Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What can a national medical examination diagnose? A cross-sectional cognitive diagnostic modelling study from Brazil

Luise Simonetti
Faculdades Pequeno Príncipe
Ana Luísa Oliveira Dornela
Faculdades Pequeno Príncipe
Etyany Lays Mariano Costa
Faculdades Pequeno Príncipe
Giulia Gil Lopes
Faculdades Pequeno Príncipe
Regiana Quinto Souza
Universidade Estadual de Santa Cruz
Carlos Fernando Collares
carlos.collares@professor.fpp.edu.br
Faculdades Pequeno Príncipe

Research Article

Keywords: Cognitive diagnostic models, Clinical competence, Content specificity, Situated cognition, Educational measurement, Psychometrics, Q-matrix, Education, medical, undergraduate, Brazil

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11406696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. C.F.C. is an Editorial Board member of BMC Medical Education and had no involvement in the editorial handling or peer review of this manuscript. The other authors declare that they have no competing interests.

Abstract

Background

Clinical competence is content-, context- and task-specific, yet national medical examinations usually report a single score. We used cognitive diagnostic modelling to compare attribute frameworks reflecting these views (clinical tasks, content areas, care settings and patient features) in the first edition of Brazil's national examination of final-year medical students (ENAMED, 2025), and to examine what the raw score reveals about attribute mastery.

Methods

Cross-sectional psychometric analysis of public microdata from 39,256 final-year medical students who answered 90 multiple-choice items (two booklets aligned to a common item order). Ten Q-matrices representing four axes (clinical tasks, content areas, care settings and patient features, in original and revised versions) were fitted with the saturated G-DINA model. We compared absolute fit ($RMSEA_2$, SRMSR), relative fit (BIC) and classification accuracy; estimated attribute prevalence, mastery profiles and latent correlations; and used receiver operating characteristic (ROC) analysis to relate raw scores to full mastery.

Results

All specifications showed close absolute fit ($RMSEA_2$ 0.011–0.013; SRMSR 0.022–0.027), but the clinical-task Q-matrix had clearly the best relative fit (BIC 3,744,145 versus 3,755,939–3,767,002 for the other axes). The five tasks were distinct but positively correlated (r = 0.03–0.71); diagnostic-test reasoning was the least mastered task (35.4%) and the one missing in two-thirds of examinees who mastered exactly four tasks. Content areas behaved largely as a single clinical competence (inter-area r = 0.72–0.80), except mental health and public health, which formed a separate pair. Among care settings, only community and territory emerged as distinct. Patient age and sex did not define distinct competences, whereas sexual and reproductive health and social determinants of health formed largely independent dimensions. Revisions based on item rereadings improved relative fit. Raw scores identified full mastery accurately (area under the ROC curve 0.965–0.990; Youden-optimal cut scores 58–63 of 90) but could not indicate which attribute an examinee lacked.

Conclusions

In this national examination, a clinical-task framework yields the most informative diagnostic description of performance. Reporting attribute profiles by clinical task, alongside the overall result, could give students and medical schools actionable feedback, particularly on diagnostic-test reasoning.

Background

Large-scale licensing and graduation examinations occupy a central place in the regulation of medical education, yet their contribution to learning depends on what they report back to students and schools [1, 2]. In 2025, Brazil introduced a national examination of graduating medical students, the *Exame Nacional de Avaliação da Formação*

Médica (National Examination for the Assessment of Medical Training, ENAMED) [3, 4]. As of the date of submission of this manuscript, ENAMED is not a licensing examination, since registration to practise does not depend on its result, but it is national in scope and consequential. It is the medical modality of the National Student Performance Examination (Enade), taken by the final-year students of every medical school; its results are used to calculate an official quality indicator for each school; and it also serves as the objective test for direct-entry residency selection [4]. Its 100 multiple-choice items are written against a common reference matrix aligned with the national curricular guidelines, which organise medical training around internal medicine, surgery, paediatrics, obstetrics and gynaecology, family and community medicine, public health and mental health [4, 5]. As in most examinations of this kind, individual results are reported as a single proficiency score, estimated with item response theory and interpreted against a standard set with a modified Angoff procedure, supplemented by counts of correct answers in five broad areas [4].

A single score supports pass/fail and accountability decisions, but it says little about which capabilities an examinee has or lacks. Two examinees with the same score may have very different strengths, and area subscores often add little information beyond the total score when areas are highly correlated [6, 7]. Contemporary views of assessment emphasise that the meaning and use of scores, rather than the instrument itself, must be validated, and that assessment should feed learning as well as certification [8–10].

Cognitive diagnostic models (CDMs), also called diagnostic classification models, address this gap by classifying each examinee as master or non-master of a set of discrete attributes, using a Q-matrix that specifies which attributes each item requires [11, 12]. CDMs have been advocated for health professions education because they turn existing examinations into sources of fine-grained feedback [13], and they have been applied to medical licensing and subject examinations, typically with content domains as attributes [14, 15], and to competency-based situational judgement tests [16]. Their usefulness, however, hinges on the Q-matrix: misspecified or poorly chosen attributes degrade parameter estimates and classification accuracy [17], and many published applications use a single, retrofitted Q-matrix without examining alternatives [18].

Choosing the attributes of a CDM is therefore a theoretical decision as much as a technical one, and current understanding of clinical competence offers several candidates. One of the most robust findings in medical education is that performance is specific rather than general: how well a clinician handles one problem predicts only modestly how well they handle another. First described as content specificity in studies of medical problem solving [19], this phenomenon is the main reason why assessments of clinical competence require broad sampling across cases [20–22]. Generalisability studies of performance assessments, in medicine and elsewhere, have identified the interaction between examinee and task as one of the largest sources of score variation [23, 24]. Situativity theories, which include situated cognition and situated learning, as well as distributed and embodied cognition, go further by locating competence in the interaction among the clinician, the patient, and the environment rather than in the individual alone [25–28]. From this perspective, the setting of an encounter and the characteristics of the patient are contextual factors that can shape clinical reasoning [29–31], and competence is better understood as context-dependent than as a stable possession of the individual [32]. Clinical competence is thus content-, context-, and task-specific.

Each of these perspectives suggests a different way of describing what a clinical vignette item measures: by its content area, by the care setting in which it is placed, by the characteristics of the patient it describes, or by the clinical task it demands (for example, recognising a syndrome or managing an emergency). Expressed as Q-matrices, these descriptions become competing measurement models whose fit to the same responses can be

compared. They also imply different diagnostic stories, and they matter for the design of feedback, for curriculum review and for equity monitoring. Whether such frameworks are equally supported by examinee responses is an empirical question that, to our knowledge, has not been examined for a national medical examination.

This study therefore applied the G-DINA model to the complete public response data from ENAMED 2025, using ten Q-matrices representing four attribute axes. We aimed to (1) compare the fit and classification quality of clinical-task, content-area, care-setting and patient-feature Q-matrices; (2) describe attribute prevalence, mastery profiles and latent attribute associations under the best-supported specification of each axis; (3) evaluate whether revisions of the Q-matrices grounded in item reading improved the models; and (4) examine how well the raw score identifies examinees who master all attributes and what it fails to reveal.

Methods

Design, setting and data

This was a cross-sectional psychometric secondary analysis of the public microdata of ENAMED 2025 released by the Brazilian National Institute for Educational Studies and Research Anísio Teixeira (INEP) [4]. The examination was administered on 19 October 2025 in 225 municipalities and consisted of 100 four-option multiple-choice items [4]. We analysed the scored (0/1) responses of final-year undergraduate medical students assessed under the ENADE component of the examination; graduates sitting ENAMED only for residency selection are released in separate files and were not included. The microdata exclude courses with fewer than two valid examinees and contain no directly identifying information [4].

Ten of the 100 items were not used: seven items were annulled by INEP and three were withdrawn from calibration because of non-positive biserial correlations, as documented in the item-parameter file of the microdata. The remaining 90 items were analysed. All items were also read in full by the research team from the published booklet.

Alignment of the two test booklets

The examination was administered in two booklets, each containing the same items in a different order. Responses from booklet 2 were re-ordered into the booklet-1 sequence using the item correspondence table supplied by INEP (booklet-1 and booklet-2 item numbers). The correspondence was verified independently by matching the full text of all 100 item stems across the two booklets, which produced a one-to-one pairing identical to INEP's table. Because booklets are distributed at random, item difficulty should be equivalent across them; after alignment, item proportions correct correlated at $r = 0.996$ between booklets, and mean scores were 55.7 (booklet 1, $n = 19{,}670$) and 55.8 (booklet 2, $n = 19{,}586$). As a further check, we compared the prevalence of each attribute between booklets under every fitted model.

Attribute axes and Q-matrices

Four attribute axes were defined to represent task specificity (clinical tasks), content specificity (content areas) and context specificity (care settings and patient features), each represented by a Q-matrix in which an entry of 1 indicates that the attribute is required to answer the item correctly (Table 1; complete matrices in Additional file

1). The matrices were developed by the authors through a full reading of each item stem and its options, and received final review by the senior authors, with disagreements resolved through discussion. The clinical-task axis comprised five attributes describing what the item asks the examinee to do: syndromic recognition (SR), diagnostic-test reasoning (DTR), emergency and urgent care management (EUM), outpatient therapeutic planning (OTP) and population-health action (PHA). This matrix had been developed and refined in previous analyses by the senior author; 57 of the 90 items required two or three tasks.

The content-area axis was first coded with nine single-attribute groupings (seven areas plus two small residual groupings). The revised version adopted the seven areas of the ENAMED reference matrix (internal medicine, surgery, obstetrics and gynaecology, paediatrics, family and community medicine, mental health and public health) [3, 4]. Item reading led to the reassignment of the principal area of 28 items (for example, epidural haematoma and complicated diverticulitis from paediatrics to surgery), seven of which had belonged to the residual groupings. Because mental health and public health are defined as interdisciplinary, one of these was added as a second attribute in 17 items; one additional item was coded for both obstetrics and gynaecology and for family and community medicine, resulting in 18 multi-attribute items.

The care-setting axis was initially coded into nine settings (primary care, specialised outpatient care, emergency care, prehospital care, hospital ward, intensive care, obstetric unit, operating theatre, and community/territory). Because prehospital care, intensive care and the operating theatre were represented by only two to four items each, the revised version merged prehospital into emergency care and ward, intensive care and operating theatre into hospital care, yielding six settings; one item set in an obstetric emergency department additionally received the emergency-care attribute.

The patient-feature axis was coded with 25 descriptors of the patient or population in each vignette, covering life stage, sex and gender identity, sociodemographic characteristics, conditions of vulnerability, sexual and reproductive health, and items without an individual patient. A saturated model with 25 attributes would require more than 33 million latent classes, and several descriptors appeared in only one to three items, so four theoretically motivated specifications were derived: life cycle (five attributes), detailed life cycle (eight), sex or gender (three) and an integrated matrix (seven) that added sexual and reproductive health (SRH) and social determinants of health and vulnerability (SDV) to the life-stage and collective attributes. In the life-cycle and integrated specifications, childhood, adolescence, adulthood, and older age are referred to as life stages, and items without an individual patient are grouped into a separate collective attribute (seven items). Adulthood was assigned to two vignettes of pregnant women whose age was not stated, and two items without an individual patient that had not been coded as collective (a prison scabies outbreak and care for riverside and Indigenous communities) were added to that attribute.

Table 1
Attribute axes and Q-matrix specifications analysed

Axis / specification	K	Attributes (number of items)	Multi-attribute items
Clinical tasks			
Tasks, original	5	SR (55), DTR (29), EUM (21), OTP (34), PHA (21)	57
Tasks, revised	5	SR (55), DTR (32), EUM (21), OTP (34), PHA (21)	57
Content areas			
Content, original	9	IM (15), SUR (9), PED (18), OBG (15), FCM (10), PH (9), MH (7), residual 1 (3), residual 2 (4)	0
Content, revised	7	IM (18), SUR (11), OBG (16), PED (17), FCM (21), MH (8), PH (17)	18
Care settings			
Settings, original	9	PC (46), SOC (21), EC (16), PHC (2), WARD (5), ICU (4), OC (5), OT (3), CT (12)	24
Settings, revised	6	PC (46), SOC (21), EC (19), HC (12), OC (5), CT (12)	25
Patient features			
D1 Life cycle	5	CHI (10), ADO (9), ADU (55), OLD (9), COL (7)	0
D2 Detailed life cycle	8	NEO (3), INF (3), CHD (4), ADO (9), YAD (36), MAD (19), OLD (9), COL (7)	0
D3 Sex or gender	3	FEM (40), MAL (32), OTH (18)	0
D4 Integrated	7	CHI (10), ADO (9), ADU (55), OLD (9), SRH (15), SDV (23), COL (7)	34

K, number of attributes. Numbers in parentheses are the number of items requiring each attribute (90 items in all). Multi-attribute items require more than one attribute. In the original content matrix, the two residual groupings (3 and 4 items) were not named areas. CHI, childhood (neonate, infant and child); ADO, adolescence; ADU, adulthood; OLD, older adults; COL, collective (no individual patient). Clinical tasks: SR, syndromic recognition; DTR, diagnostic-test reasoning; EUM, emergency and urgent care management; OTP, outpatient therapeutic planning; PHA, population-health action. Content areas: IM, internal medicine; SUR, surgery; OBG, obstetrics and gynaecology; PED, paediatrics; FCM, family and community medicine; MH, mental health; PH, public health. Care settings: PC, primary care; SOC, specialised outpatient care; EC, emergency care; PHC, prehospital care; WARD, hospital ward; ICU, intensive care; OC, obstetric care; OT, operating theatre; HC, hospital care (WARD + ICU + OT); CT, community and territory. Patient features: NEO, neonate; INF, infant; CHD, child; YAD, young adult; MAD, middle-aged adult; FEM, female; MAL, male; OTH, sex not stated, gender-diverse or collective; SRH, sexual and reproductive health; SDV, social determinants of health and vulnerability.

Cognitive diagnostic modelling

Each Q-matrix was fitted with the generalised deterministic inputs, noisy “and” gate (G-DINA) model [33], which allows each combination of the attributes required by an item to have its own success probability and nests the more restrictive DINA model. The joint attribute distribution was left saturated, so that associations among attributes were estimated freely. Parameters were estimated by marginal maximum likelihood with the expectation–maximisation algorithm, with monotonicity constraints, using the GDINA package (version 2.13.2) [34] in R (version 4.3.3) [35].

Absolute fit was assessed with the limited-information M_2 statistic, its root mean square error of approximation ($RMSEA_2$) and the standardised root mean square residual (SRMSR) [36, 37]; with almost 40,000 examinees, the M_2 test itself was expected to reject any model, so $RMSEA_2$ and SRMSR were interpreted as effect sizes. Item-level fit was examined through the standardised difference between observed and predicted proportions correct and through residual correlations and log-odds ratios for item pairs, with adjustment for multiple comparisons [38]. Relative fit across Q-matrices was compared with the Bayesian information criterion (BIC); because all models were fitted to the same responses, BIC differences indicate which attribute structure accounts for the data more parsimoniously. For Q-matrices containing multi-attribute items, G-DINA was compared with DINA; for single-attribute matrices, the two models have identical item response functions.

Examinees were classified as masters of an attribute when its marginal posterior probability was 0.5 or higher, and attribute profiles were formed from these classifications. Classification accuracy at the pattern and attribute levels was estimated using the indices implemented in GDINA [39, 40]. Latent associations between attributes were expressed as phi coefficients computed from the estimated joint attribute distribution.

Q-matrix validation

Empirical Q-matrix validation used the proportion of variance accounted for (PVAF) method with the default cut-off of 0.95 [41]. Because this cut-off is known to favour over-specified q-vectors and to suggest changes that are not substantively plausible, particularly in large samples [42], PVAF suggestions were treated as auxiliary information. Revisions were accepted only when the presence of the attribute was supported by the content of the item.

Raw score and full mastery

For each specification, the raw score (number correct, 0–90) was examined as a predictor of full mastery (classification as master of all attributes) using receiver operating characteristic (ROC) analysis with the pROC package [43]; 95% confidence intervals for the area under the curve (AUC) used DeLong’s method [44]. For every integer cut score from 30 to 85 we computed sensitivity, specificity and overall accuracy, and identified the cut scores that maximised Youden’s index [45] and accuracy. These cut scores describe the correspondence between the two scoring approaches and are not proposed as standards for the examination. Score distributions were also compared across the number of attributes mastered. Detailed results for every specification are provided in Additional file 2, the analysis code in Additional file 3 and the aligned response matrix analysed in this study (booklet-2 responses re-ordered into the booklet-1 item sequence) in Additional file 4.

Ethical considerations

The study used only public, de-identified secondary data. Under Brazilian regulations, research using publicly accessible or public-domain information is not subject to registration or review by the research ethics committee system [46].

Results

Examinees and items

The analysis included 39,256 examinees. The mean raw score was 55.7 of 90 (SD 10.9; median 56; range 0–88), and Kuder–Richardson 20 reliability was 0.87. Item proportions correct ranged from 0.10 to 0.96 (mean 0.62), and corrected item–total correlations from 0.00 to 0.50 (median 0.26).

Comparison of Q-matrices

Table 2 summarises the ten specifications, and detailed results for each are given in Additional file 2. Absolute fit was close for all of them ($RMSEA_2$ 0.011–0.013; SRMSR 0.022–0.027). Every model reproduced the observed proportion correct of all 90 items, but residual correlations for some item pairs remained statistically significant for every item in every model (median of the largest absolute residual z per item 12.2–13.3 across the selected models), indicating residual local dependence. Relative fit discriminated sharply between axes: the clinical-task matrices had the lowest BIC (3,743,972 revised; 3,744,145 original), followed at a distance of more than 11,000 points by the integrated patient-feature matrix (3,755,939) and by the revised care-setting (3,759,650) and content-area (3,759,951) matrices. Within each axis, revisions grounded in item reading improved relative fit while reducing or modestly changing the number of parameters. For every Q-matrix with multi-attribute items, G-DINA fitted better than DINA. Attribute prevalences differed by at most 1.8 percentage points between booklets in any model. Because revising the task matrix improved fit only slightly and left all findings unchanged, results below are reported for the original, pre-specified task matrix and for the revised content and setting matrices and the integrated patient-feature matrix.

Table 2
Fit and classification accuracy of the ten Q-matrix specifications (G-DINA, saturated attribute distribution)

Specification	K	Npar	BIC	ΔBIC	$RMSEA_2$	SRMSR	CA pattern [attributes]	DINA BIC	PVAF (%)
Tasks, original	5	377	3 744 145	173	0.012	0.023	0.67 [0.83–0.91]	3 762 783	10.9
Tasks, revised	5	389	**3 743 972**	0	0.011	0.022	0.66 [0.82–0.91]	3 763 109	12.0
Content, original	9	691	3 765 858	21 886	0.012	0.027	0.76 [0.87–0.93]	—	7.0
Content, revised	7	343	3 759 951	15 979	0.012	0.026	0.70 [0.85–0.93]	3 762 735	7.3
Settings, original	9	739	3 762 618	18 646	0.011	0.026	0.38 [0.67–0.94]	3 768 348	9.1
Settings, revised	6	293	3 759 650	15 678	0.012	0.026	0.62 [0.77–0.94]	3 765 281	10.7
D1 Life cycle	5	211	3 767 002	23 030	0.013	0.026	0.81 [0.87–0.96]	—	3.3
D2 Detailed life cycle	8	435	3 766 668	22 696	0.012	0.027	0.68 [0.82–0.95]	—	4.7
D3 Sex or gender	3	187	3 766 675	22 703	0.013	0.026	0.88 [0.92–0.96]	—	5.2
D4 Integrated	7	391	3 755 939	11 966	0.012	0.025	0.46 [0.72–0.96]	3 766 818	7.8

Npar, number of estimated parameters; ΔBIC, difference from the lowest BIC; $RMSEA_2$, root mean square error of approximation of M_2; SRMSR, standardised root mean square residual; CA, classification accuracy for the complete attribute pattern and, in brackets, the range across attributes; DINA BIC is reported only for matrices with multi-attribute items (for single-attribute matrices DINA and G-DINA are equivalent); PVAF, percentage of Q-matrix cells for which the PVAF method (cut-off 0.95) suggested a change. Lowest BIC in bold.

Clinical tasks

Mastery prevalence ranged from 35.4% for diagnostic-test reasoning to 50.1% for population-health action (Fig. 1A). Pattern-level classification accuracy was 0.67, and attribute-level accuracy ranged from 0.83 (DTR) to 0.91 (SR). Thirty of the 32 possible profiles were observed; the most frequent were non-mastery of all tasks (32.2%), mastery of all tasks (19.6%), and mastery of all tasks except DTR (9.8%) (Table 3). Of the 5,819 examinees (14.8%) who mastered exactly four tasks, 3,841 (66.0%) lacked DTR and 1,948 (33.5%) lacked SR. The

tasks were positively but heterogeneously associated (Fig. 2A): the strongest correlations linked SR with PHA (0.71) and DTR, EUM and OTP with one another (0.57–0.67), whereas SR and DTR were almost unrelated (0.03). PVAF suggested changes in 10.9% of the cells, many of them implausible (for example, adding SR and DTR to an item on the sensitivity and specificity of a screening test). Item reading supported adding DTR to three items whose correct answer depended on interpreting laboratory or cytology results; this revision lowered the BIC by 173 points without changing the substantive findings.

Percentage of the 39,256 examinees classified as masters (marginal posterior probability ≥ 0.5) of each attribute under (A) the clinical-task Q-matrix, (B) the revised content-area Q-matrix, (C) the revised care-setting Q-matrix, and (D) the integrated patient-feature Q-matrix, all fitted with the saturated G-DINA model. Attribute abbreviations as in Table 1.

Phi coefficients computed from the estimated joint distribution of the binary attributes for (A) clinical tasks, (B) content areas (revised), (C) care settings (revised) and (D) patient features (integrated). Blue indicates a positive and red a negative association. Attribute abbreviations as in Table 1.

Table 3
Most frequent attribute profiles under the selected specification of each axis

Specification (attribute order) and profile	Examinees (%)
Clinical tasks, original (SR-DTR-EUM-OTP-PHA)	
00000	32.2
11111	19.6
10111	9.8
10001	7.0
01110	5.9
Content areas, revised (IM-SUR-OBG-PED-FCM-MH-PH)	
0000000	37.2
1111111	34.0
0000001	4.8
0000011	4.5
1101011	0.9
Care settings, revised (PC-SOC-EC-HC-OC-CT)	
111111	29.7
000001	23.9
000000	22.4
101000	2.7
111001	2.4
Patient features, integrated (CHI-ADO-ADU-OLD-SRH-SDV-COL)	
1111111	26.4
0000110	22.1
0000000	18.3
0000111	6.8
1111101	6.4

Profiles list mastery (1) or non-mastery (0) of each attribute in the order shown in the first column. Percentages refer to all 39,256 examinees. SR, syndromic recognition; DTR, diagnostic-test reasoning; EUM, emergency and urgent care management; OTP, outpatient therapeutic planning; PHA, population-health action; IM, internal medicine; SUR, surgery; OBG, obstetrics and gynaecology; PED, paediatrics; FCM, family and community medicine; MH, mental health; PH, public health; PC, primary care; SOC, specialised outpatient care; EC, emergency care; HC, hospital care; OC, obstetric care; CT, community and territory; CHI, childhood; ADO, adolescence; ADU, adulthood;

OLD, older adults; SRH, sexual and reproductive health; SDV, social determinants of health and vulnerability; COL, collective.

Content areas

The original nine-attribute content matrix required 691 parameters; its attributes correlated uniformly (0.69–0.79), and 80.0% of examinees were classified as masters of none (45.8%) or all (34.2%) of the attributes. The revised seven-area matrix halved the number of parameters (343) and lowered the BIC by 5,907 points. Prevalence was highest for public health (57.0%) and mental health (50.6%) and lowest for obstetrics and gynaecology (41.8%) (Fig. 1B). The five clinical areas correlated strongly with one another (0.72–0.80), whereas mental health and public health formed a pair (0.70) less strongly related to the clinical areas (0.50–0.65) (Fig. 2B). Profiles remained concentrated at the extremes (none 37.2%, all 34.0%); the most frequent intermediate profiles were mastery of public health only (4.8%) and of mental health and public health only (4.5%). Attribute-level accuracy ranged from 0.85 to 0.93.

Care settings

In the original nine-setting matrix, pattern-level accuracy was only 0.38; settings represented by two to four items were classified with low accuracy (0.67–0.81) and showed implausible prevalences (for example, 72.6% for the operating theatre). The revised six-setting matrix used 293 instead of 739 parameters, lowered the BIC by 2,968 points and raised pattern-level accuracy to 0.62 (attribute level 0.77–0.94). Community and territory was the most frequently mastered setting (67.9%) and hospital care the least (33.6%) (Fig. 1C). The five clinical settings correlated substantially with one another (0.57–0.78), whereas community and territory showed weak associations with them (0.16–0.38) (Fig. 2C). Accordingly, the second most frequent profile (23.9%) was mastery of community and territory only, after mastery of all settings (29.7%) and before non-mastery of all (22.4%).

Patient features

The life-cycle, detailed life-cycle and sex or gender specifications fitted almost identically (BIC 3,766,668–3,767,002). Their attributes were highly correlated (0.84–0.87 among the four life stages and 0.75–0.78 between these and the collective attribute; 0.83–0.90 for sex or gender; 0.55–0.84 among the seven stages of the detailed life cycle, the lowest values involving infancy, which was measured by three items), and 81.0% to 91.6% of examinees fell into the all-or-none profiles. Mastery was similar for vignettes of female (42.5%) and male (43.5%) patients and across life stages (41.1%–44.4%). The integrated matrix fitted markedly better (BIC 3,755,939). In it, SRH and SDV were nearly uncorrelated with the four life stages (− 0.06 to 0.23) and moderately correlated with each other (0.38) (Fig. 2D), and they had the highest prevalences (76.6% and 69.6%) (Fig. 1D). The second most frequent profile (22.1%) was mastery of SRH and SDV only. Classification accuracy was lower for SRH (0.72) and SDV (0.81) than for the four life stages (0.88–0.96) and the collective attribute (0.85).

Raw score and full mastery

The proportion of examinees mastering all attributes ranged from 19.2% (revised clinical tasks) to 40.7% (sex or gender). The raw score discriminated full mastery very well under every specification (AUC 0.965–0.990; Table 4 and Fig. 3A). For the clinical-task matrix, the AUC was 0.979 (95% CI 0.978–0.980); a cut of 63 maximised Youden's index (sensitivity 95.3%, specificity 89.1%) and a cut of 66 maximised accuracy (93.6%) (Fig. 3B). Optimal cuts were lower for the other axes (58–61), reflecting their higher proportions of full mastery. Raw scores rose with the number of tasks mastered, but the interquartile ranges of examinees mastering one to four tasks

overlapped and together spanned only 49 to 65 points (Fig. 4); examinees with the same score could lack different tasks.

(A) Receiver operating characteristic curves of the raw score (0–90) for identifying examinees classified as masters of all attributes, under the selected specification of each axis; the inset enlarges the upper-left region. Areas under the curve (AUC) are shown in the legend. (B) Sensitivity, specificity and accuracy of each raw-score cut for identifying full mastery of the five clinical tasks; dashed lines mark the cuts that maximise Youden's index (63) and accuracy (66).

Boxes show the interquartile range and median, and whiskers the 5th and 95th percentiles, of the raw score (0–90) for examinees mastering 0 to 5 of the clinical tasks. The dashed line marks the Youden-optimal cut for full task mastery (63).

Table 4
Raw score (0–90) as a predictor of full mastery under each Q-matrix specification

Specification	Full mastery (%)	AUC (95% CI)	Youden cut (Se/Sp, %)	Accuracy cut (Acc, %)
Tasks, original	19.6	0.979 (0.978–0.980)	63 (95.3/89.1)	66 (93.6)
Tasks, revised	19.2	0.979 (0.978–0.980)	63 (95.8/88.9)	67 (93.6)
Content, original	34.2	0.986 (0.985–0.987)	60 (93.7/93.2)	61 (93.7)
Content, revised	34.0	0.986 (0.986–0.987)	60 (94.2/93.2)	61 (93.9)
Settings, original	29.1	0.980 (0.979–0.981)	61 (93.8/91.4)	62 (92.7)
Settings, revised	29.7	0.980 (0.979–0.981)	61 (93.1/91.9)	62 (92.7)
D1 Life cycle	39.0	0.987 (0.986–0.987)	58 (96.6/91.1)	59 (93.7)
D2 Detailed life cycle	34.4	0.984 (0.983–0.985)	60 (93.3/93.2)	61 (93.3)
D3 Sex or gender	40.7	0.990 (0.989–0.990)	58 (95.8/93.1)	59 (94.3)
D4 Integrated	26.4	0.965 (0.963–0.966)	61 (93.3/88.2)	63 (90.7)

Full mastery, proportion of examinees classified as masters of all attributes of the specification. AUC, area under the receiver operating characteristic curve with DeLong 95% confidence interval. Youden cut, raw-score cut (score ≥ cut) maximising sensitivity + specificity − 1; accuracy cut, cut maximising the proportion of concordant classifications.

Discussion

Fitting ten Q-matrices to the same responses from a complete national cohort showed that the attribute framework used to read an examination is not a neutral choice. Every specification reproduced the response data closely in absolute terms, yet the clinical-task framework accounted for the responses far better than frameworks based on content areas, care settings or patient features. Under that framework, the five tasks were distinguishable and positively associated, diagnostic-test reasoning emerged as the main gap among graduating students, and the raw score, while an excellent indicator of overall mastery, could not show which task an examinee lacked.

Clinical tasks as the most informative axis

The superiority of the task framework is consistent with how clinical vignette items are written: the stem frames a patient problem, and the question asks for a decision—a diagnosis, a test, an immediate action, a treatment plan, or a population-level measure. Tasks of this kind resemble the units of professional work on which competency-based frameworks such as entrustable professional activities are built [47], and they map onto the distinction between diagnostic and management reasoning in the clinical reasoning literature [48]. The near-zero association between syndromic recognition and diagnostic-test reasoning suggests that recognising a clinical pattern and reasoning about which tests to request and how to interpret them are learned separately. That diagnostic-test reasoning was the least-mastered task and the one most often missing among examinees who mastered every other task, pointing to a specific, teachable target for curricula and feedback to schools.

Content areas and care settings: one competence with two exceptions

Content areas and care settings largely behaved as facets of a single clinical competence: most examinees were classified as masters of all or none of them, and the clinical areas and settings were strongly intercorrelated. This agrees with psychometric work showing that subscores for highly correlated content areas rarely carry information beyond the total score [6, 7], and it cautions against over-interpreting area-level results, including the counts of correct answers per area currently reported for ENAMED [4]. The two exceptions are informative. Mental health and public health, defined as interdisciplinary areas in the national curriculum [5], formed a dimension of their own, and the community and territory setting was nearly independent of the clinical settings. Items on these topics appear to draw on population-oriented knowledge—surveillance, prevention, legislation, health-system organisation—that is acquired somewhat independently of clinical reasoning, which is also consistent with the strong association between syndromic recognition and population-health action on the task axis. The large gains from revising the content and setting matrices show that an item's nominal area or setting is not always the knowledge it demands, and that attributes measured by very few items are classified unreliably [17].

Patient features and equity

The absence of distinct competences linked to patients' age or sex is reassuring from an equity perspective: graduating students who mastered vignettes about one group generally mastered vignettes about the others, and mastery rates were nearly identical for female and male patients. By contrast, items involving sexual and reproductive health and social determinants of health formed largely independent dimensions with high prevalence. A plausible interpretation is that these items share features—such as guideline-based prevention, notification and protective measures—that differ from the core clinical demands of the examination, rather than reflecting a separate competence in caring for vulnerable groups. Their lower classification accuracy and the

small number of items per descriptor also call for caution. These analyses concern features of the patients described in the items, not characteristics of the examinees; fairness for examinee groups requires differential item functioning and related analyses, which are an important complement in light of current concerns about equity in assessment [49].

Content, context and task specificity in a written examination

Read against theories of clinical competence, the results suggest where specificity resides in a written national examination. Performance was organised more by the clinical task demanded than by the content area, consistent with generalisability studies that identify the interaction between examinee and task as a major source of score variation [23, 24]. The convergence of content areas on a largely single dimension does not contradict content specificity. Specificity has been described at the level of individual cases and the knowledge they require rather than of broad disciplines [21, 22], and in our models it is most plausibly reflected in item-specific variance and in the residual dependence between item pairs that no attribute structure removed. Situativity theories predict that the setting and the patient influence performance through the dynamic interaction among the clinician, patient, and environment [27, 28]. Written vignettes convey these elements only as text, which may explain why care settings and patients' age and sex did not define distinct competences. By contrast, community and territory, sexual and reproductive health, and social determinants of health, in which the context is itself part of what must be known, did form distinct dimensions. Contextual factors of the kind studied in authentic or video-based encounters [30, 31] are therefore largely beyond the reach of a multiple-choice examination, and attribute-level reporting from such an examination should complement, not replace, workplace-based assessment within a programmatic approach [2, 8].

What the raw score can and cannot tell

Because full mastery requires mastering every attribute, it is expected to correlate closely with the number of items answered correctly, and the very high AUCs confirm that the raw score is an efficient indicator of overall mastery. The ROC analysis also offers an empirical link between total-score cut scores and attribute-based definitions of competence, which may be useful when standards are reviewed. However, the overlap of score distributions for examinees who mastered between one and four tasks shows that the same score can correspond to different profiles of strengths and gaps. Reporting attribute profiles by clinical task alongside the overall result could provide students and schools with formative information, as advocated by programmatic assessment [2, 13], without altering the examination's decision function. Any such use should be supported by validity evidence appropriate to the intended interpretation [10].

Q-matrix validation in large samples

The PVAF method proposed changes to between 3% and 12% of Q-matrix cells, often adding attributes that had no plausible role in the item. With very large samples and a high cut-off, small gains in explained variance can make over-specified q-vectors appear preferable [42]. Our experience supports using empirical validation as a prompt for expert review rather than as a decision rule: revisions grounded in item reading were substantively justified and consistently improved relative fit.

Strengths and limitations

The main strengths are the use of the complete national cohort, the comparison of several theoretically distinct attribute frameworks on identical data, the explicit treatment of Q-matrix revision, and full reproducibility from public data. Several limitations should be considered. The Q-matrices were constructed retrospectively, each coded by one author and reviewed by another; inter-rater agreement was not quantified, and the examination was not designed to measure these attributes. Residual dependence between item pairs remained in all models, which may lead to overconfident classifications. Pattern-level classification accuracy was moderate (0.46–0.70 for the four selected models), whereas attribute-level accuracy ranged from 0.72 to 0.96 and exceeded 0.85 for 18 of the 25 attributes. Mastery classifications depend on the 0.5 posterior threshold. The analysis covers a single edition of a multiple-choice examination, so stability across editions and relationships with workplace-based performance remain to be established. Finally, the analysis describes individual examinees and does not separate school-level from individual variation.

Implications and future research

For examination developers, the results suggest that blueprints specifying clinical tasks, not only content areas, would allow national examinations such as ENAMED to be designed for diagnostic reporting, with enough items per task to support reliable classification. For medical schools, task-level profiles of their graduates could inform curriculum review, especially regarding diagnostic-test reasoning. Future work should replicate the analyses in later editions, examine measurement invariance across schools and examinee groups, assess the consequential validity of attribute-level feedback, and explore cognitive diagnostic computerised adaptive testing for formative use.

Conclusions

In the first edition of Brazil's national examination of graduating medical students, the clinical tasks required by the items provided a much better account of examinee responses than content areas, care settings or patient features. Graduating students' performance was organised around five distinguishable but related tasks, with diagnostic-test reasoning as the most common gap. Content areas and settings behaved largely as a single clinical competence, with population-oriented topics as the main exception, and patient age and sex did not define distinct competences. The raw score identified full mastery accurately but could not indicate which competence was missing. Task-based diagnostic reporting could make a national examination more useful for learning without changing its role in certification.

Abbreviations

AUC
Area under the receiver operating characteristic curve
BIC
Bayesian information criterion
CA
Classification accuracy
CDM
Cognitive diagnostic model
CI
Confidence interval

DINA
Deterministic inputs, noisy “and” gate
DTR
Diagnostic-test reasoning
ENADE
Exame Nacional de Desempenho dos Estudantes (National Student Performance Examination)
ENAMED
Exame Nacional de Avaliação da Formação Médica (National Examination for the Assessment of Medical Training)
EUM
Emergency and urgent care management
G-DINA
Generalised deterministic inputs, noisy “and” gate
INEP
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KR-20
Kuder–Richardson formula 20
OTP
Outpatient therapeutic planning
PHA
Population-health action
PVAF
Proportion of variance accounted for
$RMSEA_2$
Root mean square error of approximation based on M_2
ROC
Receiver operating characteristic
SDV
Social determinants of health and vulnerability
SR
Syndromic recognition
SRH
Sexual and reproductive health
SRMSR
Standardised root mean square residual

Declarations

Ethics approval and consent to participate

This study analysed publicly available, de-identified secondary data released by INEP. Under Resolution No. 510/2016 of the Brazilian National Health Council (Article 1, sole paragraph, items II and III), research that uses publicly accessible or public-domain information is not registered or evaluated by the CEP/CONEP research ethics system.

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the INEP open data repository (Microdados do Enamed 2025), https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enamed [4]. The Q-matrices, detailed model results, analysis code and aligned response matrix supporting the conclusions of this article are included within the article's additional files (Additional files 1–4).

Competing interests

CFC is an Editorial Board member of BMC Medical Education and had no involvement in the editorial handling or peer review of this manuscript. The other authors declare that they have no competing interests.

Funding

ELMC is a scholarship recipient of the Araucária Foundation's Institutional Program for Scientific Initiation Scholarships.

Authors' contributions

LS coded and revised the care-setting Q-matrix, contributed to interpreting the care-setting results, supervised the coding of the other axes, conducted psychometric analyses, and wrote the first draft of the manuscript. ALOD reviewed the clinical-task Q-matrix against the item content, conducted psychometric analyses, and contributed to the interpretation of the task results. GGL coded and revised the content-area Q-matrix, conducted psychometric analyses, and contributed to the interpretation of content-area results. ELMC coded and revised the patient-feature descriptors, performed psychometric analyses, and contributed to the interpretation of patient-feature results. RQSP reviewed the coding of all Q-matrices, contributed to the interpretation of results, and revised the manuscript. CFC conceived and designed the study, developed the clinical-task Q-matrix, supervised the coding of the other axes, conducted psychometric analyses, and wrote revised versions of the manuscript. All authors critically revised the manuscript and approved the final version.

Acknowledgements

We thank INEP for making the ENAMED microdata publicly available.

References

1. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, et al. The impact of large scale licensing examinations in highly developed countries: a systematic review. BMC Med Educ. 2016;16:212. https://doi.org/10.1186/s12909-016-0729-7.
2. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85. https://doi.org/10.3109/0142159X.2011.565828.
3. Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Portaria nº 413, de 18 de junho de 2025. Regulamenta o Exame Nacional de Avaliação da Formação Médica – Enamed. Diário Oficial

da União. 2025; Seção 1:28.

4. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Manual do usuário: microdados do Enamed, edição 2025. Brasília: INEP. 2025. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enamed. Accessed 16 Oct 2026.
5. Brasil. Ministério da Educação. Conselho Nacional de Educação. Câmara de Educação Superior. Resolução CNE/CES nº 3, de 20 de junho de 2014. Institui Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina. Diário Oficial da União. 2014; Seção 1:8–11.
6. Haberman SJ. When can subscores have value? J Educ Behav Stat. 2008;33(2):204–29. https://doi.org/10.3102/1076998607302636.
7. Sinharay S, Puhan G, Haberman SJ. An NCME instructional module on subscores. Educ Meas Issues Pract. 2011;30(3):29–40. https://doi.org/10.1111/j.1745-3992.2011.00208.x.
8. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309–. https://doi.org/10.1111/j.1365-2929.2005.02094.x. 17.
9. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206 – 14. https://doi.org/10.3109/0142159X.2011.551559
10. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50(1):1–73. https://doi.org/10.1111/jedm.12000.
11. Rupp AA, Templin J, Henson RA. Diagnostic measurement: theory, methods, and applications. New York: Guilford Press; 2010.
12. Ravand H, Baghaei P. Diagnostic classification models: recent developments, practical issues, and prospects. Int J Test. 2020;20(1):24–56. https://doi.org/10.1080/15305058.2019.1588278.
13. Collares CF. Cognitive diagnostic modelling in healthcare professions education: an eye-opener. Adv Health Sci Educ Theory Pract. 2022;27(2):427–40. https://doi.org/10.1007/s10459-022-10093-y.
14. Choi Y, Seo DG. Estimation of item parameters and examinees' mastery probability in each domain of the Korean Medical Licensing Examination using a deterministic inputs, noisy and gate (DINA) model. J Educ Eval Health Prof. 2020;17:35. https://doi.org/10.3352/jeehp.2020.17.35.
15. Matus AR, Matus LN, Hiltz A, Chen T, Kaur B, Brewster P, et al. Development of an assessment technique for basic science retention using the NBME subject exam data. BMC Med Educ. 2022;22:771. https://doi.org/10.1186/s12909-022-03842-5.
16. García PE, Olea J, de la Torre J. Application of cognitive diagnosis models to competency-based situational judgment tests. Psicothema. 2014;26(3):372–7. https://doi.org/10.7334/psicothema2013.322.
17. Rupp AA, Templin J. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. Educ Psychol Meas. 2008;68(1):78–96. https://doi.org/10.1177/0013164407301545.
18. Sessoms J, Henson RA. Applications of diagnostic classification models: a literature review and critical commentary. Measurement: Interdisciplinary Res Perspect. 2018;16(1):1–17. https://doi.org/10.1080/15366367.2018.1435104.
19. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge (MA): Harvard University Press; 1978.

20. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1(1):41–67. https://doi.org/10.1007/BF00596229.
21. Eva KW. On the generality of specificity. Med Educ. 2003;37(7):587–8. https://doi.org/10.1046/j.1365-2923.2003.01563.x.
22. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? Med Educ. 2006;40(7):618–23. https://doi.org/10.1111/j.1365-2929.2006.02511.x.
23. Shavelson RJ, Baxter GP, Gao X. Sampling variability of performance assessments. J Educ Meas. 1993;30(3):215–32. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x.
24. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educ Res. 1995;24(5):5–11. https://doi.org/10.3102/0013189X024005005.
25. Brown JS, Collins A, Duguid P. Situated cognition and the culture of learning. Educ Res. 1989;18(1):32–42. https://doi.org/10.3102/0013189X018001032.
26. Lave J, Wenger E. Situated learning: legitimate peripheral participation. Cambridge: Cambridge University Press; 1991. https://doi.org/10.1017/CBO9780511815355.
27. Durning SJ, Artino AR. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide 52. Med Teach. 2011;33(3):188–99. https://doi.org/10.3109/0142159X.2011.550965.
28. Merkebu J, Battistone M, McMains K, McOwen K, Witkop C, Konopasky A, et al. Situativity: a family of social cognitive theories for understanding clinical reasoning and diagnostic error. Diagnosis (Berl). 2020;7(3):169–76. https://doi.org/10.1515/dx-2019-0100.
29. Durning SJ, Artino AR, Pangaro LN, van der Vleuten C, Schuwirth L. Perspective: redefining context in the clinical encounter: implications for research and training in medical education. Acad Med. 2010;85(5):894–901. https://doi.org/10.1097/ACM.0b013e3181d7427c.
30. Durning S, Artino AR, Pangaro L, van der Vleuten CPM, Schuwirth L. Context and clinical reasoning: understanding the perspective of the expert's voice. Med Educ. 2011;45(9):927–38. https://doi.org/10.1111/j.1365-2923.2011.04053.x.
31. Konopasky A, Artino AR, Battista A, Ohmer M, Hemmer PA, Torre D, et al. Understanding context specificity: the effect of contextual factors on clinical reasoning. Diagnosis (Berl). 2020;7(3):257–64. https://doi.org/10.1515/dx-2020-0016.
32. Lingard L. What we see and don't see when we look at 'competence': notes on a god term. Adv Health Sci Educ Theory Pract. 2009;14(5):625–8. https://doi.org/10.1007/s10459-009-9206-y.
33. de la Torre J. The generalized DINA model framework. Psychometrika. 2011;76(2):179–. https://doi.org/10.1007/s11336-011-9207-7. 99.
34. Ma W, de la Torre J. GDINA: an R package for cognitive diagnosis modeling. J Stat Softw. 2020;93(14):1–26. https://doi.org/10.18637/jss.v093.i14.
35. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2024. https://www.R-project.org/. Accessed 16 Oct 2026.
36. Maydeu-Olivares A, Joe H. Assessing approximate fit in categorical data analysis. Multivar Behav Res. 2014;49(4):305–28. https://doi.org/10.1080/00273171.2014.911075.
37. Hu J, Miller MD, Huggins-Manley AC, Chen YH. Evaluation of model fit in cognitive diagnosis models. Int J Test. 2016;16(2):119–41. https://doi.org/10.1080/15305058.2015.1133627.

38. Chen J, de la Torre J, Zhang Z. Relative and absolute fit evaluation in cognitive diagnosis modeling. J Educ Meas. 2013;50(2):123–40. https://doi.org/10.1111/j.1745-3984.2012.00185.x.
39. Iaconangelo C. Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models [dissertation]. New Brunswick (NJ): Rutgers University; 2017. https://doi.org/10.7282/T3W95D95
40. Wang W, Song L, Chen P, Meng Y, Ding S. Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. J Educ Meas. 2015;52(4):457–76. https://doi.org/10.1111/jedm.12096.
41. de la Torre J, Chiu CY. A general method of empirical Q-matrix validation. Psychometrika. 2016;81(2):253–. https://doi.org/10.1007/s11336-015-9467-8. 73.
42. Nájera P, Sorrel MA, de la Torre J, Abad FJ. Balancing fit and parsimony to improve Q-matrix validation. Br J Math Stat Psychol. 2021;74(Suppl 1):110–30. https://doi.org/10.1111/bmsp.12228.
43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
44. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45. https://doi.org/10.2307/2531595.
45. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3.
46. Brasil. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016. https://www.gov.br/conselho-nacional-de-saude/pt-br/acesso-a-informacao/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/@@download/file. Accessed 16 Oct 2026.
47. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ. 2005;39(12):1176–7. https://doi.org/10.1111/j.1365-2929.2005.02341.x.
48. Norman G. Research in clinical reasoning: past history and current trends. Med Educ. 2005;39(4):418–. https://doi.org/10.1111/j.1365-2929.2005.02127.x. 27.
49. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: achieving equity in assessment for medical learners. Acad Med. 2020;95(12S):S98–108. https://doi.org/10.1097/ACM.0000000000003717.

Figures

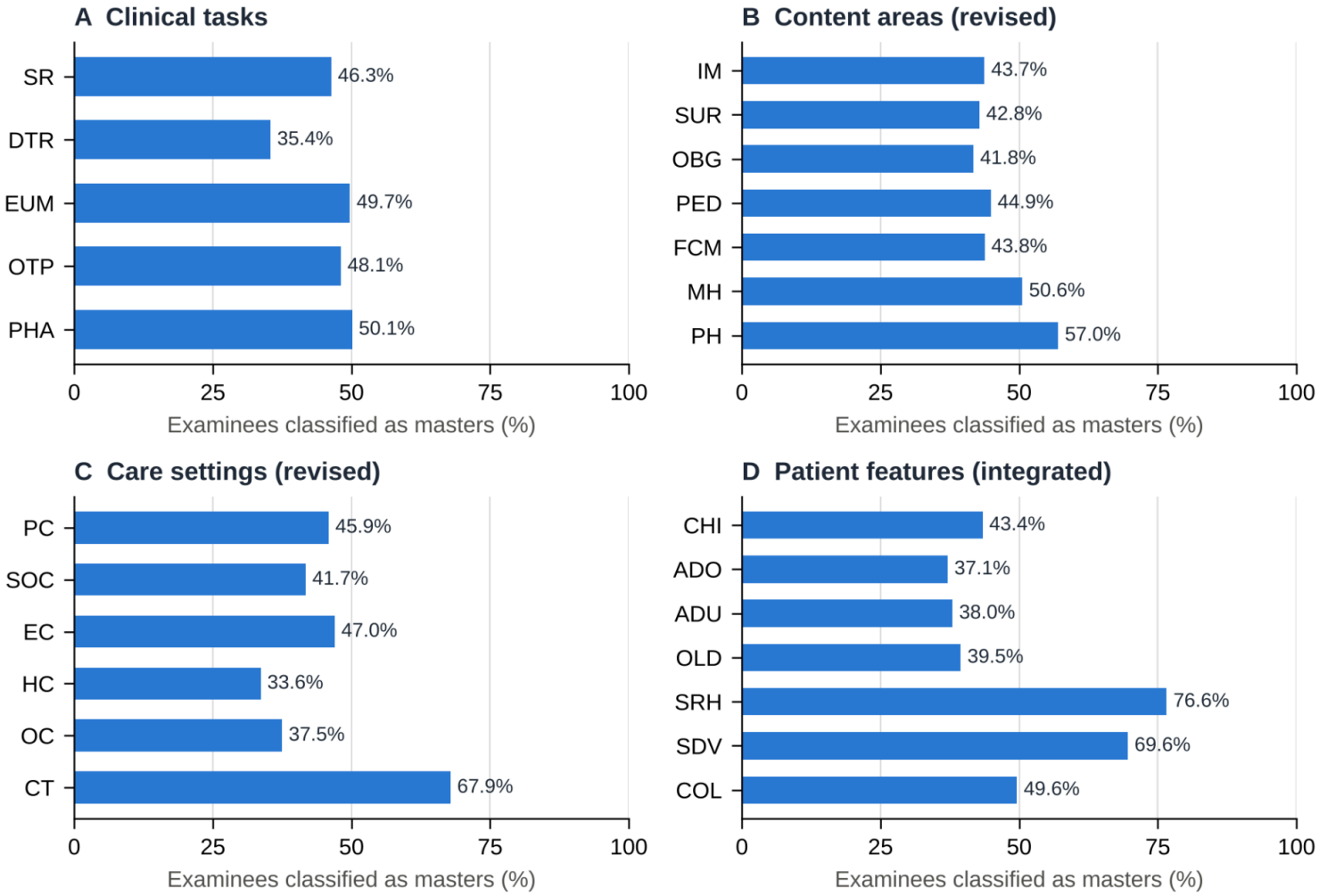


Figure 1

Attribute mastery prevalence under the selected specification of each axis

Percentage of the 39,256 examinees classified as masters (marginal posterior probability ≥ 0.5) of each attribute under (A) the clinical-task Q-matrix, (B) the revised content-area Q-matrix, (C) the revised care-setting Q-matrix, and (D) the integrated patient-feature Q-matrix, all fitted with the saturated G-DINA model. Attribute abbreviations as in Table 1.

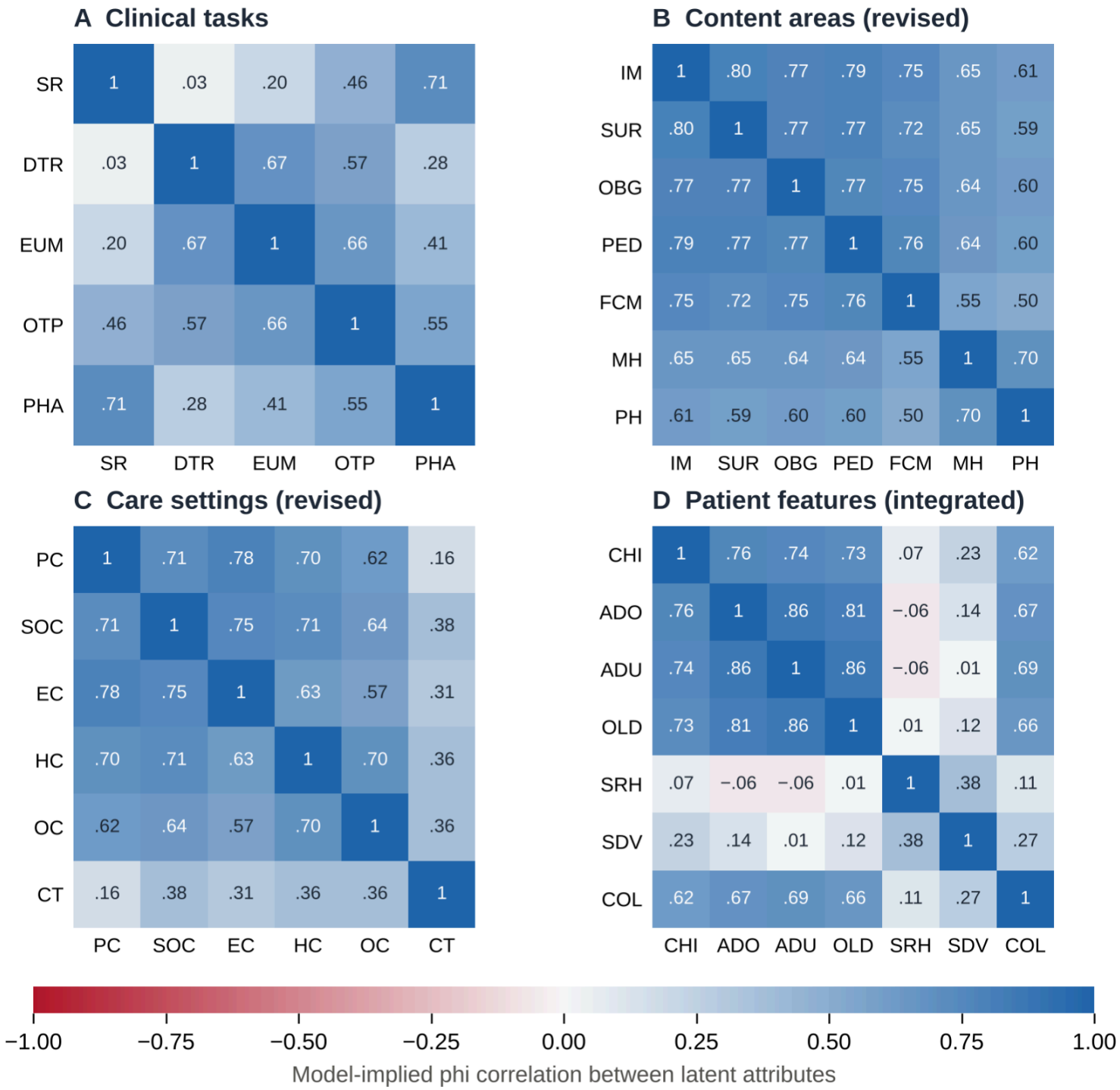


Figure 2

Latent correlations between attributes under the selected specification of each axis

Phi coefficients computed from the estimated joint distribution of the binary attributes for (A) clinical tasks, (B) content areas (revised), (C) care settings (revised) and (D) patient features (integrated). Blue indicates a positive and red a negative association. Attribute abbreviations as in Table 1.

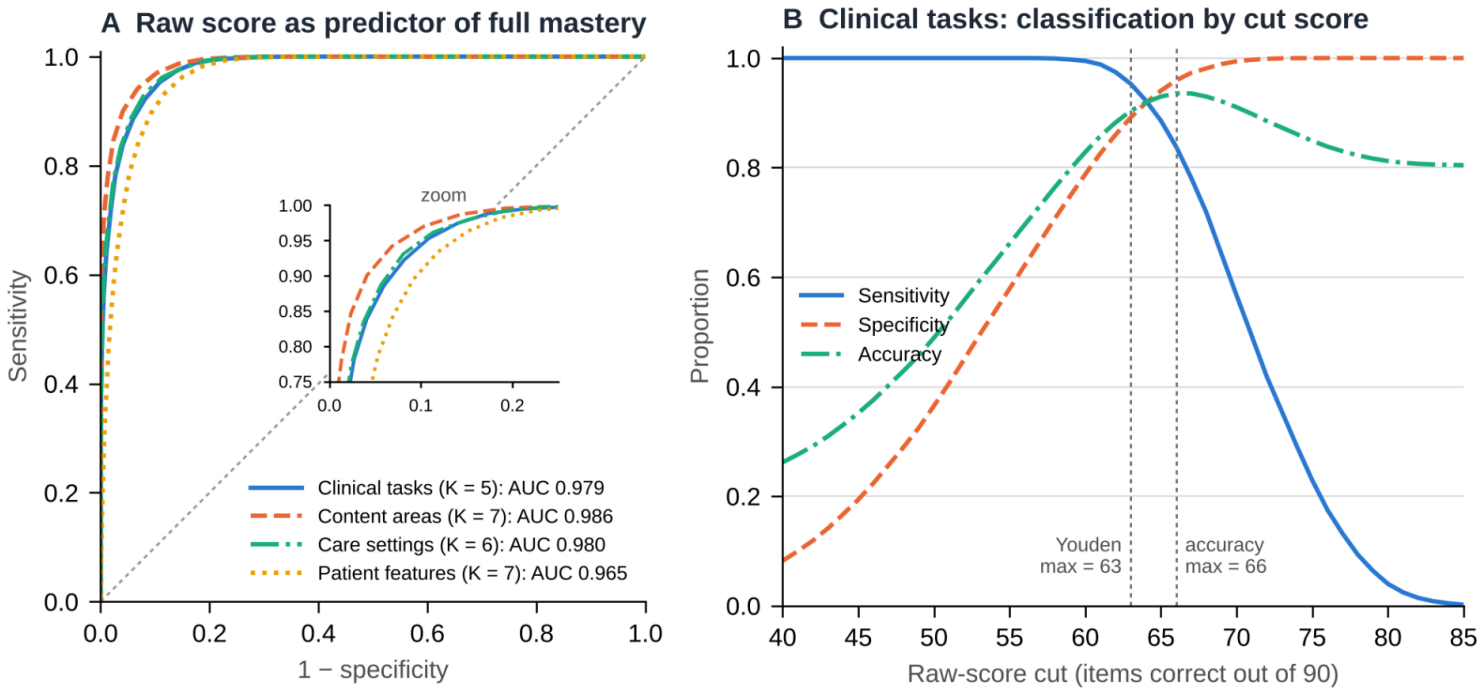


Figure 3

Raw score as predictor of full attribute mastery: ROC curves and cut-score classification

(A) Receiver operating characteristic curves of the raw score (0–90) for identifying examinees classified as masters of all attributes, under the selected specification of each axis; the inset enlarges the upper-left region. Areas under the curve (AUC) are shown in the legend. (B) Sensitivity, specificity and accuracy of each raw-score cut for identifying full mastery of the five clinical tasks; dashed lines mark the cuts that maximise Youden's index (63) and accuracy (66).

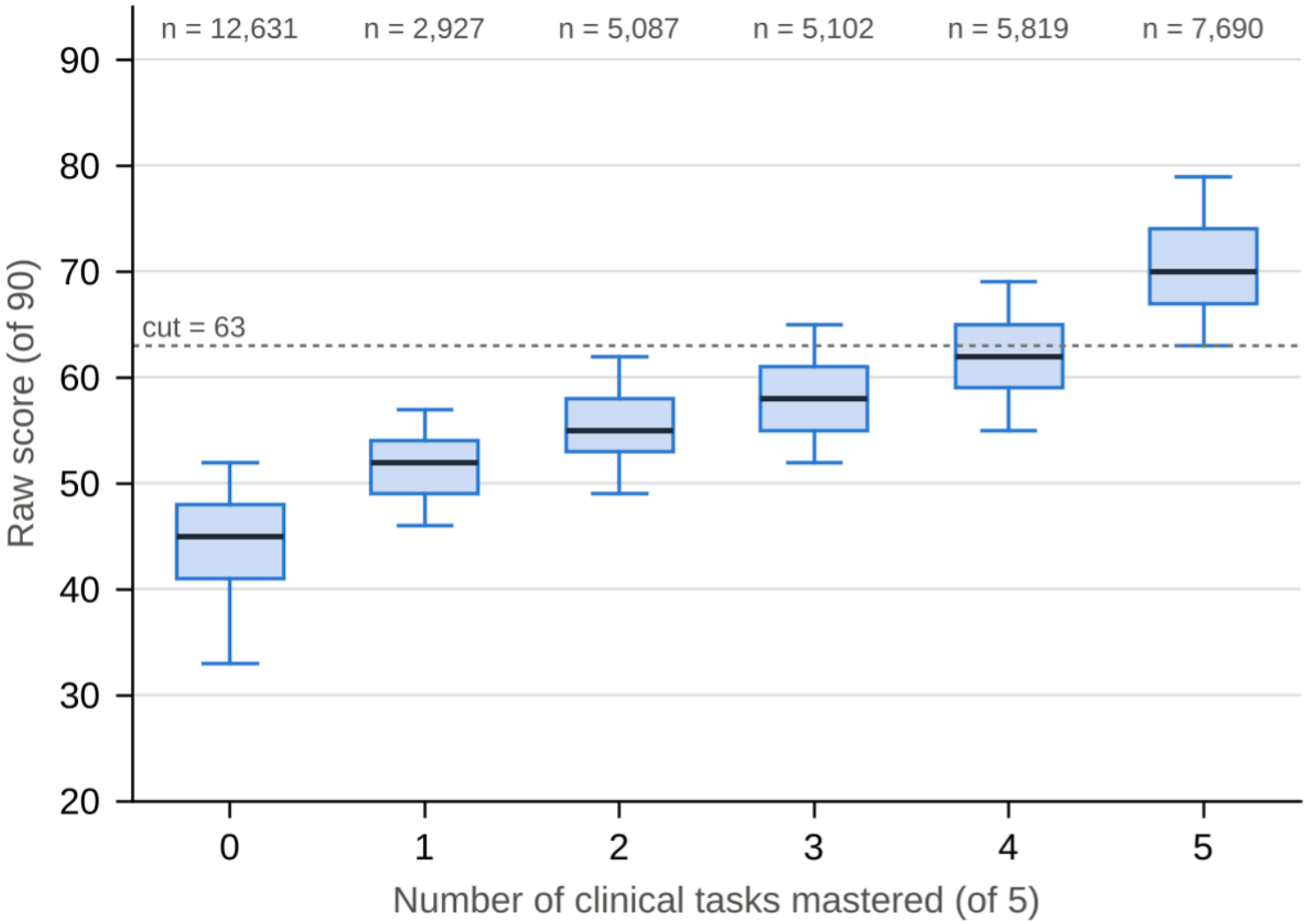


Figure 4

Raw-score distribution by number of clinical tasks mastered

Boxes show the interquartile range and median, and whiskers the 5th and 95th percentiles, of the raw score (0–90) for examinees mastering 0 to 5 of the clinical tasks. The dashed line marks the Youden-optimal cut for full task mastery (63).

Supplementary Files

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

- Additionalfile2Modelresults.xlsx
- Additionalfile3gdinaanalysis.r
- Additionalfile1Qmatrices.xlsx
- Additionalfile4alignedresponses.csv