

Applying a Cognitive Diagnostic Model to Assess Clinical Reasoning: Q- Matrix Development and Known-Groups Validity Evidence from a Nationwide Medical Examination

Seung-Joo Na

s.jna620@gachon.ac.kr

Gachon University College of Medicine

Research Article

Keywords: clinical reasoning, cognitive diagnostic model, Q-matrix, DINA model, multiple-choice questions, known-groups validity

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10412698/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

Cognitive diagnostic models (CDMs) can provide multidimensional feedback on examinees' mastery of specific cognitive attributes, but their application requires a valid Q-matrix specifying which attributes each item measures. Few studies in medical education have systematically developed and validated a Q-matrix for clinical reasoning. We developed a Q-matrix mapping cognitive attributes of clinical reasoning to 144 internal medicine multiple-choice items from a nationwide Clinical Comprehensive Examination (CCE) administered to 5,586 third- and fourth-year Korean medical students. Nine attributes derived from a literature review were mapped to items, evaluated using pairwise Jaccard coefficients and multiple regression, and refined by a four-member expert panel into seven attributes. We then fit the deterministic inputs, noisy 'and' gate (DINA) model, which showed good absolute fit (mean item-level RMSEA = 0.044), to estimate examinee-level attribute mastery and compared mastery between third- and fourth-year students as a known-groups validity check. The refined Q-matrix showed acceptable inter-attribute distinctiveness (Jaccard coefficients ≤ 0.32) and non-significant item-level regression coefficients, treated as supplementary rather than confirmatory evidence. In contrast, DINA-based mastery differed significantly between grades for all seven attributes, with uniformly large effect sizes (fourth-year: 0.83–0.91; third-year: 0.50–0.65; Cohen's $d = -1.05$ to -1.19 , all $p < .001$), consistent with the theoretically expected developmental gap as students progress through clinical clerkships toward graduation. These known-groups differences provide stronger evidence of validity than the item-level statistics alone. Findings support a parsimonious, seven-attribute framework for clinical reasoning that can inform prospective item-writing blueprints and future learner-level diagnostic feedback in medical education.

Introduction

Clinical reasoning is a core competency in medical education and is essential for safe, effective patient care (Norman, 2005; Eva, 2005; Bowen, 2006), involving the interpretation of clinical information, the generation and refinement of diagnostic hypotheses, and the selection of management strategies. Because it directly supports physicians' decision-making, the validity of its assessment is a central concern in undergraduate medical education (Bowen, 2006; Schuwirth et al., 2020; Kassirer, 2010).

Assessing clinical reasoning remains challenging because it is multidimensional, context-dependent, and often implicit (Rencic et al., 2016). Workplace-based and simulation-based assessments provide rich information about students' reasoning but are resource-intensive and have limited sampling (Daniel et al., 2019; Fürstenberg et al., 2020). In contrast, well-designed clinical vignette-style multiple-choice questions (MCQs) can efficiently assess reasoning across a broad range of clinical contexts (Fürstenberg et al., 2020; Modi et al., 2015), eliciting reasoning strategies and errors when carefully constructed (Heist et al., 2014; Surry et al., 2018).

Conventional MCQ scoring provides only total or item-level scores, offering limited information about the specific cognitive components of clinical reasoning required by each item or demonstrated by each learner. Collares (2022) has argued that the literature confirms scores alone are not effective for the

prevention or timely detection of critical learning gaps, and that overcoming the assumption of unidimensionality in favor of a more qualitative use of modern psychometrics—basing decisions on richer, attribute-level information rather than a single number—could better align assessment with competency-based curricula and programmatic assessment. Cognitive diagnostic models (CDMs) operationalize this shift by classifying learners according to mastery of multiple discrete cognitive attributes (Tatsuoka, 1983; Rupp et al., 2010). A valid Q-matrix—the incidence matrix specifying which cognitive attributes are required to answer each item correctly—is a prerequisite for applying any CDM. Because Q-matrix quality directly constrains the interpretability and classification accuracy of subsequent diagnostic modeling (Rupp & Templin, 2008), it should be developed through an explicit process integrating theoretical grounding, expert judgment, and empirical evidence, and its validity should ultimately be supported by evidence beyond item-level statistics alone.

Relatively few studies in medical education have systematically developed and evaluated the cognitive attributes of clinical reasoning for MCQ-based examinations, and even fewer have carried this work through to learner-level diagnostic modeling. In this study, we (a) developed and refined a Q-matrix for clinical reasoning attributes measured by a nationwide internal medicine MCQ examination, using a targeted literature review, item-level empirical indices, and structured expert consensus; and (b) fit the deterministic inputs, noisy ‘and’ gate (DINA) model to estimate examinee-level attribute mastery and tested whether mastery probabilities differed between third- and fourth-year students in the theoretically expected direction. Because clinical reasoning is expected to develop over the course of clinical clerkships, a significant, theoretically consistent difference in DINA-based mastery between grades would provide known-groups validity evidence for the Q-matrix—evidence that item-level indices alone cannot supply. Figure 1 summarizes the overall study procedure, distinguishing the supplementary item-level evidence (Phases 1–3) from the core learner-level validity evidence (Phase 4).

Methods

Ethics statement

This study involved a secondary analysis of previously collected, fully de-identified examination data and did not involve direct interaction with participants. The Institutional Review Board of CHA University reviewed the protocol for this secondary data analysis and determined that the study was exempt from full review (IRB No. 1044308-202507-HR-273-01). The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable institutional guidelines. Further institutional oversight details are provided in the Declarations section.

Setting, participants, and data

We used data from the 2019 second Clinical Comprehensive Examination (CCE), a nationwide written examination administered by the Korea Association of Medical Colleges, consisting of 360 MCQs across

major clinical disciplines. We analyzed the internal medicine section (144 MCQs). The dataset included response data from 5,586 third- (n = 2,603) and fourth-year (n = 2,983) medical students from all 40 medical schools in Korea who took the 2019 second CCE. All eligible responses were included; because the data were fully anonymized, no demographic variables were available.

Attribute development and Q-matrix construction

Preliminary cognitive attributes were developed through a targeted literature review of conceptual and empirical studies on clinical and diagnostic reasoning published between 2000 and 2022. Three frameworks (Simmons, 2010; Young et al., 2018; Thampy et al., 2019) informed an initial nine-attribute structure, which the researcher and an expert responsible for CCE item management mapped to the 144 internal medicine items to construct a draft Q-matrix.

Empirical evaluation of the draft Q-matrix

For two attributes A and B, each represented as the set of items requiring that attribute, the Jaccard coefficient is defined as $J(A,B) = |A \cap B| / |A \cup B|$, ranging from 0 (no shared items) to 1 (identical item sets); coefficients below 0.50 were interpreted as indicating acceptable distinctiveness (Kim et al., 2012; Hennig, 2007), and multiple regression analysis with the logit-transformed item correct-response rate as the dependent variable and dichotomous attribute indicators as independent variables, consistent with prior Q-matrix validation studies using this approach (Kim et al., 2012). Consistent with prior Q-matrix construction approaches that combine content-expert judgment with item-level statistical results as reference information during consensus review (Chen & de la Torre, 2014; Sawaki et al., 2009), these results were treated as supplementary reference for expert judgment rather than as standalone validation criteria.

Expert panel refinement

An expert consensus review, rather than an exploratory qualitative study, was conducted to refine the preliminary attributes and Q-matrix. The panel consisted of four medical school faculty members responsible for CCE item development and management, with specialties in internal medicine, family medicine, and pediatrics, and 6 to over 20 years of medical education experience. The draft Q-matrix and its Jaccard and regression results were provided to the panel as reference information. Experts reviewed each attribute for conceptual clarity, overlap, feasibility of identification, and educational interpretability, and final decisions on retention, merging, revision, or removal were made by consensus. The revised Q-matrix was then re-evaluated using the same Jaccard and regression procedures.

Cognitive diagnostic modeling and known-groups comparison

To assess whether the refined Q-matrix could support meaningful learner-level inference—evidence that item-level indices alone cannot provide—we fit the deterministic inputs, noisy ‘and’ gate (DINA) model (Haertel, 1989; de la Torre, 2009) to the item response data using the refined seven-attribute Q-matrix. The DINA model is a conjunctive, non-compensatory CDM widely used with large-scale assessment data (Kim & Park, 2013) and was considered appropriate because mastering all attributes required by an item was judged necessary for a correct response, consistent with the expert panel’s judgment during Q-matrix refinement.

For examinee j and item i , the DINA model specifies the probability of a correct response as a function of two item parameters—the slip parameter (s_i), the probability of an incorrect response despite mastering all required attributes, and the guess parameter (g_i), the probability of a correct response despite not mastering all required attributes—and a deterministic latent-response variable (η_{ij}) indicating whether examinee j has mastered every attribute required by item i :

$$P(X_{ij} = 1 \mid \alpha_j) = g_i^{(1-\eta_{ij})} \times (1 - s_i)^{\eta_{ij}}, \text{ where } \eta_{ij} = \prod_k \alpha_{jk}^{q_{ik}}$$

Here α_{jk} indicates whether examinee j has mastered attribute k (1 = mastered, 0 = not mastered), and q_{ik} is the corresponding entry of the Q-matrix (1 if item i requires attribute k , 0 otherwise); η_{ij} equals 1 only if examinee j has mastered every attribute required by item i . Examinee-level attribute mastery probabilities (posterior probabilities of mastery) were estimated for all 5,586 examinees. Because this study did not compare competing CDMs, model evaluation focused on absolute fit rather than relative fit indices; the item-level root mean square error of approximation (RMSEA) was the primary fit statistic, with values ≤ 0.05 interpreted as indicating excellent absolute fit and ≤ 0.08 as acceptable. Akaike and Bayesian information criteria (AIC, BIC) are also reported descriptively. The DINA model was estimated using the CDM package (version 8.2-6; Robitzsch et al., 2022) in R.

Because clinical reasoning is expected to develop over the course of clinical clerkships—with fourth-year students possessing more clinical experience and, in most participating schools, stronger incentives for engagement because CCE performance is incorporated into course grades and the examination immediately precedes the national licensing examination—we compared DINA-based mastery probabilities between third- and fourth-year examinees for each of the seven attributes using independent-samples t -tests. A statistically significant difference in the theoretically expected direction (fourth-year > third-year) for all attributes would constitute known-groups validity evidence for the Q-matrix (Cronbach & Meehl, 1955). As a complementary, more readily interpretable analysis, examinees with a posterior mastery probability greater than 0.50 were classified as having mastered the attribute, and the percentage classified as having mastered each attribute was compared between grades using chi-square tests.

Results

As a baseline for interpreting subsequent attribute-level analyses, Table 1 presents descriptive statistics for total internal medicine CCE scores by grade. Because Levene’s test indicated a significant violation of the equal-variances assumption ($F(1,5584) = 213, p < .001$), a Welch’s t-test was used. Fourth-year examinees scored substantially higher than third-year examinees on the total score ($M = 102.0$ vs. 82.8 , out of 144 items; Welch’s $t(4,843) = -47.3, p < .001$, Cohen’s $d = -1.28$), confirming that a known grade-based difference existed at the level of overall exam performance and providing a benchmark against which to compare the attribute-specific DINA results below.

Table 1
Descriptive statistics of total internal medicine CCE scores, by grade

Grade	n	M	SD	Max	Min	Welch's t	df	p	Cohen's d
3rd-year	2,603	82.8	16.7	130	33	-47.3	4,843	< .001	-1.28
4th-year	2,983	102.0	12.8	133	51				
Total	5,586	92.9	17.5	133	33				

Q-matrix development and refinement

Nine preliminary attributes were mapped to the 144 internal medicine items to construct the draft Q-matrix. Pairwise Jaccard coefficients ranged from 0.0000 to 0.3200, all below the prespecified threshold of 0.50. In the regression analysis ($R^2=0.047$, adjusted $R^2=-0.018$; $F(9,134) = 0.73, p=.685$), no attribute coefficient was statistically significant.

The expert panel reviewed the nine preliminary attributes alongside these empirical results and recommended merging patient history interpretation and physical examination interpretation (as these were often integrated when examinees identified clinically significant patient information), removing significant-information screening and evaluation (as written MCQ stems generally present already-selected clinical information), and removing the emergency-level component of severity assessment (as it was difficult to infer consistently in written examinations). Table 2 summarizes the panel’s decisions and rationale for each preliminary attribute.

Table 2
Expert panel review results and refinement decisions

Preliminary attribute	Decision	Rationale
p1. Patient history interpretation P2. Physical exam interpretation	Merged → a1	Often integrated in practice when examinees identify clinically significant patient information.
p3. Interpretation of laboratory findings	Retained → a2	Conceptually clear and consistently identifiable in item stems.
p4. Significant information screening and evaluation	Removed	Written MCQ stems generally present already-selected clinical information, making it difficult to distinguish this independently.
p5. Cause of disease analysis	Retained → a3	Conceptually clear and educationally interpretable.
p6. Additional medical testing evaluation	Retained → a4	Conceptually clear and educationally interpretable.
p7 Diagnostic reasoning	Retained → a5	Conceptually clear and educationally interpretable.
p8. Severity assessment	Retained (modified) → a6	Emergency-level component removed, as it was difficult to infer consistently in written examinations.
P9. Treatment decision-making	Retained → a7	Conceptually clear and educationally interpretable.

This process yielded seven final attributes: distinguishing significant patient information (a1), interpretation of laboratory findings (a2), analysis of causes of disease (a3), assessment of need for additional tests (a4), diagnostic reasoning (a5), severity assessment (a6), and decision-making for treatment and management (a7). Definitions are provided in Table 3, and the number of items measuring each attribute is provided in Table 4.

Table 3
Final seven cognitive attributes of clinical reasoning and definitions

Process	Attribute	Definition
Patient information interpretation	a1 Distinguishing significant patient information	Able to distinguish information related to the disease from medical history and physical examination findings.
	a2 Interpretation of laboratory findings	Able to correctly interpret laboratory findings.
Differential diagnosis	a3 Analysis of causes of disease	Able to explain the causes and mechanisms of disease occurrence.
	a4 Assessment of need for additional tests	Able to suggest additional tests necessary for diagnosis.
Diagnosis	a5 Diagnostic reasoning	Able to interpret, analyze, and synthesize patient information to derive a diagnosis.
Treatment and management	a6 Severity assessment	Able to assess severity based on patient information and diagnostic reasoning.
	a7 Decision-making for treatment and management	Able to plan appropriate medical or surgical treatment based on patient information and diagnostic reasoning.

Table 4
Number of items measuring each cognitive attribute

Attribute	Number of items (of 144)
a1 Distinguishing significant patient information	101
a2 Interpretation of laboratory findings	109
a3 Analysis of causes of disease	11
a4 Assessment of need for additional tests	30
a5 Diagnostic reasoning	46
a6 Severity assessment	8
a7 Decision-making for treatment and management	64
Note. Because attributes can co-occur within an item, column totals may exceed 144.	

Most items required multiple attributes simultaneously: 52.1% of items (n = 75) required three attributes, 43.8% (n = 63) required two, and only 2.8% (n = 4) and 1.4% (n = 2) required four or one attribute, respectively (Table 5), consistent with the conjunctive, multi-attribute nature of clinical reasoning items that the DINA model is designed to accommodate.

Table 5
Distribution of the number of attributes required per item

Number of attributes per item	Number of items	% of items
1	2	1.4%
2	63	43.8%
3	75	52.1%
4	4	2.8%
Total	144	100.0%

Re-evaluation of the refined Q-matrix showed comparable, non-improved item-level indices—Jaccard coefficients from 0.0000 to 0.3238, and no significant regression coefficients ($R^2=0.025$, adjusted $R^2=-0.025$; $F(7,136) = 0.51$, $p=.828$)—consistent with our treatment of these indices as supplementary reference rather than confirmatory validation criteria (see Methods).

DINA-based attribute mastery and known-groups comparison

Model fit was evaluated before interpreting mastery estimates. Because this study did not compare competing CDMs, we focused on absolute fit: item-level RMSEA ranged from 0.004 to 0.078 across the 144 items ($M = 0.044$), with every item at or below the 0.08 threshold for acceptable fit and the mean below the 0.05 threshold for excellent fit, indicating good absolute fit of the DINA model to the observed data ($AIC = 788,050$; $BIC = 790,801$).

Item parameter estimates further characterized model behavior at the item level. Across the 144 items, correct response rates ranged from 0.033 to 0.996 (58 items, 40.3%, ≥ 0.75 ; 79 items, 54.9%, $0.25 < 0.75$; 7 items, 4.9%, < 0.25). The guess parameter (probability of a correct response despite not mastering all required attributes) ranged from 0.006 to 0.993 ($M = 0.520$), and the slip parameter (probability of an incorrect response despite mastering all required attributes) ranged from 0.001 to 0.962 ($M = 0.267$). We additionally examined the item discrimination index ($IDI = 1 - s - g$), which should be positive under the DINA model's assumption that mastering all required attributes increases the probability of a correct response beyond that of guessing. IDI was negative for 7 of 144 items (range: -0.094 to -0.003), indicating that, for these items, this assumption was not empirically supported; we did not remove these items post hoc, but note that they may warrant expert content review in future refinement of the Q-matrix (see Limitations).

Across the full sample ($n = 5,586$), DINA-based attribute mastery probabilities ranged from 0.674 (a4, assessment of need for additional tests) to 0.770 (a3, analysis of causes of disease). Table 6 presents overall and separate mastery probabilities for third- and fourth-year examinees. Fourth-year students showed substantially and significantly higher mastery than third-year students on every attribute (fourth-

year M range 0.826–0.910; third-year M range 0.500–0.646; all $t > 30$, all $p < .001$), with effect sizes uniformly large (Cohen’s d ranging from -1.05 to -1.19). This pattern was consistent across all seven attributes and in the theoretically expected direction, providing known-groups validity evidence for the refined Q-matrix that complements—and is considerably stronger than—the non-significant item-level regression results reported above. As a complementary, more readily interpretable summary, Fig. 2 presents the percentage of examinees in each grade classified as having mastered each attribute (posterior mastery probability > 0.50); this dichotomous classification, compared using chi-square tests, showed the same pattern (χ^2 range 237.69–1261.48, $df = 1$, all $p < .001$).

Table 6
DINA-based attribute mastery probability, overall and by grade

Attribute	Overall mastery probability	3rd-year M (SD) (n = 2,603)	4th-year M (SD) (n = 2,983)	t	p	Effect size (d)
a1 Distinguishing significant patient information	0.765	0.598 (0.313)	0.910 (0.207)	-43.34	< .001	-1.19
a2 Interpretation of laboratory findings	0.721	0.539 (0.381)	0.892 (0.251)	-41.80	< .001	-1.15
a3 Analysis of causes of disease	0.770	0.646 (0.217)	0.880 (0.202)	-41.82	< .001	-1.13
a4 Assessment of need for additional tests	0.674	0.500 (0.335)	0.826 (0.296)	-30.02	< .001	-1.05
a5 Diagnostic reasoning	0.743	0.581 (0.308)	0.866 (0.237)	-40.93	< .001	-1.12
a6 Severity assessment	0.750	0.607 (0.288)	0.867 (0.218)	-38.74	< .001	-1.05
a7 Decision-making for treatment and management	0.697	0.504 (0.361)	0.866 (0.279)	-41.38	< .001	-1.13
Note. Mastery probabilities are DINA posterior probabilities of mastery. t-tests are independent-samples comparisons between third- and fourth-year examinees; negative t-values and effect sizes indicate higher mastery among fourth-year students. Effect sizes are Cohen’s d; $ d > 0.8$ is conventionally considered a large effect.						

Discussion

This study developed and refined a seven-attribute Q-matrix for clinical reasoning in MCQ-based written examinations and evaluated it using both item-level indices and, distinctively, learner-level DINA-based known-groups comparisons. The item-level Jaccard and regression indices were non-significant and showed no improvement after expert-driven refinement, which we interpret as expected given their

limited statistical power at the item level (144 items) and their role as supplementary reference information rather than confirmatory validation criteria (Chen & de la Torre, 2014; Sawaki et al., 2009). In marked contrast, DINA-based attribute mastery differed significantly between third- and fourth-year examinees on every attribute, in the theoretically expected direction and with very large effect sizes. This known-groups pattern is a form of validity evidence that item-level regression, by construction, cannot provide, because it depends on examinee-level classification rather than aggregate item difficulty.

The magnitude and consistency of the grade difference is notable: fourth-year students, who have completed core clinical clerkships and are approaching the national licensing examination, showed uniformly high mastery (0.83–0.91) across all seven attributes, whereas third-year students, earlier in clinical training, showed markedly lower and more variable mastery (0.50–0.65). This gradient is consistent with the theoretical expectation that clinical reasoning develops progressively through clinical exposure and is not fully established at the third-year level, and it mirrors the incentive structure of the examination itself: fourth-year CCE performance is incorporated into course grades at most participating schools and immediately precedes the national licensing examination, whereas third-year administration and stakes vary by institution. Notably, the attribute-level effect sizes ($d = -1.05$ to -1.19) were comparable in magnitude to, though somewhat smaller than, the effect size for the total internal medicine score itself ($d = 1.28$; Table 1). This suggests that DINA-based attribute mastery did not simply recapitulate the total-score gap with added noise; rather, it decomposed a known, large aggregate difference into seven attribute-specific estimates of comparable magnitude, offering the more granular, attribute-level information that a single total score cannot provide. As a further robustness check, we also compared mastery probabilities across achievement-level subgroups (top, middle, and bottom 27%, defined by a three-parameter IRT ability estimate, consistent with the classification convention used by the Korea Health Personnel Licensing Examination Institute [2020] for item discrimination analysis). Top-achieving examinees showed consistently and substantially higher mastery than bottom-achieving examinees across all seven attributes in both grades (third-year top–bottom differences ranging 0.41–0.84; fourth-year 0.33–0.57), converging with the grade-based results above. A full analysis of achievement-level mastery patterns, including several attributes for which fourth-year top- and middle-achieving examinees both approached ceiling levels, is beyond the scope of this study and is reported in a companion paper examining attribute mastery profiles and their practical utilization.

These findings extend prior work suggesting that well-constructed clinical vignette-style MCQs can assess clinical reasoning and reveal reasoning-related errors (Heist et al., 2014; Surry et al., 2018) by demonstrating that an explicit Q-matrix linking clinical reasoning attributes to MCQ items can support learner-level diagnostic inference, not merely item-level description. Collares (2022) argued conceptually that cognitive diagnostic modeling can overcome the limitations of unidimensional scores—which the literature suggests are not effective for the timely detection of critical learning gaps—by supporting decisions based on richer, attribute-level information rather than a single number, and that this shift could better align assessment with competency-based curricula and programmatic assessment. The present study provides empirical, known-groups evidence consistent with that argument: attribute mastery probabilities were not only estimable at the individual level but also behaved exactly as clinical

reasoning theory would predict across training years, a pattern that a single total score could not have revealed with comparable specificity. This is also consistent with prior Q-matrix construction studies in which item-level statistical indices served as reference information for expert judgment rather than as an independent validation standard, while the ultimate validity claim rested on evidence closer to the intended use of the instrument (Chen & de la Torre, 2014; Sawaki et al., 2009). Insofar as attribute-level mastery information may eventually inform decisions with patient-safety implications—such as identifying trainees who have not yet mastered specific clinical reasoning components before advancing to higher-stakes clinical responsibility—the kind of validity evidence demonstrated here, grounded in a theoretically expected known-groups contrast, is a necessary step before such uses could be responsibly considered.

Limitations

This study has limitations. First, the Q-matrix was developed retrospectively from items not originally designed to measure predefined attributes—a retrofitting problem documented in cognitive diagnostic assessment (Gierl & Cui, 2008)—so attribute distributions and combinations across items were not systematically controlled; consequently, some attributes were measured by relatively few items (a3, $n = 11$; a6, $n = 8$; Table 4), and mastery estimates for these attributes warrant cautious interpretation, as estimates based on fewer items may be less stable. We note, however, that these two attributes did not appear as systematic outliers in mastery probability relative to attributes measured by more items (Table 6). Second, the initial Q-matrix relied on expert judgment rather than think-aloud protocols, given the impracticality of cognitive interviews with an already-administered nationwide examination; this is consistent with established procedures for diagnostically modeling extant large-scale assessment data (Chen & de la Torre, 2014; Sawaki et al., 2009), but future prospective item development could incorporate cognitive interviews. Third, attribute refinement was based on group consensus discussion among four experts rather than independently elicited ratings, so quantitative inter-rater agreement indices could not be calculated; future studies could use structured techniques such as the Delphi method. Fourth, although the grade-based known-groups comparison provides strong evidence of validity, grade is a proxy for clinical reasoning development rather than a direct external criterion; future work should examine convergent validity against other clinical reasoning measures (e.g., objective structured clinical examination performance) where available. Fifth, item discrimination indices were negative for 7 of 144 items, indicating that the DINA model's assumption of higher correct-response probability for examinees mastering all required attributes was not empirically supported for these items; because we retained the full item set to preserve consistency with the Q-matrix used throughout, mastery estimates should be interpreted with this caveat in mind, and future refinement could involve targeted expert review or exclusion of these items. Finally, the analysis was limited to internal medicine items from a single nationwide Korean examination, which may limit its generalizability to other disciplines and contexts.

Conclusion

We developed and refined a seven-attribute Q-matrix for clinical reasoning in MCQ-based written examinations and found that, although item-level Jaccard and regression indices were limited and non-significant, DINA-based attribute mastery probabilities differed markedly and significantly between third- and fourth-year examinees on every attribute, in the theoretically expected direction. This known-groups pattern provides validity evidence for the Q-matrix that complements and substantially strengthens the item-level indices alone, and demonstrates that the framework can support learner-level diagnostic inference. Future research should examine whether this Q-matrix, applied prospectively in item development, can generate interpretable mastery profiles that support individualized feedback, curricular improvement, and standard-setting in medical education.

Abbreviations

AIC Akaike information criterion

BIC Bayesian information criterion

CCE Clinical Comprehensive Examination

CDM Cognitive diagnostic model

DINA Deterministic inputs, noisy 'and' gate (model)

MCQ Multiple-choice question

RMSEA Root mean square error of approximation

Declarations

Ethical approval.

This study was approved for exemption from full review by the Institutional Review Board of CHA University (IRB No. 1044308-202507-HR-273-01). This study involved a secondary analysis of previously collected, fully de-identified examination data and did not involve direct interaction with participants. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable institutional guidelines.

Consent to Participate.

Not applicable. This study conducted a secondary analysis of retrospective, de-identified examination data provided by the testing institution after the examination. As the study was granted an exemption from review by the Institutional Review Board of CHA University, the requirement for informed consent was waived.

Competing interests.

The author declares no competing interests.

Funding.

None

Author Contribution

S.J.N conceived and designed the study, developed the Q-matrix, conducted the statistical analyses, and drafted and revised the manuscript. S.J.N. is the sole author and meets the criteria for authorship as defined by the International Committee of Medical Journal Editors (ICMJE) guidelines.

Acknowledgements.

None.

Data Availability

Individual-level respondent data are not publicly available, as they are owned by the examination-administering institution, which did not grant permission for public release; given that response patterns across 144 items could pose re-identification risk even if de-identified, these data may be available from the corresponding author upon reasonable request and with permission from the data-providing institution. Item-level correct response rates (aggregated across respondents) and the Q-matrix will be made available as data files.

References

1. Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21), 2217–2225. <https://doi.org/10.1056/NEJMr054782>
2. Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the Programme for International Student Assessment in reading. *Psychology*, 5(18), 1967–1978. <https://doi.org/10.4236/psych.2014.518200>
3. Collares, C. F. (2022). Cognitive diagnostic modelling in healthcare professions education: An eye-opener. *Advances in Health Sciences Education*, 27, 427–440. <https://doi.org/10.1007/s10459-022-10093-y>
4. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.

5. Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., et al. (2019). Clinical reasoning assessment methods: A scoping review and practical guidance. *Academic Medicine*, 94(6), 902–912. <https://doi.org/10.1097/ACM.0000000000002618>
6. de la Torre, J. (2009). DINA model parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115–130.
7. Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical Education*, 39(1), 98–106. <https://doi.org/10.1111/j.1365-2929.2004.01972.x>
8. Fürstenberg, S., Helm, T., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2020). Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, 20, 368. <https://doi.org/10.1186/s12909-020-02260-9>
9. Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 263–268. <https://doi.org/10.1080/15366360802497762>
10. Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4), 301–321.
11. Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258–271. <https://doi.org/10.1016/j.csda.2006.11.025>
12. Heist, B. S., Gonzalo, J. D., Durning, S., Torre, D., & Elnicki, D. M. (2014). Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: A mixed methods study. *Journal of Graduate Medical Education*, 6, 709–714. <https://doi.org/10.4300/JGME-D-14-00176.1>
13. Kassirer, J. P. (2010). Teaching clinical reasoning: Case-based and coached. *Academic Medicine*, 85(7), 1118–1124. <https://doi.org/10.1097/acm.0b013e3181d5dd0d>
14. Kim, H. K., Han, J. A., Choi, S. K., & Kim, B. (2012). *The application of cognitive diagnostic model for achievement profile analysis and score report (RRE 2012-7)*. Korea Institute for Curriculum and Evaluation.
15. Kim, H. K., & Park, C. H. (2013). An exploration of the appropriate cognitive diagnostic model for NAEA (National Assessment of Educational Achievement): A comparison between Fusion and DINA models. *Journal of Curriculum and Evaluation*, 16(2), 127–148.
16. Korea Health Personnel Licensing Examination Institute. (2020). *2020 nyeondo je 84hoe uisa gukga siheom (pilgi) munhang bunseok gyeolgwa [2020 84th Korean Medical Licensing Examination (written) item analysis results]*. Author.
17. Modi, J. N., Anshu, Gupta, P., & Singh, T. (2015). Teaching and assessing clinical reasoning skills. *Indian Pediatrics*, 52, 787–794. <https://doi.org/10.1007/s13312-015-0718-7>
18. Norman, G. (2005). Research in clinical reasoning: History and current trends. *Medical Education*, 39(4), 418–427. <https://doi.org/10.1111/j.1365-2929.2005.02127.x>

19. Rencic, J., Durning, S. J., Holmboe, E., & Gruppen, L. D. (2016). Understanding the assessment of clinical reasoning. In P. Wimmers & M. Mentkowski (Eds.), *Assessing competence in professional performance across disciplines and professions* (pp. 209–235). Springer.
https://doi.org/10.1007/978-3-319-30064-1_11
20. Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2022). CDM: Cognitive diagnosis modeling (R package version 8.2-6) [Computer software]. <https://CRAN.R-project.org/package=CDM>
21. Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, 68, 78–96.
22. Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
23. Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6(3), 190–209. <https://doi.org/10.1080/15434300902801917>
24. Schuwirth, L., Durning, S., & King, S. (2020). Assessment of clinical reasoning: Three evolutions of thought. *Diagnosis*, 7(3), 191–196. <https://doi.org/10.1515/dx-2019-0096>
25. Simmons, B. (2010). Clinical reasoning: Concept analysis. *Journal of Advanced Nursing*, 66(5), 1151–1158. <https://doi.org/10.1111/j.1365-2648.2010.05262.x>
26. Surry, L. T., Torre, D., Trowbridge, R. L., & Durning, S. J. (2018). A mixed-methods exploration of cognitive dispositions to respond and clinical reasoning errors with multiple-choice questions. *BMC Medical Education*, 18, 277. <https://doi.org/10.1186/s12909-018-1372-2>
27. Tatsuoaka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345–354.
28. Thampy, H., Willert, E., & Ramani, S. (2019). Assessing clinical reasoning: Targeting the higher levels of the pyramid. *Journal of General Internal Medicine*, 34(8), 1631–1636.
<https://doi.org/10.1007/s11606-019-04953-4>
29. Young, M. E., Thomas, A., Lubarsky, S., Ballard, T., Gordon, D., Gruppen, L. D., et al. (2018). Drawing boundaries: The difficulty in defining clinical reasoning. *Academic Medicine*, 93(7), 990–995.
<https://doi.org/10.1097/ACM.0000000000002142>

Figures

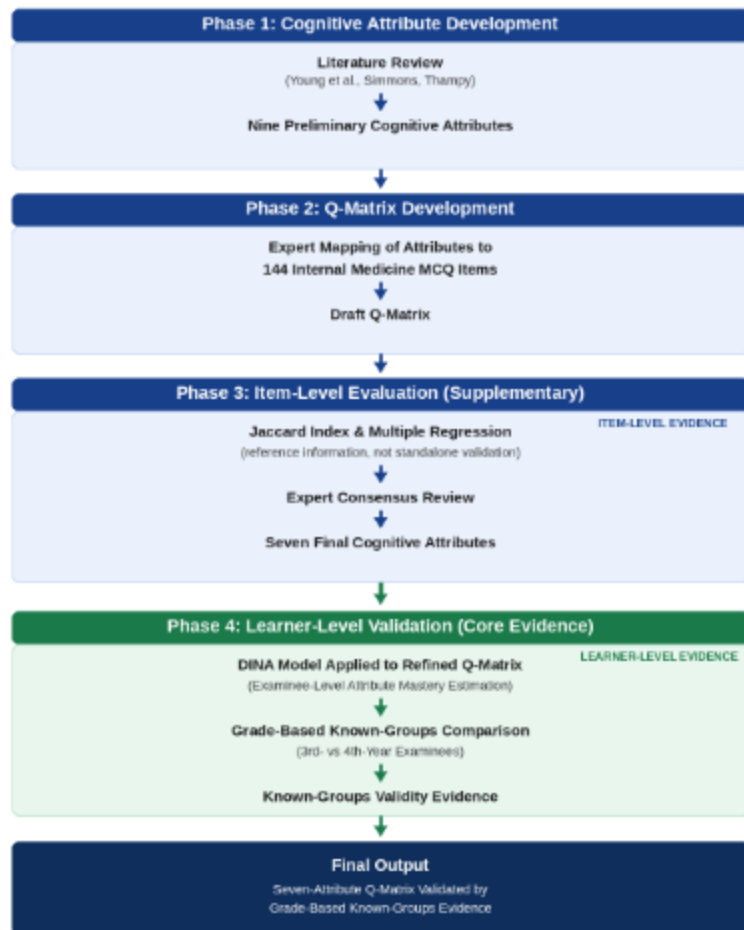


Figure 1

Overall study procedure, distinguishing supplementary item-level evidence (Phases 1–3) from core learner-level known-groups validity evidence (Phase 4).

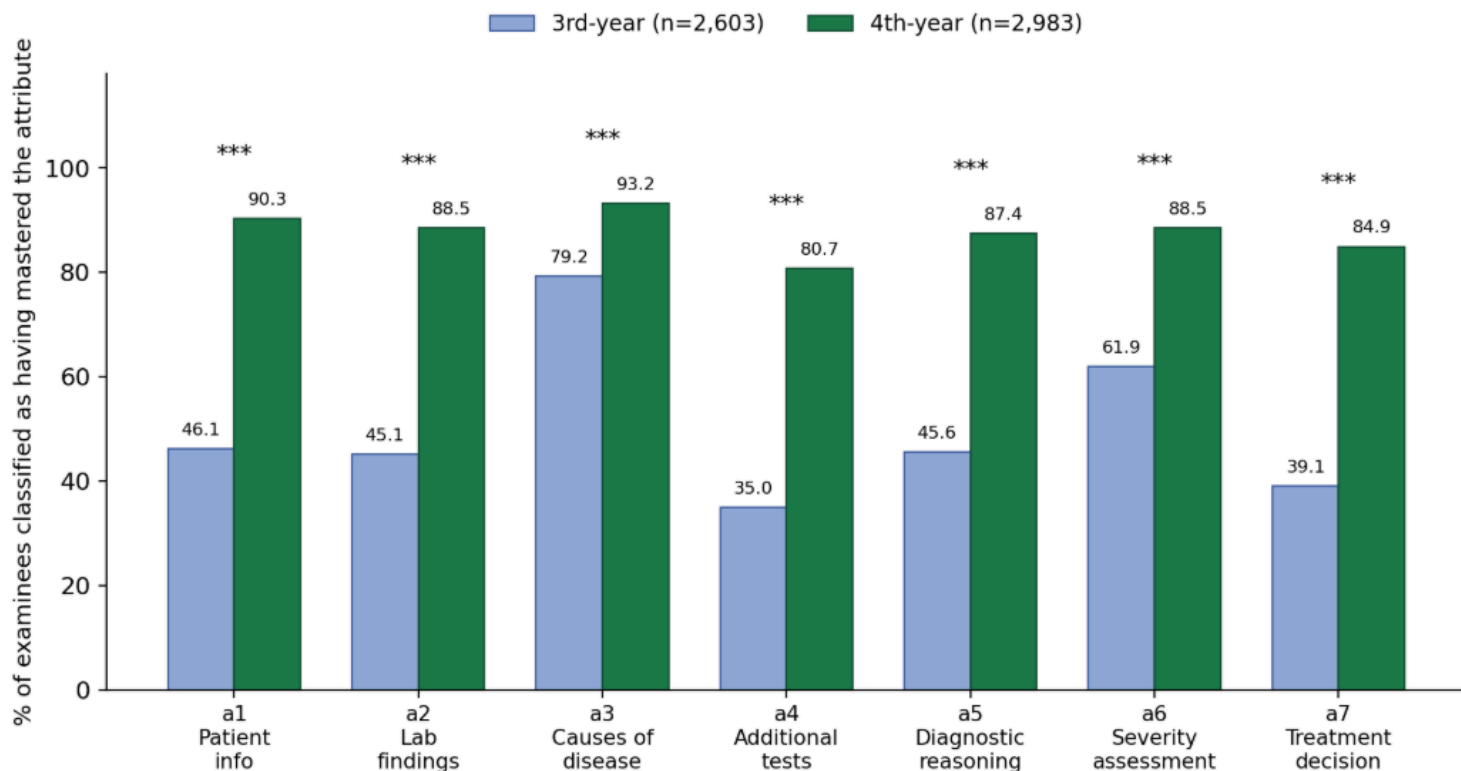


Figure 2

Percentage of examinees classified as having mastered each attribute (posterior mastery probability >0.50), by grade. *** χ^2 test $p < .001$ for the third- vs. fourth-year comparison for every attribute (χ^2 range 237.69–1261.48, $df=1$).