and psychometrics. Each item was rated on a 4-point Likert scale for relevance and clarity. Agreement was quantified using the Item-Level Content Validity Index (I-CVI), Scale-Level Content Validity Index (S-CVI/Ave), and the modified Cohen's Kappa (κ*) to adjust for chance agreement, following Polit et al. and Lynn.[10,11] Based on quantitative metrics (I-CVI threshold: 0.83) and qualitative panel feedback, one item was removed and two were revised in accordance with WHO guidelines, yielding a 26-item instrument.

Phase 3 - Pilot testing. A pilot study was conducted among 100 undergraduate medical students at Hanoi Medical University (isolated from the main validation sample). Classical Test Theory (CTT) analyses — item difficulty, discrimination, item-total correlations, and KR-20 internal consistency — were performed as an initial screening reference. The pilot demonstrated adequate preliminary reliability (KR-20 = 0.854). These CTT results informed subsequent IRT-based refinement but did not serve as definitive retention criteria.

Main Validation (Phase 4)

Of 723 responses collected, 636 were retained for the main analytical dataset following removal of 87 responses that failed attention checks (n = 71) or had a completion time of less than 3 minutes (n = 16). Response coding assigned 1 to correct answers and 0 to both incorrect and 'I don't know' responses to penalize guessing. A secondary strict-clean dataset (n = 545) was derived by additionally excluding responses with low response variation, multivariate outliers (Mahalanobis distance), long string or straightlining patterns, and extreme total scores (Fig. 1). The primary model was developed on the full dataset; the strict-clean dataset was reserved for sensitivity analyses.

2.4 Statistical Analyses

All analyses were performed in R version 4.6.0. CTT pilot analyses were conducted in SPSS. IRT analyses used the mirt package with marginal maximum likelihood (MML) estimation for unidimensional models and the quasi-Monte Carlo expectation-maximization (QMCEM) method for multidimensional models. Parallel analysis used tetrachoric-correlation-based simulation via the psych package.

Dimensionality assessment. Parallel analysis on a tetrachoric correlation matrix was performed to determine the number of latent factors. Scree plots and eigenvalue ratios were used to evaluate the dominant-factor hypothesis.

IRT model selection and item reduction. A two-parameter logistic (2PL) model was applied to estimate item discrimination (a) and difficulty (b) parameters. To identify the most parsimonious representation, seven competing models were evaluated across both the 26-item and 18-item versions, spanning unidimensional (1D) and multidimensional (2D, 3D, 4D) IRT specifications. Model selection followed the parsimony principle: the simplest model with fit indices comparable to those of more complex specifications was retained. Items were iteratively removed if they failed any of the following criteria: (i) discrimination $a < 0.80$, (ii) item-level misfit by $S\text{-}X^2$ statistic ($p < 0.05$), or (iii) local dependence by Yen's Q3 ($|Q3| > 0.20$).

Model fit evaluation. Model fit was assessed using the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Standardized Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA), with thresholds of CFI/TLI > 0.95, SRMR < 0.10, and RMSEA < 0.10. Given the binary item format and relatively low degrees of freedom, a relaxed RMSEA threshold (< 0.10) was applied to account for known upward bias in these conditions.

Differential item functioning (DIF). DIF was evaluated across sex (male vs. female) and training stage (preclinical vs. clinical) using likelihood-ratio tests within a multi-group 2PL IRT framework, with Benjamini–Hochberg false discovery rate (FDR) correction applied for multiple comparisons. In the IRT framework, item-level DIF analysis serves as the functional analog of measurement invariance testing in factor-analytic approaches, evaluating whether items function equivalently across subgroups. Measurement precision was examined via the Test Information Function (TIF) and Wright map.

Sensitivity analysis. The 2PL model was re-estimated on the strict-clean dataset, and changes in model fit (ΔCFI) and item parameters (Δa, Δb) were assessed. Stability of parameter rank ordering was interpreted as evidence of structural robustness.

3. Results

3.1 Sample characteristics

A total of 636 participants were included in the main analytical dataset (723 recruited; 87 excluded for failed attention checks [n = 71] or short completion time [n = 16]). The strict-clean dataset retained 545 participants after further removal of low-quality responses. The two samples were comparable across all demographic and outcome variables (all $p > 0.05$; Table 1). Females comprised 55.6% of the main sample; 68.2% were in the preclinical stage. The mean raw knowledge score was 15.43 ± 4.97 out of 26 (Table 1).

Table 1
Demographic Characteristics and Raw Score Distribution

Characteristic	Main dataset N = 636	Strict-clean dataset N = 545	p-value[1]
Sex, n (%)			0.352
Male	282 (44.4%)	226 (41.5%)	
Female	353 (55.6%)	318 (58.5%)	
Academic year, n (%)			0.932
Year 1	150 (23.6%)	122 (22.4%)	
Year 2	176 (27.7%)	141 (25.9%)	
Year 3	108 (17.0%)	96 (17.6%)	
Year 4	102 (16.0%)	91 (16.7%)	
Year 5	48 (7.5%)	48 (8.8%)	
Year 6	52 (8.2%)	47 (8.6%)	
Training stage, n (%)			0.423
Preclinical	434 (68.2%)	359 (65.9%)	
Clinical	202 (31.8%)	186 (34.1%)	
Institution, n (%)			0.852
Hanoi Medical University	221 (34.7%)	174 (31.9%)	
Hanoi Medical University, Thanh Hoa campus	93 (14.6%)	81 (14.9%)	
Thai Nguyen University of Medicine and Pharmacy	178 (28.0%)	166 (30.5%)	
Hai Phong University of Medicine and Pharmacy	95 (14.9%)	83 (15.2%)	
Faculty of Medicine, Phenikaa University	49 (7.7%)	41 (7.5%)	
Residence, n (%)			0.546
Urban	297 (46.7%)	244 (44.8%)	
Rural	339 (53.3%)	301 (55.2%)	
Religion, n (%)			> 0.999

[1]Pearson's Chi-squared test; Welch Two Sample t-test

Characteristic	Main dataset N = 636	Strict-clean dataset N = 545	p-value[1]
Yes	18 (2.8%)	15 (2.8%)	
No	618 (97.2%)	530 (97.2%)	
Ethnicity, n (%)			0.775
Kinh	544 (85.5%)	462 (84.8%)	
Other	92 (14.5%)	83 (15.2%)	
History of sexual experience, n (%)			0.953
Yes	128 (20.1%)	108 (19.8%)	
No	508 (79.9%)	437 (80.2%)	
History of using contraception, n (%)			0.890
Yes	128 (20.1%)	107 (19.6%)	
No	508 (79.9%)	438 (80.4%)	
Received SRH education, n (%)			0.735
Yes	492 (77.4%)	427 (78.3%)	
No	144 (22.6%)	118 (21.7%)	
Completed Ob/Gyn clinical rotation, n (%)			0.513
Yes	223 (35.1%)	202 (37.1%)	
No	413 (64.9%)	343 (62.9%)	
Self-rated confidence in contraceptive knowledge (0–10)			0.704
Mean ± SD	7.11 ± 1.77	7.15 ± 1.70	
Median (Q1, Q3)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	
Raw knowledge score (0–26)			0.520
Mean ± SD	15.43 ± 4.97	15.62 ± 5.04	
Median (Q1, Q3)	15.00 (12.00, 18.00)	15.00 (12.00, 19.00)	

[1]Pearson's Chi-squared test; Welch Two Sample t-test

3.2 Classical Test Theory analyses

Internal consistency was good in both samples (KR-20 = 0.839 and 0.845 in the main and strict-clean datasets, respectively).

3.3 Dimensionality

Parallel analysis identified seven eigenvalues exceeding 1.0 (8.804, 2.170, 1.681, 1.454, 1.285, 1.115, 1.051; Table S3), with the first eigenvalue accounting for 33.9% of variance - markedly larger than subsequent factors. A clear scree plot elbow after the first factor (Figure S1) and the dominant variance ratio collectively supported an essentially unidimensional structure, with minor secondary components reflecting local content variation rather than distinct latent dimensions.

3.4 IRT model selection and item reduction

Seven competing IRT models - spanning unidimensional (1D) and multidimensional (2D, 3D, 4D) specifications - were evaluated across both the 26-item and 18-item versions (Table 2). For the 26-item scale, the 3D MIRT and 1D models showed virtually identical fit (CFI 0.944 vs 0.940; RMSEA 0.048 vs 0.049). For the 18-item version, the 3D, 2D, and 1D models were again indistinguishable (CFI 0.983; RMSEA 0.036), and inter-factor correlations in the multidimensional solution were high (r = 0.802–0.974). Applying the parsimony principle, the reduced 18-item unidimensional 2PL model (Model 7) was retained as the final instrument (CFI = 0.983, TLI = 0.980, RMSEA = 0.036, SRMR = 0.039).

Table 2
Fit indices of the Item Response Theory (IRT) models

Model	χ^2/df	M2	df	*p*-value	RMSEA (95% CI)	SRMR	CFI	TLI
Model 1: MIRT 4D (26 items)	5.48	1,637.183	299	< 0.001	0.084 (0.080–0.088)	0.162	0.823	0.807
Model 2: MIRT 3D (26 items)	2.43	720.562	296	< 0.001	0.048 (0.043–0.052)	0.050	0.944	0.938
Model 3: MIRT 2D (26 items)	3.49	1,044.538	299	< 0.001	0.063 (0.059–0.067)	0.143	0.901	0.893
Model 4: IRT 1D (26 items)	2.51	751.156	299	< 0.001	0.049 (0.044–0.053)	0.051	0.940	0.935
Model 5: MIRT 3D (18 items)	1.84	242.230	132	< 0.001	0.036 (0.029–0.043)	0.039	0.983	0.980
Model 6: MIRT 2D (18 items)	1.81	242.692	134	< 0.001	0.036 (0.028–0.043)	0.039	0.983	0.980
Model 7: IRT 1D (18 items)	1.82	245.099	135	< 0.001	0.036 (0.029–0.043)	0.039	0.983	0.980

Note. MIRT = multidimensional item response theory; IRT = item response theory. Model 1 (MIRT 4D) represented a four-domain structure based on content areas: B1 = General reproductive knowledge; B2 = Contraceptive methods; B3 = Mechanisms of action; and B4 = Clinical/WHO Medical Eligibility Criteria (MEC). Models 2 and 5 (MIRT 3D) represented a three-factor structure based on Bloom's taxonomy domains: Recall, Understanding, and Application. Models 3 and 6 (MIRT 2D) represented a two-factor cognitive structure in which Recall and Understanding were combined as lower-order cognition, while Application represented higher-order cognition. Models 4 and 7 (IRT 1D) were unidimensional IRT models that assumed a single latent trait.

Item reduction from 26 to 18 items proceeded iteratively: six items were removed for poor discrimination ($a < 0.80$; Table S4), and two for local dependence (B.2.1–B.3.4: $|Q3| = 0.272$; B.4.8–B.4.7: $|Q3| = 0.456$; Table S5). Among the 18 retained items, three required additional consideration: item B.2.6 had discrimination marginally below threshold ($a = 0.687$); items B.3.3 and B.3.5 showed emergent misfit in the 18-item model (S-X^2 $p = 0.006$ and 0.044, respectively) despite adequate fit in the 26-item pool; and item pair B.3.2–B.3.6 marginally exceeded the local independence threshold ($|Q3| = 0.203$). These items were retained on substantive grounds; the rationale is discussed below.

Item parameters for the CCKT-18 are shown in Table 3. Discrimination ranged from 0.687 (B.2.6) to 2.374 (B.3.2), and difficulty from − 3.86 (B.3.4) to 1.08 (B.2.3).

Table 3
IRT 2PL model parameters for the CCKT-18

Item	Discrimination	Difficulty	S-X2	p S-X2
B.1.6	0.805	-3.342	15.875	0.146
B.2.2	2.155	0.735	19.028	0.088
B.2.3	1.605	1.081	15.773	0.202
B.2.5	0.913	-0.433	11.205	0.670
B.2.6	0.687	-1.615	20.927	0.104
B.2.7	1.408	0.801	13.305	0.425
B.3.1	1.971	-0.539	9.624	0.474
B.3.2	2.374	0.589	7.453	0.761
B.3.3	1.675	0.432	29.309	0.006
B.3.4	0.895	-3.860	7.069	0.630
B.3.5	2.175	0.379	20.091	0.044
B.3.6	1.402	0.231	16.200	0.239
B.4.1	1.270	-0.312	6.287	0.935
B.4.2	1.844	-0.007	4.512	0.952
B.4.3	2.259	0.744	9.646	0.562
B.4.4	1.665	0.094	9.148	0.690
B.4.6	1.718	0.443	7.162	0.894
B.4.7	1.423	-0.258	13.823	0.243

The person–item map demonstrated good alignment between student ability and item difficulty in the final 18-item mode (median $\theta = -0.03$; Fig. 2).

3.5 Differential item functioning

No items demonstrated significant DIF across sex after Benjamini–Hochberg FDR correction (all adjusted $p > 0.05$), supporting measurement invariance between male and female participants. Four items showed significant DIF across the training stage: B.2.2 (adjusted $p = 0.026$), B.2.3 ($p < 0.001$), B.2.7 ($p < 0.001$), and B.4.2 ($p = 0.049$) - all involving guideline-based clinical decision-making competencies (Table 4).

Table 4
Differential item functioning (DIF) analysis of the CCKT-18

Item	Gender			Training stage		
	X^2	Non-adjusted *p*-value	Adjusted *p*-value	X^2	Non-adjusted *p*-value	Adjusted *p*-value
B.1.6	4.19	0.123	0.325	0.19	0.909	0.963
B.2.2	2.21	0.331	0.663	10.90	0.004	0.026
B.2.3	0.28	0.868	0.894	18.30	0.000	0.001
B.2.5	4.42	0.110	0.325	4.19	0.123	0.317
B.2.6	1.01	0.602	0.774	1.69	0.429	0.536
B.2.7	8.51	0.014	0.135	21.69	0.000	0.000
B.3.1	1.68	0.431	0.705	4.63	0.099	0.297
B.3.2	1.50	0.473	0.709	0.20	0.905	0.963
B.3.3	8.41	0.015	0.135	2.81	0.246	0.491
B.3.4	1.31	0.520	0.720	6.95	0.031	0.112
B.3.5	1.97	0.373	0.672	1.74	0.419	0.536
B.3.6	6.28	0.043	0.260	1.61	0.446	0.536
B.4.1	5.03	0.081	0.325	3.32	0.190	0.427
B.4.2	0.22	0.894	0.894	9.04	0.011	0.049
B.4.3	3.28	0.194	0.437	1.81	0.405	0.536
B.4.4	0.33	0.847	0.894	2.37	0.305	0.500
B.4.6	0.47	0.791	0.894	2.50	0.286	0.500
B.4.7	4.14	0.126	0.325	0.07	0.967	0.967

3.6 Scale reliability and measurement precision

The CCKT-18 demonstrated good overall measurement precision (marginal reliability = 0.844). The TIF peaked at θ = 0.462 (information = 10.4, SE = 0.310, conditional reliability = 0.913). High precision (SE < 0.45; reliability ≥ 0.80) was maintained across θ = −0.784 to + 1.51, covering the ability range of most medical students (Fig. 3). Precision declined at extreme ability levels, consistent with the item difficulty ceiling (maximum b = 1.08).

3.7 Sensitivity analysis

Model fit was highly stable across datasets (ΔCFI = 0.003, well below the 0.01 threshold; Table 5).

Table 5
Sensitivity analysis of model fit indices for the CCKT-18

Model	χ^2/df	M2	df	*p*-value	RMSEA (95% CI)	SRMR	CFI	TLI
Main dataset	1.82	245.099	135	< 0.001	0.036 (0.029–0.043)	0.039	0.983	0.980
Strict-clean dataset	1.64	222.072	135	< 0.001	0.034 (0.026–0.042)	0.040	0.985	0.983

Item parameters showed negligible variation: most $|\Delta a| < 0.15$ and $|\Delta b| < 0.10$, with the largest shift observed for B.3.4 ($|\Delta b| = 0.408$). Parameter rank ordering was fully preserved across datasets (Table 6), confirming structural robustness.

Table 6
Sensitivity analysis of item parameters for the CCKT-18

Item	Discrimination			Difficulty		
	Main dataset	Strict-clean dataset	Index change	Main dataset	Strict-clean dataset	Index change
B.1.6	0.805	0.820	0.015	-3.342	-3.448	-0.107
B.2.2	2.155	2.310	0.155	0.735	0.661	-0.074
B.2.3	1.605	1.722	0.117	1.081	1.049	-0.033
B.2.5	0.913	0.860	-0.053	-0.433	-0.533	-0.100
B.2.6	0.687	0.681	-0.006	-1.615	-1.637	-0.023
B.2.7	1.408	1.403	-0.005	0.801	0.760	-0.041
B.3.1	1.971	1.949	-0.022	-0.539	-0.557	-0.018
B.3.2	2.374	2.691	0.317	0.589	0.513	-0.076
B.3.3	1.675	1.711	0.036	0.432	0.373	-0.059
B.3.4	0.895	1.136	0.241	-3.860	-3.452	0.408
B.3.5	2.175	2.321	0.146	0.379	0.338	-0.041
B.3.6	1.402	1.428	0.027	0.231	0.215	-0.016
B.4.1	1.270	1.388	0.118	-0.312	-0.323	-0.010
B.4.2	1.844	1.950	0.106	-0.007	-0.024	-0.017
B.4.3	2.259	2.531	0.273	0.744	0.724	-0.020
B.4.4	1.665	1.644	-0.021	0.094	0.051	-0.043
B.4.6	1.718	1.719	0.001	0.443	0.392	-0.051
B.4.7	1.423	1.551	0.128	-0.258	-0.306	-0.048

4. Discussion

4.1 Principal Findings

This study developed and validated the CCKT using an IRT framework, resulting in a refined 18-item unidimensional instrument (CCKT-18) derived from the original 26-item pool. The final scale demonstrated three core psychometric strengths: high measurement precision across a wide range of latent ability levels, strong structural validity, and robust model stability.[12] Specifically regarding structural validity, although exploratory analyses initially suggested multidimensionality, the presence of

a strong dominant factor, exceptionally high inter-factor correlations, and near-identical fit indices between the 3D and 1D models strongly justified adopting a parsimonious unidimensional structure.[13] By moving beyond traditional raw scores, the CCKT-18 provides a precise, clinically meaningful assessment of contraceptive knowledge, serving as a psychometrically sound tool for competency evaluation.

4.2 Structural Refinement and Item Reduction

The reduction from 26 to 18 items represents a deliberate psychometric optimization process rather than a limitation in data-capturing capacity. Notably, items were not excluded based on extreme difficulty levels; in fact, highly accommodating items (such as item B.3.4) were retained to assess baseline competencies and ensure comprehensive coverage across the ability continuum.[14] Instead, the elimination process strictly targeted items with low discrimination or violations of local independence, both of which compromise measurement efficiency in IRT-based instruments.[12] Eliminating these poorly discriminating and redundant items improved model performance and resolved local dependence issues that could otherwise distort parameter estimation. From an educational standpoint, this refinement reflects a shift from assessing superficial knowledge toward evaluating higher-order cognitive processes.[15] By balancing statistical rigor with substantive validity - allowing rare exceptions for marginal discrimination (item B.2.6) - the CCKT-18 aligns with competency-based frameworks. It effectively differentiates students capable of clinical application from those dependent on rote memorization.

4.3 Psychometric Flexibility and Content Validity

While IRT optimization typically eliminates extreme items, retaining B.1.6 and B.3.4, which exhibited extreme difficulty parameters in the full item pool ($b = -3.2$ and -3.732, respectively; Table S4), was a substantive decision. These items assess critical baseline competencies; removing them would truncate the lower end of the ability continuum and bias the instrument against lower-performing students. A robust scale must capture the full ability spectrum to identify foundational knowledge gaps.[12] Furthermore, item B.2.6 was retained despite the discrimination parameter falling marginally below the 0.80 threshold (Table S4). The expert panel determined that its essential clinical content outweighed the minor statistical deviation. In addition, the item pair B.3.2 and B.3.6 showed a marginal exceedance of the conventional local dependence threshold in the CCKT-18 ($|Q3| = 0.203$), although it remained acceptable in the original 26-item model. This slight increase likely reflects residual correlation shifts following scale reduction rather than meaningful content redundancy. Similarly, B.3.3 and B.3.5 exhibited emergent misfit ($p < 0.05$) in the CCKT-18, despite adequate fit in the original 26-item model (Table S4). This is likely a statistical artifact of scale reduction, where narrowed latent trait distributions heighten fit statistical sensitivity to non-substantive deviations. Ultimately, these decisions underscore the critical balance between empirical metrics and substantive judgment.

DIF analyses revealed no significant sex-based bias following false discovery rate adjustment, confirming measurement invariance. Conversely, significant DIF emerged across training stages,

specifically for items evaluating contraceptive method selection (B.2.2), copper IUD emergency contraception (B.2.3), missed pill management (B.2.7), and postpartum combined oral contraceptive eligibility (B.4.2). These topics require guideline-based clinical decision-making - competencies predominantly reinforced during clinical rotations. Thus, this DIF likely reflects true differences in curricular exposure rather than construct-irrelevant bias. Nonetheless, future studies should delineate the practical implications of these findings.

4.4 Dimensionality and Model Fit

Although initial theoretical alignment with Bloom's taxonomy and preliminary parallel analyses suggested a potential multidimensional structure (spanning recall, understanding, and application), subsequent empirical evaluation strongly supported a unidimensional framework for the final CCKT-18.[13]. Specifically, comparative fit indices demonstrated that the unidimensional 2PL model provided a parsimonious and highly adequate fit to the data, performing comparably to more complex multidimensional specifications.[12] Crucially, in the multidimensional models, the estimated latent traits exhibited exceptionally high inter-factor correlations ($r > 0.96$). Such extreme overlap indicates poor discriminability between the hypothesized subdomains, suggesting they capture a single, dominant latent construct rather than distinct dimensions. This empirical finding aligns closely with educational realities: while recall, understanding, and application are distinct in pedagogical theory, they operate as a highly integrated, indivisible cognitive competency when medical students conduct clinical contraceptive counseling. Consequently, the adoption of the parsimonious unidimensional model was both empirically justified and educationally sound, providing a robust representation of overarching contraceptive knowledge.

4.5 Model Robustness

A key strength of the CCKT-18 lies in its demonstrated structural stability and robustness. Sensitivity analyses comparing the main dataset with a strictly cleaned dataset confirmed the stability of model parameters and fit indices, with minimal changes observed ($\Delta CFI < 0.01$ and negligible variation in item parameters). This consistency demonstrates that the CCKT-18 is highly robust to data quality issues, such as inattentive responding and straightlining patterns, thereby enhancing confidence in its reliability, reproducibility, and generalizability across diverse educational settings.[16]

4.6 Comparison with Legacy Instruments and Alignment with International Trends

The superiority of the CCKT-18 over legacy instruments, such as the Contraceptive Knowledge Assessment (CKA), stems primarily from its advanced psychometric framework. While the CKA relies on CTT, which evaluates competency based on aggregated raw scores and assumes equal weighting for all items, this approach yields an imprecise measure of latent clinical competence.[8] In contrast, the CCKT-18 leverages a 2PL-IRT model to position students along a continuous latent ability scale (θ). By independently estimating item difficulty and discrimination, the CCKT-18 offers a highly precise, ability-

sensitive evaluation. This methodological shift allows educators to pinpoint specific cognitive deficits rather than simply generating an undifferentiated total score, providing actionable diagnostic insights that legacy tools fundamentally lack.

Furthermore, the development of the CCKT-18 contextualizes and advances observations from international medical education research. Previous multinational studies from India, Malaysia, and Egypt have consistently highlighted an educational paradox.[5,17,18] Despite excelling academically, medical and health science students often lack the practical knowledge and clinical skills needed to apply WHO MEC standards. Our findings closely align with this global trend; analysis of the CCKT-18's item parameters revealed that while Vietnamese students generally perform well on foundational recall items, they exhibit a relative deficit in higher-order application skills. Importantly, whereas earlier international research predominantly relied on traditional surveys or CTT-based assessments to report these gaps,[5,17,18] our study introduces a rigorously validated IRT framework. By precisely quantifying the discrepancy between theoretical knowledge and practical readiness, the CCKT-18 not only corroborates previous global findings but also offers a standardized methodological template that could be adapted for cross-cultural educational assessments.

4.7 Implications for Medical Education

Implementing the CCKT-18 has important implications for advancing medical education. By leveraging IRT-based scoring, the instrument enables measurement beyond students' raw scores, positioning students along a latent ability continuum (θ) and providing a more precise understanding of competency distribution. Analysis of the Wright map and TIF indicates that while students generally perform well on recall-based items, there is a relative deficit in higher-order application skills, highlighting a critical gap in current curricula. This level of diagnostic precision allows educators to move beyond generalized evaluation toward targeted instructional interventions, focusing on areas where students demonstrate the greatest need.[15] Consequently, the CCKT-18 supports a transition from traditional score-based assessment to competency-based evaluation, aligning with contemporary medical education frameworks that emphasize clinical reasoning and applied knowledge.

4.8 Strengths and Limitations

This study's principal strengths lie in its multicenter design, systematic four-phase instrument development, and the application of a rigorous IRT framework incorporating both unidimensional and multidimensional model comparisons, sensitivity analyses, and DIF testing - a combination rarely reported in medical education validation studies in this region.

However, several limitations should be considered. First, although the multicenter approach included multiple institutions, the reliance on convenience sampling may mean the sample does not fully represent all medical training environments nationwide. Second, the cross-sectional design limits the ability to assess longitudinal changes in students' knowledge. Third, the use of dichotomous response data limits the capacity to capture partial knowledge states, suggesting that future research could

benefit from incorporating polytomous response models. Fourth, given the sensitive nature of contraceptive topics, self-reported responses may be susceptible to social desirability bias, as participants might alter their answers due to concerns about social stigma. Fifth, criterion-related validity was not assessed, as the CCKT-18 was not compared against an established external benchmark such as another validated contraceptive knowledge measure or observed clinical performance. Finally, the current item set was concentrated in the moderate difficulty range, with the most difficult item posing only a limited challenge for respondents at higher ability levels. As a result, the instrument may provide less discrimination among top-performing students, and future revisions may benefit from adding more difficult items to improve measurement precision at the upper end of the trait continuum.

5. Conclusion

The CCKT-18 is a concise, valid, and reliable instrument for measuring contraceptive knowledge among undergraduate medical students, capturing a spectrum of competencies from basic recall to guideline-based clinical application. Longitudinal and criterion validity studies in independent samples are warranted to further establish its temporal stability and predictive utility, alongside the investigation of training-stage measurement differences identified in this study. Within the broader V-COC Study framework, the CCKT-18 will serve as a foundational assessment tool in a planned nationwide survey characterizing the contraceptive knowledge landscape among Vietnamese medical students. These data will equip medical educators with actionable evidence to evaluate and reform reproductive health curricula across institutions, ultimately contributing to a better-prepared future healthcare workforce.

Abbreviations

Abbreviation	Full Term
1D	Unidimensional
2PL	Two-parameter logistic
CCKT	Clinical Contraceptive Knowledge Test
CFI	Comparative Fit Index
CKA	Contraceptive Knowledge Assessment
CTT	Classical Test Theory
DIF	Differential item functioning
FDR	False discovery rate
I-CVI	Item-Level Content Validity Index
IRT	Item Response Theory
IUD	Intrauterine device
MEC	Medical Eligibility Criteria
MIRT	Multidimensional item response theory
MML	Marginal maximum likelihood
QMCEM	Quasi-Monte Carlo expectation-maximization
RMSEA	Root Mean Square Error of Approximation
S-CVI/Ave	Scale-Level Content Validity Index
SRMR	Standardized Root Mean Square Residual
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TIF	Test Information Function
TLI	Tucker-Lewis Index
V-COC	Readiness and determinants of contraceptive counseling behavior among Vietnamese medical students
WHO	World Health Organization

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board for Ethics in Biomedical Research (Approval ID: HMUIRB2446;

Decision No. 120/GCN-KHCN). Before accessing the questionnaire, all participants were provided with detailed study information on the survey's landing page and gave electronic informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

The authors declare that no funding was received for this study.

Authors' contributions

N.D.K., N.Q.A., and N.N.M. conceptualized the study, designed the methodology, performed the statistical and psychometric analysis (IRT modeling), and were major contributors in writing the original draft of the manuscript. P.T.L., D.H.C., T.K.L., and N.D.L.C. contributed to data collection, data curation, and critically reviewed the manuscript. P.H.A. and N.T.H.H. coordinated and facilitated the multicenter data collection at their respective institutions and assisted in data validation. N.T.T. and N.T.H.A. contributed to the expert review phase and methodology refinement. N.M.T. supervised the entire project, provided critical intellectual content, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the six specialists who contributed their expertise to the expert review panel during instrument development, whose critical appraisal of item relevance and clarity was instrumental in establishing the content validity of the CCKT. We also extend our sincere gratitude to all undergraduate medical students who voluntarily participated in this study. Their engagement and willingness to contribute time during an active academic period made this research possible.

Authors' information

Hanoi Medical University, Hanoi, Vietnam

Nguyen Dinh Ky, Nguyen Quynh Anh, Nguyen Ngoc Minh, Pham The Lam, Do Hoa Cham, To Khanh Linh, Nguyen Dang Linh Chi, Nguyen Thi Huyen Anh & Nguyen Manh Thang.

Hai Phong University of Medicine and Pharmacy, Hai Phong, Vietnam

Pham Hong Anh.

Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

Nguyen Thi Hong Hanh.

The University of Texas Medical Branch, Galveston, Texas, United States

Nguyen Tien Thang.

References

1. United Nations Population Fund. State of World Population 2012: By Choice, Not by Chance - Family Planning, Human Rights and Development. United Nations; 2012.
2. Bearak J, Popinchalk A, Ganatra B, et al. Unintended pregnancy and abortion by income, region, and the legal status of abortion: estimates from a comprehensive model for 1990–2019. Lancet Glob Health. 2020;8(9):e1152–61. 10.1016/S2214-109X(20)30315-6.
3. Hồ HN, Hoàng T, Phạm VH, Nguyễn THT. Khảo sát kiến thức, thái độ và hành vi của sinh viên khối ngành Khoa học sức khỏe tại Thành phố Hồ Chí Minh về các biện pháp tránh thai. HOC TP HO CHI MINH. 2025;28(9):152. 10.32895/hcjm.m.2025.09.18.
4. Special Programme of Research, Development, and Research Training in Human Reproduction (World Health Organization). World Health Organization, eds. Medical Eligibility Criteria for Contraceptive Use. Sixth edition. World Health Organization, Human Reproduction Programme;; 2025.
5. Hogmark S, Klingberg-Allvin M, Gemzell-Danielsson K, Ohlsson H, Essén B. Medical students' knowledge, attitudes and perceptions towards contraceptive use and counselling: a cross-sectional survey in Maharashtra, India. BMJ Open. 2013;3(12):e003739. 10.1136/bmjopen-2013-003739.
6. Htay MNN, Latt SS, Abas AL, Chuni N, Soe HHK, Moe S. Medical students' knowledge and perception toward family planning services: A preliminary intervention study. J Educ Health Promot. 2018;7:137. 10.4103/jehp.jehp_104_18.
7. Tang I, Rupley D. An Interactive Curriculum to Teach Person-Centered Contraceptive Counseling. MedEdPORTAL J Teach Learn Resour. 2023;19:11368. 10.15766/mep_2374-8265.11368.
8. Haynes MC, Ryan N, Saleh M, Winkel AF, Ades V. Contraceptive Knowledge Assessment: validity and reliability of a novel contraceptive research tool. Contraception. 2017;95(2):190–7. 10.1016/j.contraception.2016.09.002.
9. World Health Organization. Family Planning: A Global Handbook for Providers. 3rd ed. Geneva: World Health Organization; 2022.

10. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459–67. 10.1002/nur.20199.
11. Lynn MR, Determination, Quantification Of Content Validity. Nurs Res. 1986;35(6):386. 10.1097/00006199-198611000-00017.
12. Embretson SE, Reise SP. Item Response Theory for Psychologists. Psychology; 2013. 10.4324/9781410605269.
13. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychol Assess. 2000;12(3):287–97. 10.1037/1040-3590.12.3.287.
14. Baker FB. The Basics of Item Response Theory. 2nd ed. ERIC Clearinghouse on Assessment and Evaluation; 2001.
15. Gogus A. Bloom's Taxonomy of Learning Objectives. Encyclopedia of the Sciences of Learning. Boston, MA: Springer; 2012. pp. 469–73. 10.1007/978-1-4419-1428-6_141.
16. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods. 2012;17(3):437–55. 10.1037/a0028085.
17. Azzam H, Kamal B, Elsayed A, et al. Knowledge and attitude on family planning among medical students in Egypt: a multicentric cross-sectional study. BMC Public Health. 2025;25(1). 10.1186/s12889-024-20827-9.
18. Elkalmi RM, Khan MU, Ahmad A, et al. Knowledge, awareness, and perception of contraception among senior pharmacy students in Malaysia: A pilot study. J Res Pharm Pract. 2015;4(2):94–8. 10.4103/2279-042X.155760.

Figures

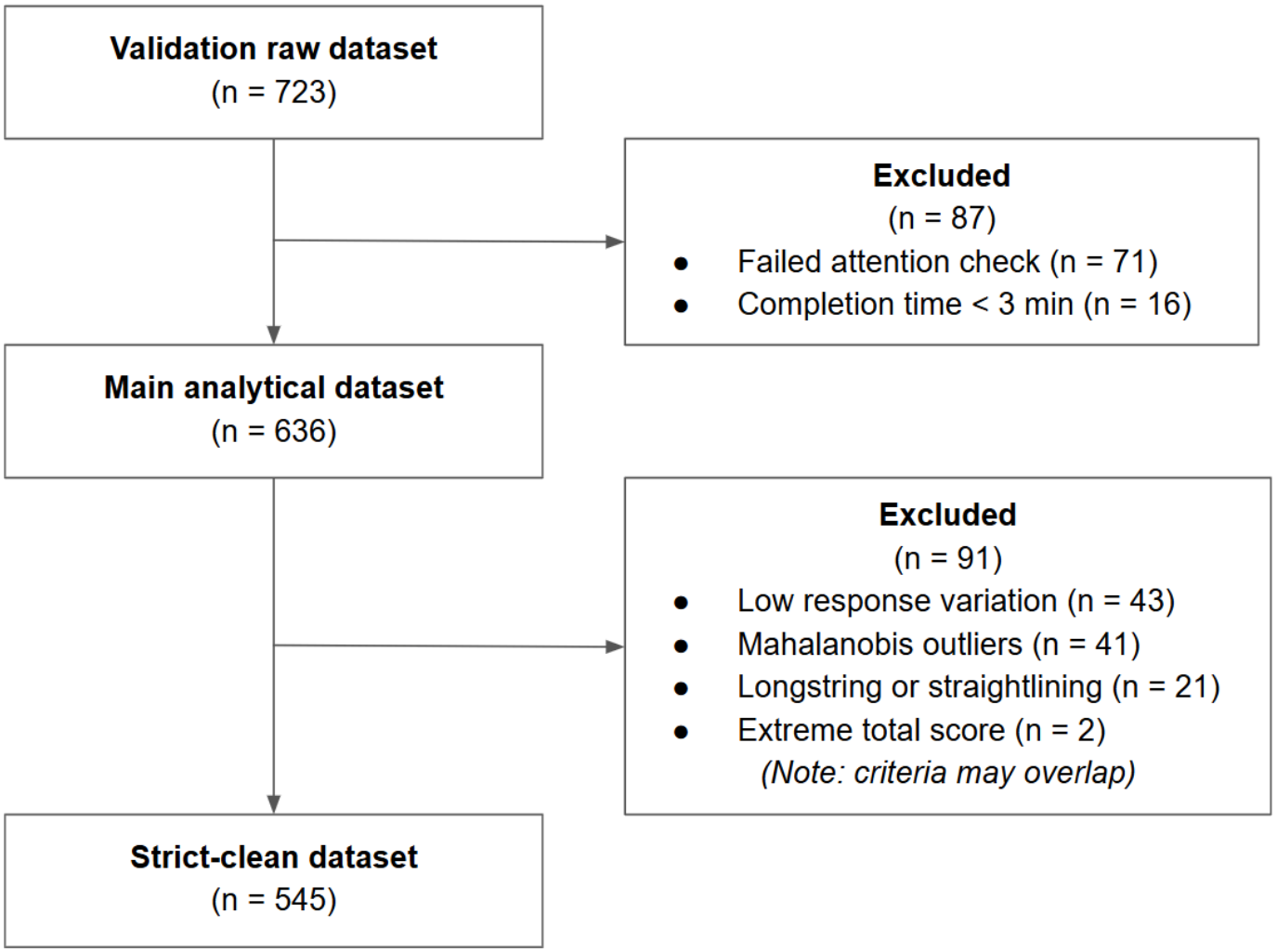


Figure 1

Flowchart of the data screening process for deriving the main and strict-clean datasets

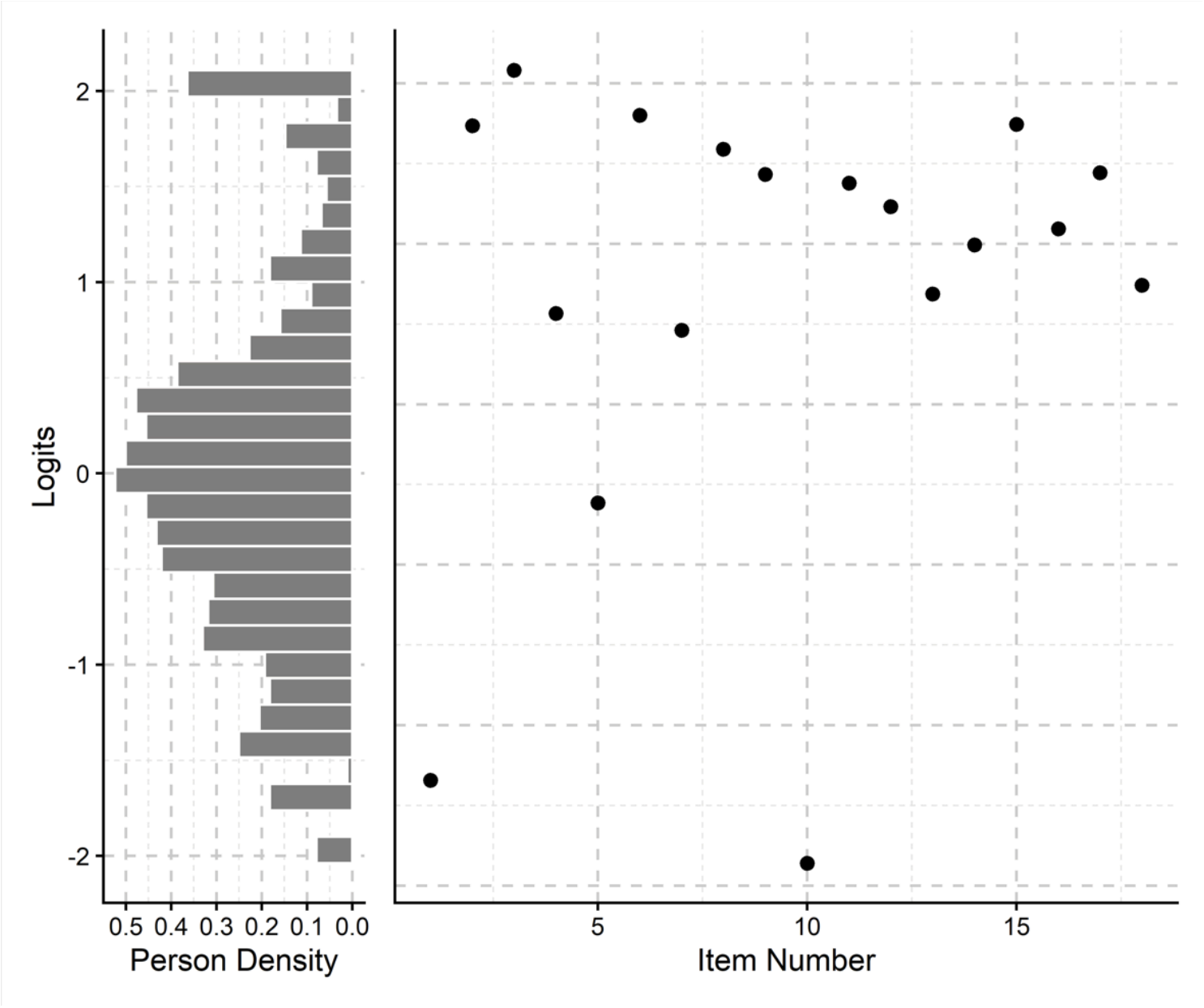


Figure 2

Person-item map for the CCKT-18

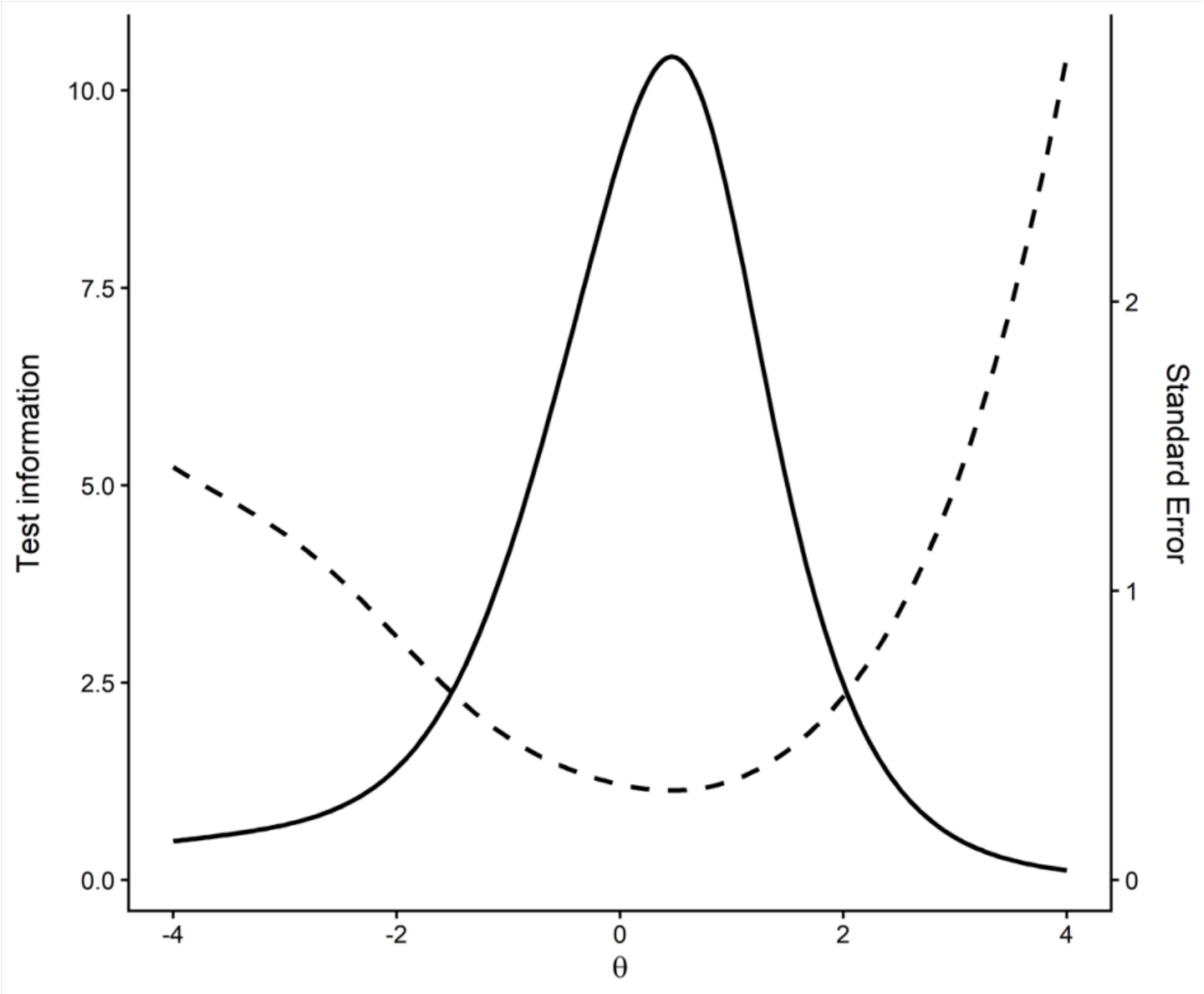


Figure 3

Test Information Function (TIF) and conditional reliability of the CCKT-18

Supplementary Files

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

- TableS1.ContentvalidityindicesandKappastatisticsfortheCCKT27Phase2.xlsx
- TableS2.ClassicalTestTheoryCTTItemStatisticsforthePilotPhase.xlsx
- TableS3.EigenvaluesandVarianceExplainedforDimensionalityAssessmentoftheScale.xlsx
- TableS4.ItemParametersFitStatisticsandSelectionOutcomesforthe26itemInstrument.xlsx
- TableS5.YensQ3ResidualCorrelationMatrixforAssessingLocalItemDependence.xlsx

- TableS6.itemblueprintandanswerkeyforthe27itemcckitempool
- FigureS1.ParallelAnalysisScreePlotforthe26itemScale.tiff
- STROBEChecklist.docx
- SupplementaryMaterials.docx