

Global Validation and Cognitive Diagnosis Evaluation of the Pharmacology Concept Inventory: Multinational, Multi-Stage Psychometric and Invariance Analysis

Adeladlew K. Netere

Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Melbourne, Parkville, VIC 3052, Australia

Tony Hughes

Faculty of Pharmacy and Pharmaceutical Sciences, Melbourne, Monash University, Parkville, Victoria, Australia.

Clare Guilding

School of Medicine, Faculty of Medical Science, Newcastle University, UK

Anna-Marie Babey

University of New England

Willmann Liang

University of Hong Kong

Lynette B. Fernandes

The University of Western Australia

Shola Olafuyi

University of Nottingham

Shahida Mallah

University of Nottingham

Fabiana Caetano Crowley

Department of Physiology & Pharmacology, Western University, London, Ontario, Canada

Anita Woods

Western University

Carolina Restini

Department of Pharmacology & Toxicology, College of Osteopathic Medicine, Michigan State University, East Lansing, MI-USA

Farhan Ahmad Khan

Department of Pharmacology, JN Medical College, Aligarh Muslim University

Janet Mifsud

Department of Clinical Pharmacology and Therapeutics, Faculty of Medicine and Surgery, University of Malta, Msida, Malta

Jennifer Koenig

University of Nottingham

Lindsay Cormier

University of Kentucky

Jennelle Durnett Richardson

Indiana University School of Medicine

Mark Hernandez

East Tennessee State University

Kelly Karpa

East Tennessee State University

Anneke Houwelingen

Utrecht University

Patrik Aronsson

University of Gothenburg

Martin Hawes

University of Surrey

Wael Eldahshan

Southwestern Oklahoma State University

Reem Atawia

Southwestern Oklahoma State University

Sudhir Chandra Sarangi

All India Institute of Ayurveda, New Delhi

Tori Llewelyn

Flinders University

Ullamari Pesonen

University of Turku

Christine Edmead

University of Bath

Sarah Bailey

University of Bath

Lorena Dima

Transylvania University of Braşov

Anikó Görbe

Semmelweis University

Zsófia Onódi

Semmelweis University

Pál Riba

Semmelweis University

Zoltán Varga

Semmelweis University

Dax Vendrig

Utrecht University

Arno Werners

Washington State University

Marina Junqueira Santiago

Macquarie University

Graeme Sills

University of Glasgow

Paul J. White

paul.white@monash.edu

Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Melbourne, Parkville, VIC 3052, Australia

Research Article

Keywords: core concepts of pharmacology, pharmacology concept inventory, validation, psychometric properties

Posted Date: July 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10049511/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

The validity, psychometric robustness, and educational effectiveness of the Pharmacology Concept Inventory (PCI), including its ability to identify students' misconceptions, were assessed across diverse educational settings and student cohorts to evaluate its potential as a concept-based assessment tool in pharmacology education. The PCI comprised 24 multiple-choice items, developed through a triangulated strategy that integrated core pharmacology concepts, expert input, and empirical evidence, and was informed by documented student misconceptions. This multinational validation and effectiveness evaluation included 1211 students from 23 institutions across 12 countries. The analyses examined the instrument's structure, item- and test-level psychometric properties, and measurement invariance across demographic and educational subgroups. Factor analysis indicated that the PCI primarily measures a single underlying dimension of pharmacological proficiency, with some weak secondary multidimensional groupings of items. The PCI demonstrated consistently strong item discrimination and a moderate level of difficulty, with convergent findings across factor analysis, cognitive diagnostic assessment, and other psychometric approaches. Exploratory cognitive diagnostic estimates showed that mastery varied across concepts, with drug tolerance having the highest probability (68%) and the structural activity relationship the lowest (33%). Year of study and academic program were significant predictors of pharmacological proficiency, with early-year and medical students more likely to achieve mastery than late-year and non-medical students. The PCI also identified a proficiency distribution, providing insight into the extent of variability and the gaps in learners' understanding of core pharmacology. Overall, this study establishes the PCI as a valid, reliable, and diagnostically informative tool for assessing conceptual understanding and supporting evidence-based pharmacology education across international contexts.

Introduction

The importance of pharmacology education across medical and health care professional programs [1] underscores the need for a valid and reliable concept-based assessment tool to evaluate students' knowledge and understanding [2]. Consequently, a theoretically grounded, expert-reviewed, learner-centered concept assessment tool, the Pharmacology Concept Inventory (PCI), was developed [3]. This study examines the PCI, an instrument designed to support concept-based learning and teaching by providing a structured way to assess understanding, evaluate learning outcomes, and measure the effectiveness of pharmacology curriculum instruction. Grounded in established learning theories and informed by expert and student perspectives, the PCI offers insights into learners' understanding of core pharmacology concepts, thereby supporting learner-centered instructional practices that are vital to current pharmacology training. Furthermore, it aims to assess how students apply fundamental pharmacology concepts to explanations, reasoning and problem-solving. This focus on core concepts enhances the PCI's diagnostic usefulness for educators and researchers, supporting evidence-based, cross-institutional and multi-context curriculum planning, targeted instructional improvements, and the promotion of concept-focused assessment methods in pharmacology education [2].

Because the implementation and evaluation of concept-based education depend on valid measures of conceptual understanding, assessment tools such as concept inventories play an important role in supporting broader curricular development. Constructing an internationally accepted concept-based pharmacology curriculum requires systematically designed, misconception-informed, and multi-contextually validated educational resources and assessment tools [2, 4, 5]. These resources must align with internationally agreed-upon core concepts and sub-concepts of pharmacology to harmonize concept-based education across institutions, disciplines, and countries and to create a shared framework, learning expectations, and assessment standards. Analysis of the PCI has demonstrated high item quality, strong content validity, satisfactory item performance, and strong reliability in previous studies [3]. Learner interviews provided evidence that students interpreted and responded to the items as intended, helped identify both persistent and emerging misconceptions, and informed further expert-led refinement of the items [6]. To build on these findings and expand the instrument's utility, a broader approach to investigation was required. At present, the instrument lacks large-scale evaluation across multinational settings, diverse course structures, or varied pedagogical approaches. As international pharmacology educators continue to advance concept-based curricula that integrate related disciplines, promote cross-institutional alignment, and strengthen the application of pharmacological knowledge to professional practice [4, 5, 7], further evaluation of the PCI across diverse

educational contexts is critical. Consequently, to support benchmarking, identify misconceptions, assess student reasoning, and ensure consistent application of key pharmacology concepts globally, an international, multi-cohort evaluation of the PCI's structure, psychometric properties, diagnostic capacity, and measurement invariance is essential for establishing its stability, consistency, and educational impact across institutions. Such a large-scale evaluation would help assess whether the PCI is suitable for broader implementation and sustainable adoption as a standardised assessment resource in pharmacology education.

Robust empirical validation is essential to determine whether the PCI is an accurate and reliable tool for assessing students' conceptual understanding; without it, results may be compromised by measurement error or external influences, undermining the validity of the conclusions [8]. A comprehensive validation of the PCI requires integrating multiple analytical approaches, including evaluation of the internal structure (i.e., how items group to reflect underlying pharmacological concepts), psychometric analyses using classical test theory (CTT) and item response theory (IRT), diagnostic insight from cognitive diagnostic assessment (CDA), also known as the cognitive diagnostic model (CDM) [9], and assessment of consistency across student groups.

The two primary psychometric paradigms, CTT and IRT, were used to assess item performance, reliability, and validity. CTT emphasises score-based analysis and internal consistency, offering a straightforward, foundational assessment of overall reliability, but is constrained by its dependence on sample characteristics [10–12]. IRT models the relationship between individuals' latent traits, such as ability or knowledge, and their response probabilities [13]. This method addresses the limitations of the CTT framework by providing detailed, sample-independent item-level insights that accurately capture the relationship between an examinee's latent ability and their responses [14, 15]. It includes models such as the one-parameter logistic (1PL/Rasch), which estimates difficulty assuming equal discrimination [16]; the two-parameter logistic (2PL), which also estimates discrimination [17]; and the three-parameter logistic (3PL), which accounts for guessing and is suited to multiple-choice questions [18]. Although both CTT and IRT offer valuable insights into item performance and measurement quality, neither directly identifies individual misconceptions nor evaluates how well items reveal underlying conceptual misunderstandings. To address this gap, cognitive diagnostic models provide detailed information about learners' conceptual strengths, weaknesses, and misconceptions [19].

Furthermore, CDA strengthens this statistical foundation by linking item responses to specific knowledge attributes, providing detailed diagnostic data essential for targeted instruction [20, 21]. CDA comprises a family of psychometric models that classify examinees as having mastered or not mastered cognitive attributes [9, 22]. These attributes, categorical latent variables inferred from item response patterns, allow models to estimate response likelihoods based on unobservable (latent) categorical variables [9, 23]. Here, "attribute" denotes the cognitive skills or knowledge required to solve an item, and CDAs aim to deliver concept-level diagnostic feedback aligned with these skills [24]. When integrated into assessment tools, CDA provides detailed insights into learners' cognitive strengths and weaknesses [25], particularly regarding core pharmacology concepts [4, 5]. This approach enables nuanced analysis of conceptual understanding by pinpointing misconceptions and knowledge gaps that underlie students' responses [9, 19]. Collectively, these paradigms form a comprehensive framework that supports the PCI's reliability, validity, and diagnostic value. This integrated method aims to generate detailed, practical evidence to inform educational decisions, guide curriculum development, and enhance teaching and learning outcomes.

This study evaluated the PCI's validity, psychometric quality, and educational impact across diverse educational settings and student cohorts. Our primary goal was to assess its ability to detect learners' misconceptions and to evaluate their conceptual understanding of pharmacology. We administered the PCI to a large international student cohort and conducted psychometric analyses using CTT, IRT, and CDA. These analyses examined the instrument's internal structure, validity evidence, and overall effectiveness in identifying key patterns of student understanding and misconceptions. Ultimately, this study seeks to support the broader implementation of the PCI as a scalable tool for tracking learning outcomes, evaluating teaching effectiveness, and guiding curriculum development across diverse educational settings.

Methods

Research design

A quantitative, cross-sectional study design was used to evaluate the psychometric properties and educational impact of the PCI. The study focused on large-scale, multi-institutional instrument validation, including assessments of dimensionality, measurement characteristics, and structural validity.

Participants characteristics

This large-scale, multinational study, conducted in Hungary, India, the USA, Australia, Romania, Malta, Canada, the UK, Sweden, Finland, Germany, and the Netherlands, was administered via the Qualtrics platform (Qualtrics LLC, Version XM, 2026) across all participating institutions. Recruitment was facilitated by each institution's members through in-person briefings and online platforms, including announcements in course modules, classroom presentations, and workshops. Participants were recruited from health-professional and biomedical science programs across multiple countries. Eligibility was limited to students who had completed coursework on pharmacology concepts within the internationally recognized core-concept framework [4, 5]. As a result, the sample included a wide range of educational levels, from students who had recently studied pharmacology to those reflecting on concepts learned years earlier. Demographic data were collected via a standardized questionnaire in which participants provided details about their academic backgrounds. Given the diversity of nationalities and educational pathways, academic level (undergraduate or postgraduate) and program of study were analyzed as separate, independent variables. Participant classification was based on self-reported educational status within their institutional and national contexts, acknowledging that professional programs such as medicine, pharmacy, and veterinary medicine may be classified as either undergraduate- or graduate-entry, depending on the country and educational system, as well as professional students who completed pharmacology coursework and long-term conceptual recall and application.

At the start of each session, the local member(s) of the research team delivered a briefing outlining the study's purpose and the test-completion procedure, after which students were invited to participate. The Qualtrics link and/or the survey quick response (QR) code were shared in several ways, including being displayed on the projected screen, sent via email, included in course announcements and the letter of information, and posted on online noticeboards, along with a link to the information sheet. Students received an explanatory statement and consent form detailing the project's objectives, eligibility criteria, potential risks, and confidentiality provisions. Multiple cohorts from different countries and institutions, at varying stages of training, were included to evaluate how the PCI items functioned across a range of learner groups. Participants were informed that participation was voluntary, had no academic consequences, and was not linked to a required assessment. No incentives were offered, and students could withdraw at any time without penalty.

PCI administration and guidance

The PCI is a 24-item multiple-choice (MCQ) assessment designed to evaluate seven core pharmacological concepts [4, 5]. It includes four items on drug clearance (DCI1 to DCI4), three on drug bioavailability (DBio2, DBio3, and DBio5), four on steady-state concentration (Css1 to Css4), three on volume of distribution (Vd2, Vd3, and Vd6), four on drug efficacy (DE4 to DE7), three on drug tolerance (DT1, DT2, and DT4), and three on structure–activity relationships (SAR1 to SAR3). Each item required participants to select a single best answer from multiple options, with distractors representing common misconceptions. An "unsure" option was included to reduce random guessing (see Appendix S1). Responses were classified as correct if the participant chose the scientifically accurate answer, incorrect if they selected a distractor reflecting misconceptions, or uncertain if they chose the "unsure" option, intended for those unsure of their understanding. The PCI was administered online via Qualtrics from November 2025 to April 2026. Participants accessed the assessment via a secure link, received standardized instructions for independent completion, and were guided to select the most appropriate response.

Response quality control

Response data were screened to ensure analytical validity before processing. Multiple indicators were used to identify low-quality or disingenuous submissions. Specifically, six responses were excluded because of Qualtrics' automated bot-detection flags, implausibly short completion times relative to the sample distribution, and patterned or invariant responding. Responses were also checked for completeness of all PCI items, and predefined criteria were applied to exclude incomplete or low-engagement submissions. For this analysis, only responses with complete data on all 24 PCI items were included to ensure consistency across all psychometric and modelling analyses, avoid assumptions about missing data, and obtain stable estimates of item, correlation, factor, and attribute parameters. This applied to full response patterns. This was particularly important for the CDA analysis, which entirely relies on complete item responses. As a result, 534 incomplete responses were excluded. Only cleaned and validated datasets were retained for subsequent analysis.

Analytical framework

Analyses followed a three-stage, model-driven process (see Fig. 1): (i) assessing the instrument's latent structure; (ii) examining psychometric properties at both item and test levels; and (iii) performing structural modeling and invariance analysis. The final step aimed to explore how participant traits influence results and whether measurements remain consistent across different subgroups.

Dimensionality assessment

To evaluate the internal structure of the PCI, the total sample was randomly divided into two equal, independent subsamples. This split-sample method ensures the replicability of the factor structure and prevents the final model from overfitting to sample-specific characteristics. The first subsample was used to explore the dimensionality of the data, and the second subsample served to independently verify structural validity [20–23]. In stage one, a subsample ($n = 607$) of cases was used for an exploratory factor analysis (EFA) to identify the underlying factor. The dimensionality of the PCI was assessed using a multi-step approach suitable for dichotomous responses. Inter-item associations were estimated using a tetrachoric correlation matrix [24], which assumes that binary responses reflect an underlying continuous latent variable related to pharmacological understanding. Data suitability for factor analysis was assessed using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity [25, 26]. Parallel analysis (PA) was employed to identify the optimal number of latent factors by comparing observed eigenvalues to those from 1000 randomly generated datasets. Factors were kept only if the observed eigenvalues surpassed the 95th percentile of the simulated thresholds [27, 28]. By using the 95th percentile, this method controls for random variation specific to this exploratory sample, ensuring that only factors that explain more variance than random patterns are retained. This approach helps reduce the risk of over-extracting factors. The decision to retain factors was based on the criterion that the observed eigenvalues must exceed the simulated values.

Following the PA, EFA was conducted using the minimum residual (minres) extraction method and the direct oblimin rotation; this rotation was chosen to account for the expected overlap among pharmacological domains. Item-factor relationships were identified using a primary loading threshold of 0.30–0.40, with a threshold > 0.30 used to detect potential cross-loadings [25, 26]. This tiered approach served to distinguish items that aligned with specific conceptual clusters from those showing “concept-drift” toward a more unidimensional structure. To move from discovery to structural validation, exploratory structural equation modelling (ESEM) was subsequently applied to this exploratory subsample. ESEM integrates the descriptive flexibility of EFA with the rigorous statistical goodness-of-fit assessments of traditional structural equation modelling (SEM) [29, 30]. Standard confirmatory approaches are often overly restrictive for educational tests, as forcing cross-loadings to zero can artificially inflate factor correlations and distort the true structural validity [31]. For the PCI, ESEM accommodates the reality that the core pharmacological concepts are inherently interconnected, allowing items to cross-load across multiple domains while simultaneously evaluating overall model fit.

The dimensionality identified in the EFA was cross-validated in stage two using a second independent subsample ($n = 604$). Building on the exploratory structure, PCI performance was assessed against the pharmacology reasoning or proficiency

framework. In this study, pharmacology proficiency refers to the capacity to understand, reason about, and apply core pharmacology concepts and PK-PD principles. Confirmatory factor analysis (CFA) was conducted to assess model fit, and a more stringent threshold was applied, prioritizing loadings of 0.70 or higher to ensure high factor determinacy and stable parameter estimates [32]. This sequential transition from EFA to CFA provided a rigorous test of the instrument's dimensionality. The model fit, indicating the average difference between the model's predicted correlation and the actual correlation matrix, was evaluated using various fit indices. Specifically, model fit was assessed through the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMSR).

Measurement modeling

To evaluate measurement properties, we used a multifaceted approach that combined CTT, IRT, CFA, and CDA. Specifically, CTT and IRT were applied to assess item-level parametric properties and tool reliability [33, 34], while CFA verified structural characteristics, and CDA investigated diagnostic capabilities, construct alignment, and item characteristics. Under CTT, item difficulty was defined as the proportion of correct responses (p -value) and classified as easy (> 0.70) or difficult (< 0.30). Discrimination indices (e.g., point-biserial (r_{pb})) assessed items' ability to differentiate high- and low-performing students and were interpreted using established thresholds, such as excellent (≥ 0.40), good (0.30 – 0.39), fair (0.20 – 0.29), or poor (< 0.20) [35, 36]. Scale internal consistency reliability was assessed using Cronbach's alpha, with values interpreted according to established thresholds. IRT models were fitted to estimate item-level parameters. Depending on the model selected, item difficulty and discrimination parameters were estimated using the Rasch (1PL) or two-parameter logistic (2PL) models, depending on the optimal statistical fit. Item difficulty and discrimination ranges were also classified with the recommended ranges [37, 38].

Selection of an appropriate CDA model

CDA are generally categorized as compensatory or non-compensatory based on attributes related to successfully answering an item [39, 40]. In a compensatory model, strong mastery of one attribute can compensate for weaker mastery of another, increasing the likelihood of a correct response through skill trade-offs. This allows an examinee to still answer correctly even if they lack one attribute, with success influenced by the combination of mastered attributes in models such as deterministic inputs, the noisy-or-gate (DINO), and the additive cognitive diagnostic model (A-CDM) [9, 39]. Conversely, a non-compensatory model assumes that all attributes must be mastered for success and that the absence of any essential attribute cannot be offset by others, leading to a lower probability of a correct answer. This approach underpins models such as the deterministic inputs, noisy and gate (DINA), and the reduced reparametrized unified model (R-RUM) [40, 41]. The decision to use a compensatory or non-compensatory model depends on the task's theoretical framework and the interaction of skills. The generalized DINA (G-DINA) model is a comprehensive, saturated model that can handle both relationships and is often used to empirically determine which assumption best aligns with the data [40].

CDA was implemented as an exploratory validation of the PCI to examine alignment between items and underlying constructs, assess item behavior, and evaluate the tool's potential to generate a diagnostic mastery profile. The CDA framework, with an expert-defined Q-matrix linking each item to a single attribute, was used to assess item-attribute (core pharmacology concept) alignment, item behavior, and the PCI's diagnostic capability. A theory- and expert-based Q-matrix (Table 1) was developed to specify the relationships between items and core pharmacology attributes [4, 5]. The matrix was finalized through expert evaluation, with 11 content experts mapping the 24 PCI items across seven pharmacological domains. Inter-rater reliability, assessed via Fleiss' kappa, ranged from 0.7 to 1.0, indicating substantial to almost perfect agreement [42]. As a result, a simple-structure Q-matrix was produced, in which each item was assigned to a single attribute (core concept), and each attribute was represented by three to four items.

Table 1

The table for the expert refined Q-matrix for each pharmacology concept inventory (PCI) item to concept map

Items	DCI	DBio	Css	Vd	DE	DT	SAR
DCI1	1	0	0	0	0	0	0
DCI2	1	0	0	0	0	0	0
DCI3	1	0	0	0	0	0	0
DCI4	1	0	0	0	0	0	0
DBio2	0	1	0	0	0	0	0
DBio3	0	1	0	0	0	0	0
DBio5	0	1	0	0	0	0	0
Css1	0	0	1	0	0	0	0
Css2	0	0	1	0	0	0	0
Css3	0	0	1	0	0	0	0
Css4	0	0	1	0	0	0	0
Vd2	0	0	0	1	0	0	0
Vd3	0	0	0	1	0	0	0
Vd6	0	0	0	1	0	0	0
DE4	0	0	0	0	1	0	0
DE5	0	0	0	0	1	0	0
DE6	0	0	0	0	1	0	0
DE7	0	0	0	0	1	0	0
DT1	0	0	0	0	0	1	0
DT2	0	0	0	0	0	1	0
DT4	0	0	0	0	0	1	0
SAR1	0	0	0	0	0	0	1
SAR2	0	0	0	0	0	0	1
SAR3	0	0	0	0	0	0	1
NB: DCI-drug clearance, DBio-drug bioavailability, Css-steady-state concentration, Vd-volume of distribution, DE- drug efficacy, DT- drug tolerance, SAR-structure–activity relationships.							

Model comparison and item-attribute structure alignment

To determine the mathematical relationships between core concepts and students' responses, multiple CDA models (GDINA, DINO, A-CDM, DINA, and R-RUM) were estimated and compared. The comparison used absolute fit indices (SRMSR, RMSEA) and relative fit measures (Akaike information criterion (AIC) and Bayesian information criterion (BIC)) to evaluate exploratory CDA models of the PCI. Item discrimination was operationalized as the difference in success probabilities between the non-mastery profile (P (0)) and the mastery profile (P (1)). These separation values were used to characterize the extent to which items differentiated across levels of attribute mastery. An integrated evaluation framework was specified to jointly consider

item fit and discrimination, enabling interpretation of whether items functioned consistently with their intended constructs while retaining diagnostic sensitivity.

Structural modelling and invariance testing

Structural equation modelling (SEM) was employed to examine relationships between latent constructs and observed participant characteristics [43]. The model integrated measurement and structural components to evaluate the influence of external variables on estimates of overall pharmacologic understanding. Multiple indicators and multiple causes (MIMIC) models were used to assess the effects of participant characteristics (e.g., demographic or educational variables) on latent constructs and individual item responses [43]. Direct effects from covariates to items were examined to identify potential sources of measurement bias. Extensions of the MIMIC model were applied to formally test for differential item functioning (DIF) [44, 45]. Items exhibiting significant direct effects from covariates after controlling for latent traits were identified as potentially exhibiting DIF. These analyses were used to evaluate measurement invariance across subgroups. All statistical analyses were performed using the CDM package in R (R Foundation, Vienna, Austria) [46].

Results

Participant characteristics

A total of 1751 started the Qualtrics survey, of whom 1211 completed the PCI. Participants' demographic information was collected via a structured questionnaire administered alongside the study instrument. The PCI was conducted in 12 countries, namely Hungary, India, the USA, Australia, Romania, Malta, Canada, the UK, Sweden, Finland, Germany, and the Netherlands, across 23 institutions, reflecting substantial geographic and institutional diversity (Fig. 2). Most participants were enrolled in general or introductory pharmacology courses (68.4%), with others in pharmacokinetics (PK) and pharmacodynamics (PD) (10.3%). Individuals who identify as female made up 57.8% of the sample. Participants self-identified their backgrounds, with medical students comprising the largest group (62.7%). Most participants self-identified as undergraduate students (78.8%), with third-year students the largest year-of-study category (61.6%), reflecting differences in education systems between countries. Most respondents were recruited from undergraduate-entry health professional programs, including medicine, pharmacy, and related disciplines. Education levels were categorized to reflect the conventions of each participant's institution and country. Language backgrounds varied, with most reporting a primary language other than English (61%) or being multilingual (11.5%) (Table S1).

Dimensionality and Factorability

For latent structure modeling evaluation, the tetrachoric correlation matrix was estimated using the exploratory subsample ($n = 607$). The matrix showed moderate inter-item associations ($r = 0.30-0.49$), with several strong relationships ($r \geq 0.5$) between conceptually related items, such as DBio3 and DBio5 ($r = 0.64$) and DT2 and DT4 ($r = 0.63$). Notably, significant correlations across theoretically discrete domains, such as the association between Vd3 and DT2 ($r = 0.61$), provided clear evidence of a unified latent construct rather than isolated conceptual compartments (Table S2). Data factorability was confirmed by excellent overall sampling adequacy ($KMO = 0.84$) and a statistically significant Bartlett's test of sphericity ($\chi^2_{276} = 11588.04$, $p < 0.001$), indicating that the inter-item correlation matrix was not an identity matrix and that items shared substantial common variance, which was appropriate for latent structure modelling.

Exploratory factor analysis (EFA)

The PCI demonstrated that students' responses reflected robust overall pharmacological reasoning and grouped into distinct clusters within a multidimensional framework. Parallel analysis (PA) using tetrachoric correlations identified a 7-factor structure. The 8th factor was too small to be counted, and its eigenvalue differed from the random eigenvalue by less than 0.02. A very strong primary factor (eigenvalue = 8.47) exceeded the 95th percentile of simulated values (0.44), followed by a steep decline to the second factor (eigenvalue = 1.06) and smaller residual dimensions. PA indicated that one dominant factor and six smaller multidimensional clusters were appropriate (Table S3). The scree plot also showed a sharp drop after

the first factor, with smaller declines among the weaker factors, supporting the presence of additional meaningful factors. This suggests that the PCI includes both a strong general factor and a multidimensional structure aligned with pharmacological domains, rather than being purely unidimensional. Therefore, subsequent analyses focused on meaningful models rather than testing all configurations. A 1-factor model represented the main reasoning component; a 2-factor model examined whether residual multidimensionality reflected a broader PK-PD distinction; and a 7-factor model matched the PCI core concepts. These models aimed to determine whether the PCI functions primarily as a unidimensional measure or as a conceptually distinct, multidimensional construct.

Further analyses using EFA and ESEM showed gradual improvements in model fit as the number of factors increased. ESEM combines EFA flexibility with SEM's confirmatory approach, allowing items to load on multiple factors and enabling assessment of overall fit. Consequently, ESEM was used to explore whether PCI items formed multiple pharmacological domains with cross-loadings, thereby indicating interconnected core concepts that require integrated reasoning. The EFA/ESEM results suggested that a two-factor model fit best, implying possible multidimensionality. However, the two factors were moderately correlated ($r = 0.55$), and some items cross-loaded or split across factors, indicating that the structure did not clearly separate distinct latent domains. A 7-factor model had near-perfect fit indices (CFI = 1.00, TLI = 1.00, RMSEA = 0.00, SRMR = 0.030) but was unstable and over-factored. Many items cross-loaded on multiple factors; some factors were weakly defined, with few key items, and the negative residual variance for item SAR1 (-1.75) indicated estimation issues and limited interpretability. Additionally, expert-designed items for the item-concept matrix did not cluster neatly into the proposed factor groups; several items from C_{ss}, D_{Bio}, D_{CI}, and D_E appeared across multiple dimensions (Table S4). These findings indicate that, while the PCI primarily presented a single overarching proficiency measure, residual multidimensionality revealed that its structure also encompassed overlapping and interconnected reasoning processes. However, this multidimensionality did not clearly present as seven separate, stable latent domains

PCI structure with confirmatory factor analysis

The CFA showed that PCI performance is primarily driven by a dominant overarching factor related to pharmacology reasoning or proficiency. The one-factor CFA showed acceptable fit indices (CFI = 0.936, TLI = 0.930, RMSEA = 0.045, SRMR = 0.079) (Table 2). Consistent with the one-factor CFA model, all 24 items had significant standardized loadings (ranging from 0.45 to 0.73), supporting the measure's unidimensional structure (see Tables S5A and S5B). The two-factor CFA provided a slightly improved fit, suggesting residual multidimensionality beyond the main factor. However, the two-factor were highly correlated ($r = 0.85$), indicating they are not distinct constructs. The seven-factor CFA achieved the best fit among the models tested, with all items loading well on their respective domains. Yet the factor correlation matrix was not positive definite, and some inter-factor correlations exceeded 1 (e.g., C_{ss}-V_d = 1.10; D_{CI}-V_d = 1.1), indicating serious problems with discriminant validity and that the pharmacology core concepts were not empirically separable. The higher-order CFA highlighted the dominance of a general factor, with all seven domain factors loading very strongly (0.82–1.00), suggesting the domains are largely explained by a single overarching pharmacology-reasoning construct. However, estimation warnings and near-perfect correlations suggested redundancy among several core concepts. This confirmatory analysis indicates that PCI functions best as a mostly unidimensional assessment, with some secondary structures overlapping, rather than as seven entirely separate domains. The strongest evidence favours interpreting PCI scores at the general pharmacology reasoning level (see Table S5).

Table 2

Comparative confirmatory factor analysis (CFA) fit indices for the 1-factor, 2-factor, and 7-factor models of the PCI

Model	Factor correlation	Chi-square (χ^2) (df)	CFI	TLI	RMSEA	RMSEA Lower	RMSEA Upper	SRMR	P-value
1_Factor_CFA	*	564.27 (252)	0.936	0.930	0.045	0.040	0.050	0.079	0
2_Factor_CFA	0.85	505.46 (251)	0.948	0.942	0.041	0.036	0.046	0.074	0
7_Factor_CFA	1.0 (several)	465.97 (231)	0.952	0.942	0.041	0.036	0.046	0.071	0
NB: * Correlation not applicable									

Measurement modeling

Exploratory cognitive diagnostic assessment modeling of the PCI

The CDA framework assessment yielded very similar relative fit indices across all models (GDINA, DINO, ACDM, DINA, and RRUM), with AIC and BIC differing by up to 0.07 points (Table 3). Similarly, the log-likelihoods showed minor differences. Because these differences were below the standard threshold of 2.0, no model was statistically superior to the generalized framework [47, 48]. All models had good and consistent absolute fit indices (RMSEA = 0.044, 90% CI [0.039, 0.049]; SRMSR = 0.074) and met accepted standards of fit [49]. Given this empirical equivalence, the generalized GDINA model was selected for the final parameter estimation [50]. The GDINA model avoids restrictive assumptions about cognitive diagnosis rules, allowing for all possible item interactions. It also accommodates additive cognitive processes and compensatory reasoning rather than strict all-or-nothing mastery of attributes. This aligns with prior findings, which indicate that pharmacology concepts are intercorrelated and that the cognitive exploratory construct-diagnosis solution showed considerable overlap among concepts.

Table 3

Absolute and relative fit statistics for exploratory Cognitive Diagnostic Models (CDM) of the PCI

Model	Absolute fit statistics		Relative fit statistics		
	SRMSR	RMSEA	AIC	BIC	LL
ACDM	0.0742	0.0438	33525.08	34417.44	-16587.54
RRUM	0.0742	0.0422	33525.09	34417.45	-16587.55
GDINA	0.0742	0.0438	33525.15	34417.51	-16587.58
DINO	0.0742	0.0438	33525.15	34417.51	-16587.58
DINA	0.0742	0.0438	33525.15	34417.51	-16587.58

Test-level accuracy was moderate at 0.69, whereas attribute-level accuracy was high, ranging from 0.91 to 0.92. Item fit analysis revealed significant diagnostic misfit across all items, consistent with earlier PA, EFA/ESEM, and CFA results that identified a dominant general pharmacology reasoning factor alongside a broader multidimensional structure. The mastery classification analysis showed differences among PCI domains: DT had the highest endorsement probability, with DCI and SAR the lowest (Fig. 3). Mastery profiles tended to cluster at either extreme, with slightly less than one-third (30%) of students demonstrating no mastery and approximately one-quarter (26%) demonstrating mastery of all seven attributes. Students with mastery across all PCI items generally had higher success probabilities, though the magnitude of the differences varied.

Overall test and individual item properties

Overall test performance

The average score on the tool was 12.4 out of 24 items (SD = 5.6), with a 51.7% correct rate, reflecting good variability and discrimination across levels of conceptual understanding. Most items showed moderate to high variance (0.20–0.25), indicating notable differences in responses and effective separation of students' knowledge of pharmacology. Higher variance suggests better discrimination, while lower variance indicates more uniform responses, which could reduce the tool's ability to differentiate between performance levels. This resulted in a diverse distribution of responses across various pharmacology concept areas. Participants performed best on DT, with 75% correct answers, followed by DBio (57%) and DE (54%). Meanwhile, distractor choices were most frequent for Css (40%), followed by DBio (36%) and Vd (35%), reflecting misconceptions. The SAR (29%) and DCI (20%) domains exhibited higher response uncertainty, possibly indicating lower confidence or greater variation in understanding and/or concept coverage (see Fig. 4). In most domains, misconception-driven responses were more common than uncertain ones, meaning incorrect answers often stemmed from misunderstandings rather than doubt. The tool also demonstrated strong internal consistency, with Cronbach's alpha of 0.86 and Omega total of 0.93, confirming the PCI's reliable internal structure.

Individual item quality

Individual items were assessed for discrimination and difficulty using CDA, CTT, and IRT models. As illustrated in Table 4, the PCI consistently showed strong item discrimination and moderate difficulty. Exploratory CDA analysis indicated moderate to very good diagnostic separation, with item discrimination indices (IDI) ranging from 0.31 to 0.65 (mean = 0.48). Of the 24 items, 19 (79%) had very good discrimination (≥ 0.40), while the remaining items were reasonably good, except for one item, Css4, which showed the lowest IDI (0.31) (Table 4). Similarly, within the CTT framework, item difficulty values ranged from 0.20 to 0.86, with most items ($n = 19$; 79%) demonstrating difficulty values ≥ 0.40 . A lower difficulty value indicates items are more challenging; except for one item, Css4, with a marginal item difficulty value of 0.20, the remaining items fell within acceptable difficulty ranges. Likewise, the CTT-based item discrimination indices ranged from 0.32 to 0.51; most items ($n = 15$, 63%) had indices between 0.41 and 0.51, and the remaining items showed good discrimination. These discrimination values suggest that each item's score was positively correlated with overall test performance, supporting the items' ability to differentiate between lower- and higher-performing respondents.

Table 4
Item quality indices from CDM model, CTT and IRT

Item	Item Discrimination CDA	Item Difficulty CTT	Item Discrimination CTT	Item Difficulty 2PL	Item Discrimination 2PL
DCI1	0.33	0.74	0.32	-1.30	0.97
DCI2	0.54	0.31	0.46	0.74	1.45
DCI3	0.48	0.52	0.44	-0.10	1.33
DCI4	0.41	0.40	0.37	0.47	1.04
DBio2	0.34	0.68	0.35	-0.91	0.99
DBio3	0.59	0.41	0.47	0.31	1.47
DBio5	0.65	0.61	0.50	-0.38	1.81
Css1	0.41	0.36	0.39	0.65	1.07
Css2	0.54	0.55	0.51	-0.18	1.78
Css3	0.47	0.63	0.49	-0.45	1.74
Css4	0.31	0.20	0.39	1.37	1.30
Vd2	0.50	0.43	0.45	0.27	1.29
Vd3	0.51	0.65	0.47	-0.55	1.65
Vd6	0.52	0.34	0.48	0.56	1.58
DE4	0.36	0.59	0.34	-0.49	0.89
DE5	0.49	0.49	0.41	0.01	1.18
DE6	0.49	0.56	0.43	-0.24	1.28
DE7	0.53	0.50	0.45	-0.01	1.31
DT1	0.37	0.86	0.42	-1.39	1.94
DT2	0.60	0.76	0.47	-0.91	2.02
DT4	0.63	0.63	0.45	-0.51	1.50
SAR1	0.50	0.41	0.40	0.39	1.11
SAR2	0.48	0.33	0.34	0.88	0.95
SAR3	0.52	0.46	0.40	0.17	1.11

Using the BIC, a comparison of 1PL, 2PL, and 3PL models was conducted, and the 2PL model showed the best fit (1PL, BIC = 32897.0; 2PL, BIC = 32880.3; 3PL, BIC = 32914.5). Therefore, the 2PL model was deemed the most reliable in this analysis, aligning with sample-size planning and parameter-estimation guidelines for a robust item-response framework [37]. In IRT, item difficulty indices range from - 2.00 to 2.00 (Fig. 5). As shown in Table 4, item difficulty indices ranged from - 1.39 to 1.37, indicating that almost all items exhibited acceptable difficulty levels and were classified as moderate. Regarding item discrimination, all indices fell within the recommended range, except for item DT2, which showed very high discrimination (2.02). Overall, the item discrimination parameter fell well within acceptable psychometric boundaries ($\alpha \geq 0.50$), specifically matching the recommended operational range of 0.50 to 2.0 [37], with the majority indicating high or very high discrimination power according to standard thresholds. Most items ($n = 20$, 83%) had discrimination values above 1.00, while the remaining ($n = 4$, 17%) had values ≥ 0.89 (Baker et al., 2017; Mi et al., 2023). The IRT findings suggest that the items could reasonably distinguish between students with abilities below and above the item's location. The quality parameters demonstrated that

the PCI item successfully discriminated across classification, latent trait, and exploratory diagnostic frameworks, even though the concepts were integrated [13].

Structural validity and demographic effects on pharmacology proficiency

The baseline MIMIC model, used to assess whether demographic variables significantly predicted the main latent factors in pharmacology reasoning and performance, showed acceptable overall fit (CFI = 0.912, TLI = 0.934, RMSAE = 0.042, SRMR = 0.070). This supports the interpretation that the PCI primarily assesses a single underlying dimension of pharmacological proficiency. The MIMIC regression analysis indicated that program type and year of study were significant predictors of latent PCI performance, whereas gender, language background, and academic level were not. Specifically, nonmedical students scored significantly lower on latent pharmacology reasoning than medical students ($\beta = -0.084$, $p = 0.016$). Additionally, students in later years had significantly lower scores than first- and second-year students ($\beta = -0.182$, $p < 0.001$) (see Table 5).

Table 5
Multiple indicators and multiple causes (MIMIC) regression analysis of demographic predictors of the latent PCI performance

Predictor	Reference group	Comparison group	β (Estimate)	SE	Z	Standardize β	P-value
Gender	Male	Female	-0.064	0.071	-0.904	-0.031	
Language	English	Non-English	-0.079	0.092	-0.857	-0.035	
Program type	Medical students	Nonmedical students	-0.178	0.074	-2.411	-0.084	0.016
Academic level	Undergraduate	Postgraduate	-0.030	0.086	-0.351	-0.012	
Year of study	First/second year	Third and above	-0.434	0.099	-4.396	-0.182	0.000
NB: (-Ve) β coefficient indicates lower latent PCI performance for the comparison group relative to the reference group							
NB: Interpreting the effects of the year of study demands caution, as educational pathways differ significantly between countries and programs. For instance, graduate-entry medical programs might categorize students as first-year students even if they have considerable prior university experience. Therefore, the year-of-study variable should be seen as a general indicator of training stage rather than a precise measure of progression along a standardized preclinical-to-clinical curriculum.							

Differential item functioning (DIF) across demographic groups

Combined MIMIC-DIF and IRT-DIF analyses identified DIF for item DCI3 across program types and for items Css1, DE5, and Vd2 across participants' reported year of study within their current program. All four items showed significant DIF after adjustment in both analytical frameworks, highlighting consistent subgroup-specific differences in item responses when latent PCI ability was held constant. The magnitude of the item response differences was small to moderate, with the highest unstandardized effect size for item Vd2 across the year of study groups ($\beta = -0.564$, IRT, $\chi^2 = 50.99$, adjusted $p < 0.001$) (Table 6). Item characteristic curve (ICC) plots in Fig. 6 showed a consistent rightward shift for non-medical students on item DCI3. Likewise, as illustrated in Fig. 7, rightward shifts for third- and fourth-year students relative to first- and second-year students were observed in items Css1, DE5, and Vd2. These shift patterns indicate that, at equivalent levels of latent paragraph ability, students in the comparison groups were less likely to answer these items correctly than those in the corresponding reference groups. Across all confirmed DIF items, slope patterns were similar across groups, indicating comparable item discrimination and predominantly uniform DIF (i.e., systematic differences in item difficulty, whereby one group consistently found the item harder across the entire ability continuum). This pattern suggested that the observed DIF was primarily attributable to differences in item difficulty rather than to differences in discrimination. The most notable subgroup separation was observed in Vd2, consistent with its stronger MIMIC and IRT DIF statistics.

Table 6

Confirmed differential item functioning identified by combined multiple indicators and multiple causes with differential item functioning (MIMIC-DIF and item response theory (IRT-DIF) analyses

Item	Predictor	Reference group	Comparison group	MIMIC β (Unstandardized)	SE	Z	IRT χ^2 (df)	Adjusted p (MIMIC)	Adjusted p (IRT)
DCI3	Program type	Medical students	Nonmedical students	-0.282	0.078	-3.61	28.31 (2)	0.007	< 0.001
Css1	Year of study	First/second year	Third and above	-0.344	0.099	-3.47	24.79 (2)	0.006	< 0.001
DE5	Year of study	First/second year	Third and above	-0.307	0.1	-3.08	12.47 (2)	0.017	0.009
Vd2	Year of study	First/second year	Third and above	-0.564	0.099	-5.7	50.99 (2)	< 0.001	< 0.001
NB: (-Ve) β coefficient indicates lower item performance for the comparison group relative to the reference group									

Discussion

The goal of this study was to evaluate the validity and reliability, structural integrity, item-concept alignment, and predictive capability of the 24-item PCI, which covers seven fundamental pharmacology concepts, using an integrated, model-driven framework. As this tool aims to assess learners' conceptual understanding, identify misconceptions, measure learning outcomes, and evaluate the effectiveness of teaching, interventions, and curriculum design, a detailed assessment across multiple cohorts was warranted. The results provided comprehensive evidence that the PCI is a robust, validated instrument for diagnosing concept-based pharmacology learning across diverse educational contexts and multiple student cohorts. Our findings showed that the PCI reliably diagnoses students' grasp of key pharmacology concepts across various learning environments and learner groups. Moreover, these results clearly demonstrate the instrument's usefulness in a wide range of educational settings.

The PCI tool's factor structure offers insight into how students might organize and apply their reasoning when solving pharmacological problems. Rather than viewing core concepts as discrete, independent domains, the data revealed significant cross-domain associations, indicating an interconnected approach to reasoning. For example, the strong correlation between Vd3 and DT2 ($r = 0.61$) suggests that students draw on an integrated knowledge structure rather than relying solely on isolated conceptual elements. This interpretation is further supported by a dominant primary factor representing broader reasoning ability. Although a seven-factor model emerged in PA, the unidimensional model was inadequate, and the two-factor model showed only moderate correlations, with several items cross-loading or splitting across factors (concept domains). This lack of clear separation suggests that the knowledge structure does not delineate distinct, independent domains. Instead, proficiency in pharmacology appears hierarchical, characterized by strong general reasoning ability alongside specific core structures of concepts related to pharmacology and PK-PD principles [26].

Concept-based pharmacology studies have demonstrated relationships among main concepts and sub-concepts, core ideas, and key PK-PD principles, highlighting the interconnectedness of pharmacology concepts [2, 4, 5]. For example, pharmacokinetic models show that steady-state concentration is directly dependent on clearance through an inverse relationship, rather than being an isolated phenomenon. The findings suggest that students organize pharmacological knowledge as an interconnected conceptual network rather than as a collection of isolated mechanisms. Consistent with this interpretation, analyses of dimensionality and latent structure support the proposal that the overall PCI score summarizes general conceptual understanding in pharmacology, with some interconnected multidimensional concepts. This tool was designed to assess understanding at the concept level, aiming to identify areas of strong understanding, ongoing misconceptions, and topics that need focused instructional support. The students' response pattern showed a strong overall grasp of pharmacology proficiency and reflected a locally interconnected network of related concepts. Since reasoning

involves multiple dimensions, this should be considered part of a wider educational assessment, not solely a measure of unidimensional pharmacological knowledge. The interconnected nature of pharmacology also creates opportunities to develop future test items that more effectively address multiple related concepts to assess mastery, moving beyond simple pharmacological proficiency. This underscores the potential to further refine alignment and to use concept-specific measures to more accurately identify and evaluate mastery of individual concepts. Overall, the structure reflects pharmacology's practical, integrative nature, in which effective reasoning involves cross-domain connections, thereby affirming the PCI as a reliable measure of conceptual knowledge. This multidimensional framework, though less comprehensive than the test's primary factor structure, supports a concept-based approach to teaching and assessment, offering a perspective that may help identify patterns of strength and weakness in students' understanding, albeit within an interconnected structure.

The exploratory CDM analyses provided further empirical evidence that students were combining multiple pharmacological concepts, as measured by the PCI. The GDINA framework is particularly effective at capturing the full complexity of pharmacological reasoning without requiring a single processing rule across items [50]. The generalized model also confirmed the presence of a dominant, highly integrated latent trait, suggesting that PCI performance is better explained by additive and compensatory reasoning processes rather than by strict mastery of individual concepts. This aligns with findings from both exploratory and confirmatory structural models, which consistently identify a primary pharmacological reasoning factor alongside a multidimensional structure. Although overall CDA fit indices were acceptable, item-level diagnostic misfit indicates that students' response patterns do not fully conform to a simple one-attribute-per-item Q-matrix. Instead, these patterns likely reflect the interconnected nature of PK–PD pharmacological concepts and reasoning. For example, C_{ss} depends on factors such as drug dose, bioavailability, dosing frequency, and clearance, arguing that pharmacological concepts are cognitively integrated rather than applied independently [5, 51, 52]. Most PCI items effectively distinguish between mastery and non-mastery, with mastery levels clustering at low or high proficiency. These results suggest that the PCI is best used as an integrated measure of pharmacological reasoning and conceptual understanding, while also providing opportunities to explore broader patterns of mastery rather than strictly attribute-specific diagnoses. A consistent pattern was observed across comparative evaluations of the concepts' response distributions and CDA probabilistic mastery classifications. Domains with higher correct-response proportions and lower distractor or uncertainty responses, such as DT, also showed the highest probabilities of mastery classification. Domains with elevated distractor categories and uncertainty responses, such as C_{ss}, DBio, SAR, and DCI scores, were associated with lower mastery probabilities and higher non-mastery classifications. As illustrated in Fig. 4, the descriptive response distributions revealed that distractor responses were more frequent than uncertainty responses across all concept domains, indicating that non-mastery was often associated with misconception-driven reasoning rather than uncertainty alone. Together, observable response patterns and the CDM-based latent mastery structure across core concept domains provided complementary evidence for students' understanding and misconceptions. Overall, the PCI successfully assesses general pharmacological proficiency, but its simple matrix structure may limit detailed, attribute-specific diagnoses by failing to account for the interconnectedness of these concepts.

After evaluating the item-attribute structure, the item-quality analysis across the CDA, CTT, and ITR models indicates that the items consistently exhibit good discrimination, a strong internal structure, and the ability to distinguish among different pharmacological skill levels. This consistency across the models reinforces construct validity and reliability [41]. The difficulty distribution of the items is generally balanced, with three simpler items, such as DCI1, DT1, and DT2, that measure basic knowledge. Most items fall within the optimal, moderate-difficulty range. The IRT results showed that many items had high discrimination power, effectively distinguishing between different ability levels. This indicates that these items can accurately identify students with abilities either below or above the item's specific point (Baker et al., 2017; Mi et al., 2023). Additionally, the quality metrics demonstrated that the PCI item performed well in differentiating across various classification, latent trait, and exploratory diagnostic frameworks [13].

However, within the CDA framework, C_{ss}4 showed marginal discriminative ability (0.31). This is likely because it requires an integrated understanding of concepts such as steady-state concentrations, half-life, and dosing relationships, which a single-attribute model matrix may not fully capture. Similarly, in the CTT framework, C_{ss}4 was marginally difficult ($p = 0.20$), with only a small proportion of students answering correctly, and showed good discrimination (0.39). Importantly, it demonstrated

high IRT-based discrimination ($a = 1.39$), indicating its ability to distinguish students across different proficiency levels. This pattern suggests that C_{ss}4 could be a good discriminating item for higher-ability students under a latent trait model, likely because it involves complex reasoning that links multiple concepts rather than relying on single-concept reasoning [5, 51, 52]. More importantly, the current one-item-to-one-attribute Q-matrix may not fully capture this integrative process, leading to reduced discrimination and potential misalignment at the attribute level. This does not necessarily mean the item is of poor quality; instead, it indicates a need for a more comprehensive understanding of PK to ensure proper item performance. The IRT results support this, showing that test-takers with higher overall proficiency and better ability to integrate multiple concepts are more likely to answer correctly. As demonstrated by C_{ss}4's observed CDA performance, there is a mismatch between its integrated conceptual demands and the discrete attribute structure of the CDM framework, underscoring that PK reasoning involves interconnected processes rather than isolated concepts. Furthermore, a slightly more challenging test structure could more effectively map students' learning gains onto the post-test, provided it remains valid and reliable [41, 53]. Therefore, despite its difficulty, the test is a vital tool for assessing students' conceptual understanding and identifying misconceptions in key concepts.

Furthermore, students' test scores were assessed, revealing an average of 51.7%, suggesting that many students either did not fully understand these concepts, held misconceptions, or were uncertain about them. This could reflect the current educational stage of several student cohorts, in which full conceptual integration is still in progress, especially among students with different clinical backgrounds and experiences. Although the CDA framework used in this study was not used to perform deterministic classification of mastery and non-mastery profiles, the concept (attribute)-level probabilistic estimates indicate varying levels of engagement across core concept domains. The highest endorsement probabilities were for DT, even though items were found to be easier than others. The lowest probabilistic mastery levels were observed for SAR and DCI, with the remaining being moderate. This pattern and the relative concept-based difficulty levels indicate that students have difficulty with appropriate conceptual understanding, the integration of core concepts, and overall proficiency. These conceptual difficulties persisted across cohorts and diverse settings, as evidenced by a previous study in which the concepts of drug clearance and efficacy were the most difficult to comprehend [54, 55]. Similarly, a prior qualitative study on students' conceptual understanding and misconceptions revealed that students viewed items measuring the SAR concept as unclear or fragmented [6]. However, these studies should be treated as exploratory indicators of alignment between response patterns and the intended constructs, reflecting probabilistic estimation of proficiency trends rather than discrete conceptual mastery, consistent with the integrated and interdependent nature of pharmacology knowledge.

This study not only evaluated test-level performance and probabilistic mastery status but also examined group- and demographic-level invariance across contexts and cohorts. We employed a MIMIC model to investigate potential performance disparities across groups by study year, program type, and academic level, which could influence overall understanding of essential pharmacology concepts as measured by the PCI instrument. The latent factors were regressed on these variables to assess their effects on pharmacologic proficiency. The MIMIC findings suggest that the year of study and program type were the most relevant demographic predictors of PCI performance. The lowest scores among later-year students may reflect increasingly heterogeneous reasoning strategies, curricular specialization, and/or differences in retention of key pharmacology concepts over time. This might also be related to less reinforcement as they progress through the later years of their degree, which could lead to forgetting core concepts. These findings indicate that pharmacology reasoning may not develop uniformly across educational pathways or stages of progression [56]. Likewise, nonmedical students had lower latent pharmacology reasoning scores than medical students, potentially reflecting differences in curricular emphasis, the integration of pharmacology across disciplines, or exposure to clinically oriented PK reasoning. By contrast, medical students often study pharmacology in a practical, patient-centred approach [57, 58].

Additionally, item design fairness was evaluated across demographic groups using a triangulated DIF analysis comparing results from the MIMIC (SEM) and IRT (2PL) models. Four items (DCI3, C_{ss}1, Vd2, and DE5) showed confirmed DIF in both models, indicating that responses to these items varied systematically across some learner groups after accounting for overall pharmacology proficiency. Although the number of affected items was small, these findings suggest that certain item features may be interpreted differently by learners with different educational backgrounds or training experiences.

Consequently, these items have been prioritized for further review and refinement through expert evaluation and learner feedback. Overall, the limited extent of DIF supports the fairness of the PCI across diverse student populations while highlighting opportunities for ongoing improvement.

Implications for the education community

The large-scale, multinational validation and effectiveness evaluation of the PCI tool establish it as a psychometrically robust and educationally informative instrument for concept-based pharmacology education across diverse health, biomedical, and pharmaceutical science programs at both undergraduate and postgraduate levels. Rather than relying solely on a single total score, this multi-stage framework uses structural confirmatory analysis and CDA to provide educators with a highly detailed diagnostic view of students' conceptual learning and understanding. For example, this instrument could be used as a pre- and post-test to measure learning gains, inform the design of mitigation strategies for potential misconceptions, and assess teaching effectiveness. The group invariance analysis using MIMIC modelling and DIF item fairness analysis would further support measurement consistency, item fairness, and predictive power across diverse global cohorts. This empirical evidence supports the view that the PCI might serve as an objective benchmark for assessing the learning of interrelated, concept-based core pharmacology concepts within a holistic framework. This instrument facilitates a structured approach to teaching key concepts and helps identify persistent misunderstandings in basic pharmacology education. It offers prompt, practical diagnostic feedback that faculty can use to improve teaching methods, apply corrective strategies, and enhance curriculum planning. Additionally, it serves as a validated research instrument.

Design and implementation flexibility allow the PCI to be used for both formative and summative assessments (Netere et al., 2026) and provide a versatile, validated, and dependable assessment framework suitable for various groups and teaching environments. Its strong consistency across contexts and curricular alignment supports an integrated approach to concept-level and overall general pharmacology training, aiding the assessment of complex learning outcomes, informing instructional strategies, and evaluating interventions. Anchored in robust psychometric techniques and measurement approaches, the PCI provides clear insights into learners' thought processes, delivering actionable information that improves students' performance monitoring and the quality of institutional teaching. Ultimately, this comprehensively designed and validated tool provides wide-ranging evidence of reliability, encouraging educators and researchers to adopt and evaluate concept-based curricula combined with evidence-based diagnostic assessments.

Limitations and Future Direction

When interpreting the student grouping findings, certain methodological factors should be considered. First, the international, multi-cohort design naturally led to differences in program structures, institutional curricula, topic coverage, and instructional schedules. Educational pathways vary across countries and disciplines; for instance, some professional programs begin immediately after secondary education, whereas others require prior undergraduate experience. Consequently, background characteristics such as program type and year of study are better viewed as general markers of educational level rather than precise, comparable measures for all participants. Second, voluntary participation likely introduced self-selection and motivational biases not typically present in graded assessments. Third, limiting the analysis to complete responses, though safe for correlation and modeling, might introduce selection bias in descriptive results. Students who completed the PCI could differ systematically from those who did not, particularly in engagement, motivation, and academic preparedness. Thus, findings should be interpreted with caution, particularly for student groups that may have disengaged due to fatigue, test difficulty, or technical issues. Finally, because the expert-defined Q-matrix used a simple structure, mapping each item to a single concept, and the exploratory and confirmatory factor analyses revealed a dominant primary factor alongside some multidimensionality, our implementation of the CDA framework serves as an exploratory tool, particularly for concept-level mastery evaluation. It estimates the probability of concept mastery rather than providing an absolute, deterministic classification profile.

To address these limitations, future iterations may better implement a continuous, expert-led item-attribute refinement process, increase students' motivation, ensure tests are appropriately timed, and mitigate potential bias. Moreover,

subsequent validation and refinement studies may support the test's cross-national use and help distinguish differences across diverse educational contexts. In addition to improving the instrument, the PCI provides educators with a tool to identify persistent misconceptions and gaps in pharmacology knowledge. These insights could inform targeted teaching strategies, such as concept-focused learning, formative feedback, and activities to address misconceptions. Consequently, the PCI functions not only as an assessment tool but also as a framework for evaluating educational interventions and monitoring conceptual development over time. This continuous process could improve the PCI's long-term validity, educational value, and its role as a standardized, concept-based pharmacology assessment across diverse international curricula.

Conclusion

This study assessed the PCI's validity, reliability, and educational usefulness across diverse educational settings and student groups, and it can be implemented more broadly as a concept-based assessment in pharmacology. Findings showed that the PCI effectively measured pharmacological proficiency, conceptual understanding, and persistent difficulties and misconceptions. Consistent results across learner groups and settings suggest that the PCI can be widely applied across diverse educational contexts. Additionally, performance results and patterns at the item and concept levels showed that students have a consistent understanding of concepts, ongoing misunderstandings, and difficulties at the concept level. These findings help pinpoint learning needs, support educators in enhancing teaching methods, assess the success of interventions, and inform curriculum development and educational choices. Moreover, cognitive diagnostic analyses provided further insights into item–concept relationships and patterns of student understanding, highlighting areas for potential refinement of the instrument. Overall, the study affirms the PCI as a concept-based assessment tool that supports and evaluates pharmacology learning.

Abbreviations

CDM, cognitive diagnostic model; CFA, confirmatory factor analysis; CFI, Comparative Fit Index; C_{ss}, steady-state concentration; CTT, classical test theory; DBio, drug bioavailability; DCI, drug clearance; DE, drug efficacy; DIF, differential item functioning; DT, drug tolerance; EFA, exploratory factor analysis; ESEM, exploratory structural equation modelling; IRT, item response theory; IUPHAR-Ed CCP: International Society for Basic and Clinical Pharmacology Education Section Core Concepts of Pharmacology; PK: pharmacokinetic(s); MIMIC, multiple indicator multiple causes; PCI: Pharmacology concept inventory; PD: pharmacodynamic(s); RMSEA, Root Mean Square Error of Approximation; SAR, structure–activity relationships; SEM, Structural equation modelling; SRMSR, Standardized Root Mean Square Residual (SRMR); TLI, Tucker–Lewis Index; V_d, volume of distribution.

Declarations

Supporting information

The supporting information accompanying this study includes additional materials related to the PCI validations and cognitive evaluation: Table S1 (Demographic characteristics of participants in the survey), Table S2 Tetrachoric correlation matrix used for latent structure and dimensionality analyses of the PCI, Table S3 (Parallel Analysis (PA) comparison of observed and simulated eigenvalues for the PCI), Table S4 (Pattern matrix of PCI item loadings and psychometric classification), and Table S5 (Confirmatory factor unidimensional model and psychometric classification). These materials are available as supplementary files online.

Acknowledgment

The authors acknowledge the international expert panel members for their insights in the creation and evaluation of the PCI items. We are also grateful to all international students for their contributions to the survey, to all participating institutions and their staff, and to the course coordinators for their support in conducting it.

Authors' contributions

Adeladlew K. Netere: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, writing – review and editing. **Tony Hughes:** conceptualization, methodology, supervision, validation, writing – original draft, writing – review and editing. **Clare Guilding:** Conceptualization; investigation; methodology; visualization; validation; writing – original draft; writing – review and editing. **Anna-Marie Babey:** conceptualization, formal analysis; investigation; methodology; validation, visualization; writing – original draft; writing – review and editing. **Willmann Liang:** conceptualization, formal analysis, methodology, validation, visualization, writing – original draft, writing – review and editing. **Lynette B. Fernandes:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Olusola Olafuyi:** methodology, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Shahida Mallah:** methodology, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Fabiana A. Caetano Crowley:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Anita Woods:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Carolina Restini:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Farhan Ahmad Khan:** Investigation; methodology; visualisation; validation; writing – original draft; writing – review and editing. **Janet Mifsud:** Conceptualization; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Jennifer Koenig:** Conceptualization; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Lindsay Cormier:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Jennelle Durnett Richardson:** methodology, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Mark Hernandez:** Conceptualization; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Kelly Karpa:** Conceptualisation; investigation; visualisation; writing – original draft; writing–review and editing. **Anneke van Houwelingen:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Patrik Aronsson:** Conceptualisation; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **Martin Hawes:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Wael Eldahshan:** conceptualization, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Reem Atawia:** validation, visualization, writing – original draft, writing – review and editing. **Sudhir Chandra Sarangi:** validation, visualization, writing – original draft, writing – review and editing. **Tori Llewelyn:** validation, visualization, writing – original draft, writing – review and editing. **Ullamari Pesonen:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Christine Edmead:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Sarah Bailey:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Lorena Dima:** data curation; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **Anikó Görbe:** data curation; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **Zsófia Onódi:** data curation; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **Pál Riba:** data curation; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **Zoltán Varga:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Dax Vendrig:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Arno Werners:** Investigation; visualisation; validation; writing – original draft; writing – review and editing. **Marina Junqueira Santiago:** methodology, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Graeme J. Sills:** methodology, formal analysis, validation, visualization, writing – original draft, writing – review and editing. **Paul J. White:** conceptualization, methodology; software; investigation; validation; visualisation; project administration; supervision, resources; writing – original draft; writing – review and editing.

Funding

This research received no specific funding from any organization or institution

Data availability and materials

All working data for this study will be available on reasonable request to the corresponding author

Ethics approval and consent from participants

Ethics approval for the study was granted by the lead center and obtained from the Monash University Human Research Ethics Committee (MUHREC; approval protocol ID: 45836). In addition, participating institutions completed local ethical review. Participating institutions relied on the MUHREC approval and, when required, obtained additional local ethics approval or institutional permissions in accordance with their local governance requirements. These included approvals from the University of Surrey Ethics Committee (RM Ref: 3238), approvals from Universiteit Utrecht, an approval from Semmelweis University (SE-RKEB-29/2026), and a Letter of cooperation between Monash University and Southwestern Oklahoma State University (SWOSU). All study procedures were conducted in accordance with the MUHREC approval, and informed consent was obtained from all participants prior to participation.

Consent for publication

Not Applicable

Data availability

All working data for this study will be available on reasonable request to the corresponding author

Conflict of Interest

The authors declare they have no conflict of interest.

References

1. Fasinu, P. S., & Wilborn, T. W. (2024). Pharmacology education in the medical curriculum: Challenges and opportunities for improvement. *Pharmacology Research & Perspectives*, 12(1), e1178.
2. Guilding, C., et al. (2025). Developing an international concept-based curriculum for pharmacology education: The promise of core concepts and concept inventories. *British Journal of Clinical Pharmacology*, 91(8), 2142–2150.
3. Netere, A. K., et al. (2026). Development and validation of Pharmacology Concept Inventory for Concept-Based Learning: Leveraging theory, expert insights, and student perspectives. *Pharmacology Research & Perspectives*, 14(2), e70237.
4. White, P. J., et al. (2023). Identifying the core concepts of pharmacology education: A global initiative. *British Journal of Pharmacology*, 180(9), 1197–1209.
5. Guilding, C., et al. (2024). Defining and unpacking the core concepts of pharmacology: A global initiative. *British Journal of Pharmacology*, 181(3), 375–392.
6. Netere, A. K., & Babey, H. T. (2026). AM., et al., *Inside the learners' minds: Exploring response-process validity and diagnosing misconceptions in the Pharmacology Concept Inventory [under-review manuscript]*. BMC Medical Education.
7. Quesnelle, K. M., et al. (2021). Design of a foundational sciences curriculum: Applying the ICAP framework to pharmacology education in integrated medical curricula. *Pharmacology Research & Perspectives*, 9(3), e.
8. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
9. Arican, M., & Kuzu, O. (2020). Diagnosing preservice teachers' understanding of statistics and probability: Developing a test for cognitive assessment. *International Journal of Science and Mathematics Education*, 18(4), 771–790.
10. Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17–27.
11. Brown, J. D. (2013). *Classical test theory*, *The Routledge Handbook of Language Testing*. Routledge.
12. Crocker, L. M. A. (2008). J., *Introduction to Classical and Modern Test Theory*. Cengage Learning.
13. De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford.

14. Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Jingshen Yixue*, 26(3), 171–177.
15. Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
16. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.
17. Denison, D. B., & Secolsky, C. (2012). *Item Response Theory in Measurement, Assessment, and Evaluation for Higher Education: Steven J. Osterlind and Ze Wang* (pp. 170–180). Routledge.
18. Adedoyin, O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4), 992–1011.
19. Bradshaw, L., & Templin, J. (2014). Combining Item Response Theory and Diagnostic Classification Models: A Psychometric Model for Scaling Ability and Diagnosing Misconceptions. *Psychometrika*, 79(3), 403–425.
20. Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling*, 3(1), 62–72.
21. Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the journal of the society for social work and research. *Journal of the Society for Social Work and Research*, 1(2), 99–103.
22. Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838.
23. Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411–423.
24. Hershberger, S., Everitt, B. S., & Howell, D. C. (2005). *Tetrachoric Correlation* (pp. 2027–2028). John Wiley & Sons, Ltd.
25. Kyereh, F. K., et al. (2026). Development and psychometric validation of a questionnaire assessing perception and acceptance of micronutrient-fortified bouillon cubes among non-index household members aged ≥ 15 years in northern Ghana. *Bmc Public Health*, 26(1), 456.
26. Matsunaga, M. (2010). How to factor-analyse your data right: Do's, Don'ts, and How-To's. *International Journal of Psychological Research*, 3(1), 97–110.
27. Longman, R. S., et al. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behav Res*, 24(1), 59–69.
28. Glorfeld, L. W. (1995). An improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. *Educational and Psychological Measurement*, 55(3), 377–393.
29. Marsh, H. W., et al. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and Set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, 55(1), 102–119.
30. Marsh, H., & Alamer, A. (2024). When and how to use set-exploratory structural equation modelling to test structural models: A tutorial using the R package lavaan. *British Journal of Mathematical & Statistical Psychology*, 77(3), 459–476.
31. Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, 23(1), 116–139.
32. Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491.
33. Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education*, 90(5), 536–545.
34. Wren, D., & Barbera, J. (2013). Gathering evidence for validity during the design, development, and qualitative evaluation of thermochemistry concept inventory items. *Journal of Chemical Education*, 90(12), 1590–1601.
35. Haladyna, T., and and, & Rodriguez, M. (2013). *Developing and validating test items*. Routledge.
36. Rezigalla, A. A., et al. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(1), 445.

37. Schroeders, U., & Gnamb, T. (2025). Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science*, 8(1), 25152459251314798.
38. Affandy, H. (2021). *Calibration for instrument argumentation skills on the subject of fluid statics using item response theory*. *Journal of Physics: Conference Series*, 1842(1): p. 012032.
39. Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391–409.
40. Dong, Y., et al. (2021). An optimal choice of cognitive diagnostic model for second language listening comprehension test. *Frontiers in Psychology*, 12, 608320.
41. Mi, S., et al. (2023). Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics. *Physical Review Physics Education Research*, 19(1), 010114.
42. Figueroa, A., et al. (2023). *Generalized Cohen's Kappa: A novel inter-rater reliability metric for non-mutually exclusive categories* (pp. 19–34). Springer.
43. Schumacker, R. E., & Lomax, R. G. (2016). *SEM Basics' in A Beginner's Guide to Structural Equation Modeling* (pp. 106–143). Routledge, Editor.
44. Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1), 1–27.
45. Martinková, P., et al. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sciences Education*, 16(2), rm2.
46. Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2025). *CDM: Cognitive diagnosis modeling (Version 8.3–14) [R package]*. CRAN.
47. Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. 2002, New York, NY: Springer New York.
48. Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2), 123–140.
49. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.
50. de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, 76(2), 179–199.
51. Bauer, L. A. (2014). *Applied clinical pharmacokinetics*. Third edition. ed. McGraw-Hill's AccessPharmacy. New York, N.Y: McGraw-Hill Education LLC.
52. Whalen, K., Feild, C., & Radhakrishnan, R. (2019). *Pharmacology*. Seventh edition. ed. Lippincott illustrated reviews. Philadelphia: Wolters Kluwer.
53. Kiviniemi, T. (2018). *Development of a chemistry concept inventory for general chemistry students at Norwegian and Finnish universities*.
54. Babey, A. M., et al. (2025). Evaluating student understanding of core pharmacokinetic concepts. *European Journal of Pharmacology*, 990, 177256.
55. Kelly-Laubscher, R., et al. (2025). Evaluating student understanding of pharmacodynamics core concepts. *European Journal of Pharmacology*, 990, 177257.
56. Keijsers, C. J. P. W., et al. (2014). A comparison of medical and pharmacy students' knowledge and skills of pharmacology and pharmacotherapy. *British Journal of Clinical Pharmacology*, 78(4), 781–788.
57. Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12), 1133–1139.
58. Manias, E., & Bullock, S. (2002). The educational preparation of undergraduate nursing students in pharmacology: clinical nurses' perceptions and experiences of graduate nurses' medication knowledge. *International Journal of Nursing Studies*, 39(8), 773–784.

Figures

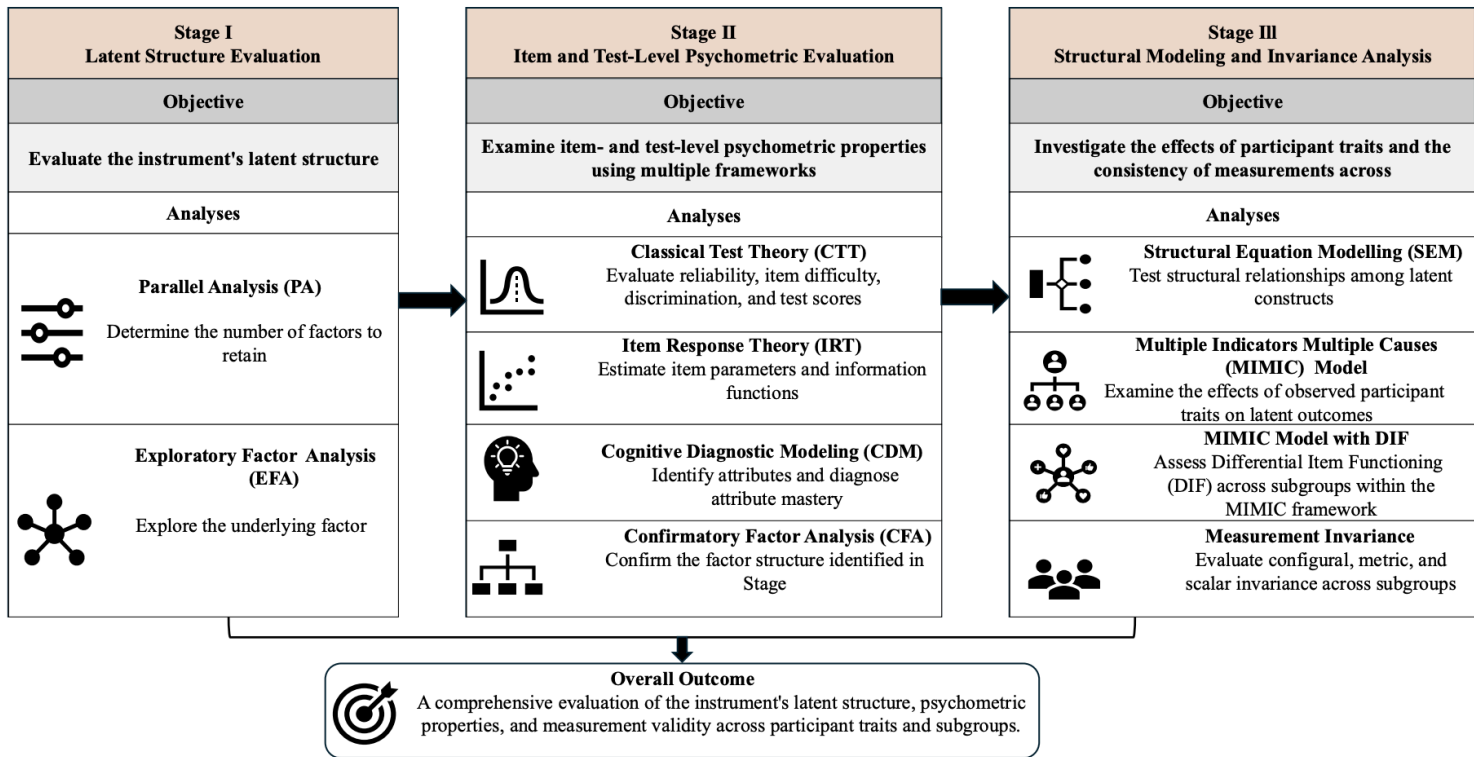


Figure 1

The PCI structural, psychometric and diagnostic evaluation

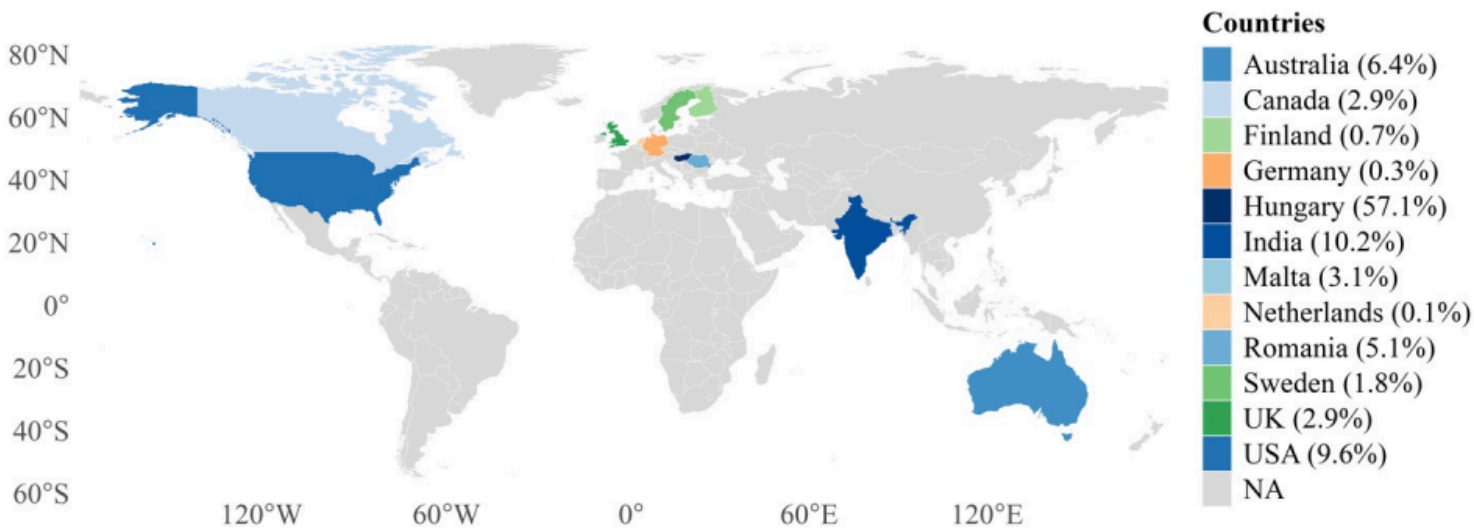


Figure 2

Geographic distribution of participants by country

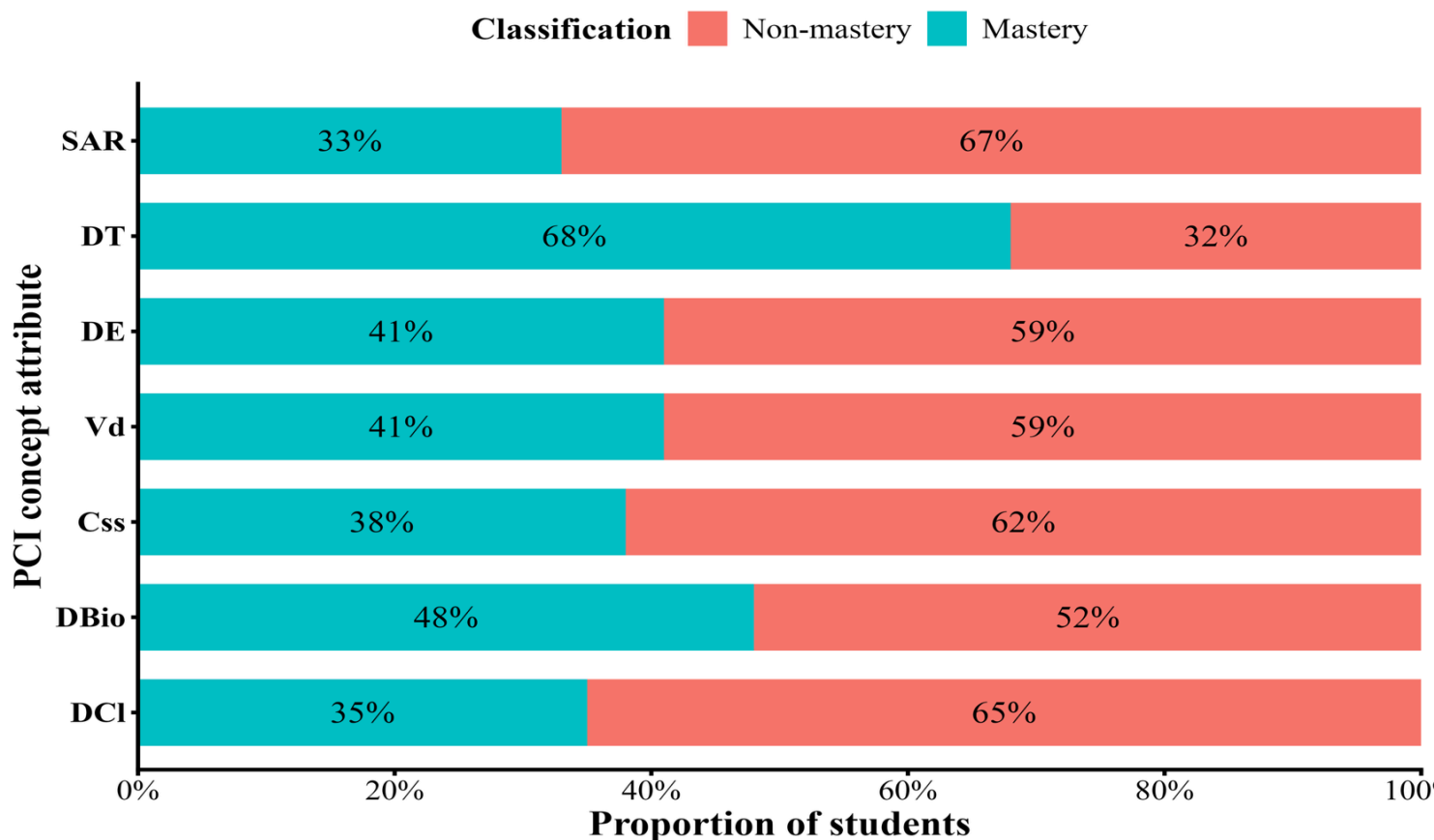


Figure 3

Probabilistic concept mastery profile derived from the exploratory cognitive diagnostic mode

Note: This figure shows the proportions representing the estimated likelihood of mastery for each pharmacological concept attribute (drug clearance (DCI), drug bioavailability (DBio), steady-state concentration (C_{ss}), volume of distribution (V_d), drug efficacy (DE), drug tolerance (DT), and structure–activity relationships (SAR)). Given the substantial item misfit and item–attribute misalignment identified in the CDM analysis, these estimates should be interpreted cautiously as indicative of general proficiency patterns across interconnected pharmacological concepts, rather than as precise measures of domain-specific mastery.

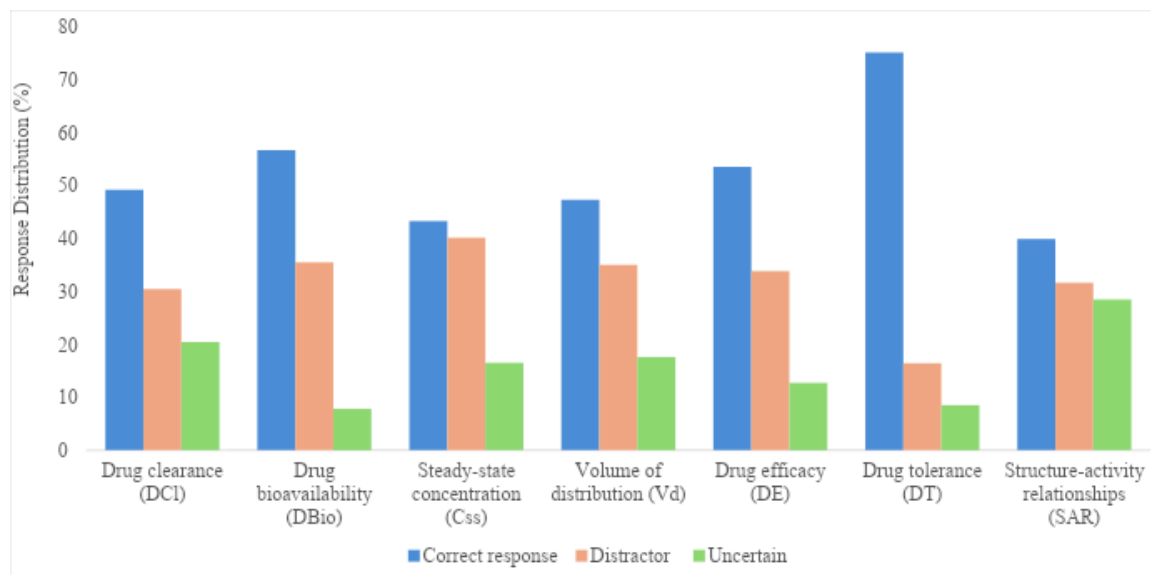


Figure 4

Response distribution across Pharmacology Concept Inventory (PCI) concept domains

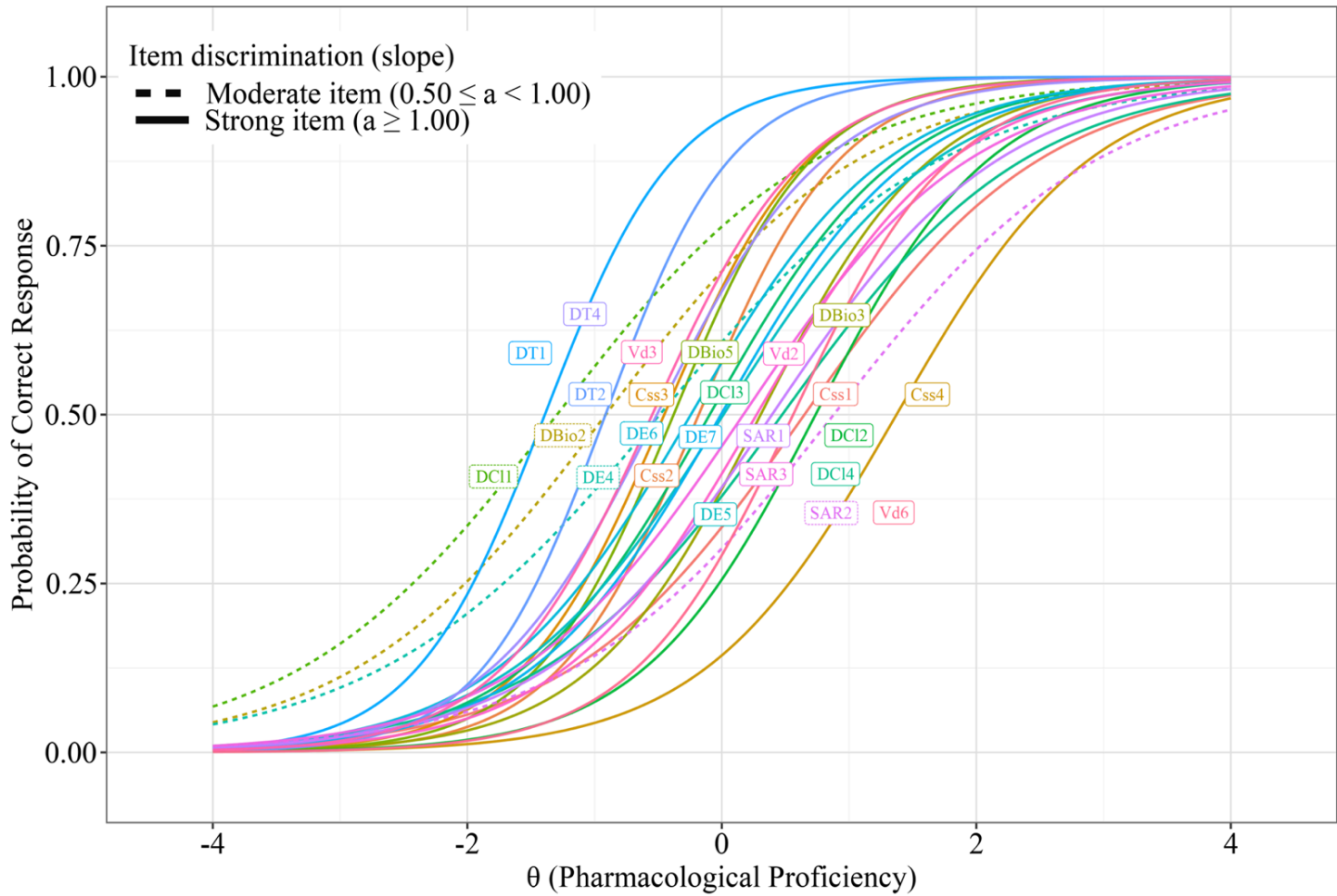


Figure 5

Item characteristic curves (ICCs) of all items of the Pharmacology Concept Inventory (PCI) plot against latent ability

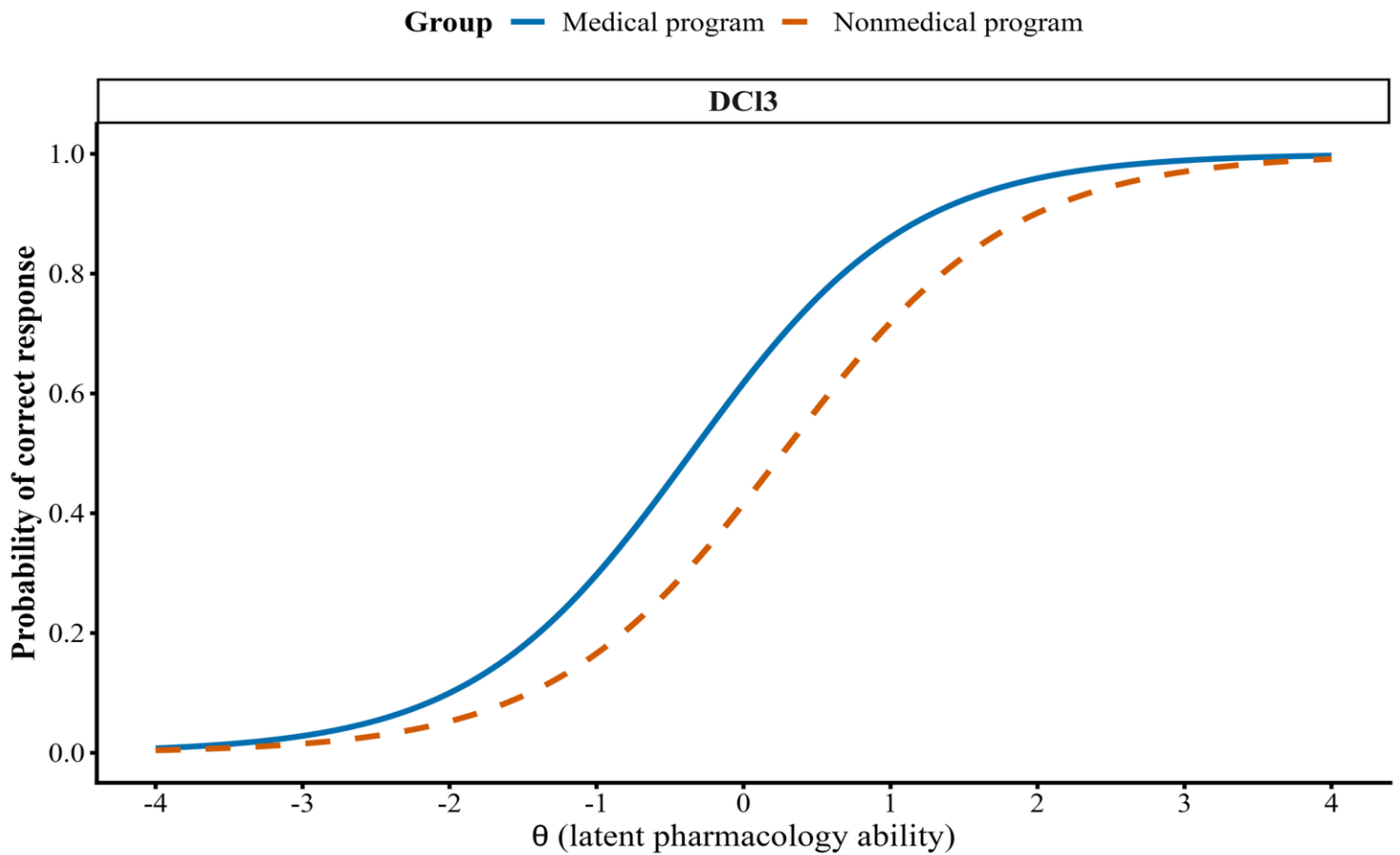


Figure 6

Item characteristic curve for DCI3 demonstrating confirmed DIF across program type

Note: The program type was determined by participants' self-reported field of study. The medical program group included students studying medicine, while the non-medical group comprised students from various disciplines, including pharmacy, nursing, veterinary medicine, biomedical sciences, and related health and life sciences. These classifications were based on the educational systems specific to participants' local institutions and countries.

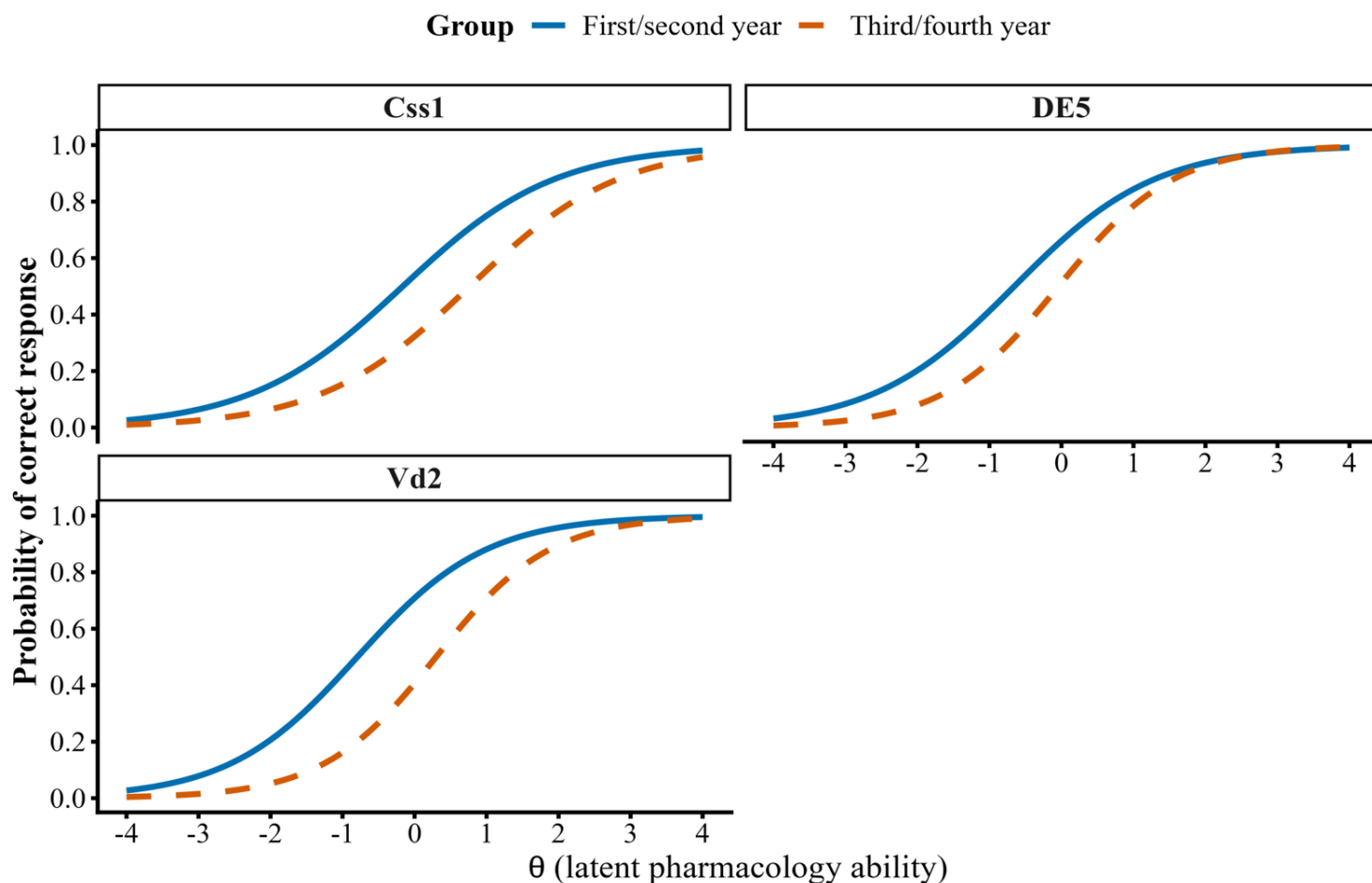


Figure 7

Item characteristic curves for Css1, DE5, and Vd2 demonstrating confirmed DIF across year of study

NB: Participants' year of study was based on self-reported progress in their current program and was categorized as either first/second year or third/fourth year. These categories reflect progress in the current program, not total years of higher education. For example, students in graduate-entry professional programs were classified by their current program year, regardless of prior degrees. A first-year student in a graduate-entry medical program was classified as a first-year student, regardless of any previously completed undergraduate degree.

Supplementary Files

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

- [AppendixS1Sampleitemdesignstructureanditerativerefinementprocessacrossvariousphases.docx](#)
- [AppendixS2CognitiveDiagnosticAssessmentCDAModels.docx](#)
- [TableS1Demographiccharacteristicsofparticipantsinthesurvey.docx](#)
- [TableS2TetrachoriccorrelationmatrixusedforlatentstructureanddimensionalityanalysesofthePCI.docx](#)
- [TableS3ParallelAnalysisPAcomparisonofobservedandsimulatedeigenvaluesforthePCI.docx](#)
- [TableS4PatternmatrixofPCIitemloadingsandpsychometricclassification.docx](#)
- [TableS5Confirmatoryfactorunidimensionalmodelandpsychometricclassification.docx](#)