

I Know How I Know: Mapping the Declarative Topography of Confidence in Higher-Order Assessment

Colin John Greengrass

cgreengrass@rcsi.com

Royal College of Surgeons in Ireland - Bahrain

Kianna Thelen

University of Iowa Health Care

Research Article

Keywords: Declarative Domain Metacognitive Instrument, confidence judgement, metacognitive calibration, dual-process theory, feeling-of-knowing, response-process validity, clinical reasoning, time-on-task

Posted Date: June 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9868995/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

Assessment outcomes coupled with confidence ratings can identify confidence–accuracy calibration, but they do not explain whether confidence is grounded in explicit evidence, reasoned justification, or intuitive judgement. This distinction is important in health-professions and higher-order assessment contexts, where learners must monitor the reliability, completeness and evidential basis of their own knowledge. This paper examines validity evidence for the Declarative Domain Metacognitive Instrument (DDMI), a four-category instrument where learners report their metacognitive articulation state: a guess with no statable rationale (DDMI 1); felt sense of the answer with no articulable basis (DDMI 2); partial reasoning stated but remaining incomplete (DDMI 3); or an explicit reason, rule, or mechanism the learner can state (DDMI 4). The DDMI is grounded in Koriatic's distinction between experience-based and information-based monitoring and dual-process metacognitive theory. The instrument was embedded in pharmacology assessments in one nursing cohort (76 students, 4,258 responses), DDMI categories showed an ordered relationship with accuracy, rising from 26.4% to 53.9% ($\gamma = 0.268$); the lowest category performed at chance, the next decisively above it. The gradient persisted after adjustment for item difficulty ($OR = 1.403$, 95% CI 1.310–1.502), within individual students and under clustering, and was not reducible to conventional confidence magnitude. Reported articulation was fallible: even DDMI 4 responses were inaccurate almost half the time, identifying cases in which learners reported an explicit basis for knowing but selected an incorrect response. The DDMI offers formative value for identifying misplaced certainty in health professions and higher-order assessment; response-process validation is the immediate next step.

Introduction

In assessment, responses receiving the same score are often interpreted as equivalent for grading, progression, or summary reporting. Yet an accurate response may be produced because a learner understands the underlying principle, because a response appears familiar, because partial knowledge has been used successfully, or because the learner has guessed. Similarly, an inaccurate response may be accompanied by appropriate uncertainty, or by strong confidence in a flawed but internally coherent line of reasoning. These states carry different educational implications and should therefore inform different interpretations and feedback priorities.

This argument has a clear basis in validity theory and in the formative assessment tradition. In validity theory, scores are not self-interpreting; their meaning depends on the inferences and uses attached to them (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1995), so that a score may be adequate for ranking or certification yet insufficient for diagnostic interpretation. Formative assessment theory makes a complementary point: feedback is most useful when it helps learners understand the nature of the gap between current and desired performance and how that gap might be closed (Black & Wiliam, 1998; Hattie & Timperley, 2007; Sadler, 1989). A raw accuracy score records whether a response met a criterion; it does not, on its own, indicate whether the learner guessed, recognised a pattern, reasoned partially, or explicitly constructed a flawed justification.

This distinction is especially important in higher-order assessment, where the educational aim extends beyond recognition of information to the use of knowledge in interpretation, problem-solving, justification, and transfer to unfamiliar situations. Conventional scoring records whether a response is accurate, and confidence scales add useful information by asking learners how certain they are. However, confidence magnitude – how certain a learner feels – is not the same as the basis of confidence – whether the learner believes they can identify why they gave that response. A conventional confidence rating may assign the same value to qualitatively different epistemic states: where one may be able to explain the response from first principles, while another may only recognise the response as familiar.

The present paper distinguishes confidence magnitude from perceived metacognitive articulation: the learner's item-level judgement of how far the basis for confidence appears statable, explainable, or justifiable at the point of answering.

Four converging bodies of theory and research

The following theoretical frameworks approach this issue from complementary perspectives and together provide the conceptual foundation for the construct examined in this study.

Metacognitive monitoring and calibration

Metacognition refers to knowledge, monitoring, and regulation of one's own cognitive activity (Flavell, 1979; Nelson & Narens, 1990). In assessment contexts, confidence judgements function as item-level metacognitive monitoring: they ask learners to evaluate the status of their response at the point of commitment (Fleming & Lau, 2014; Schraw, 2009). This item-level emphasis is consistent with evidence from health professions education that learners may show useful insight when reflecting during or immediately after performance, even where broader global self-assessment is limited (Eva & Regehr, 2005; Tweed et al., 2017). Recent work using MCQ response certainty has shown that such response-level judgements can be used to define and track aspects of medical student self-monitoring across training (Tweed et al., 2020; Tweed et al., 2024). Calibration – the alignment between subjective confidence and objective accuracy – has long been the primary metric through which the educational value of confidence judgements is assessed (Lichtenstein & Fischhoff, 1977). Poor calibration takes two distinct forms with different educational implications. Overconfidence in inaccurate responses may indicate misconception, premature closure, misplaced fluency, or failure to recognise the limits of understanding (Berner & Graber, 2008; Curtis et al., 2013). Underconfidence in accurate responses may indicate fragile knowledge, incomplete consolidation, or limited trust in otherwise sound reasoning (Eva & Regehr, 2005).

Metacognitive monitoring is particularly relevant in higher-order problem-solving contexts because monitoring operates at the interface of intuitive and deliberate reasoning. An initial response may be generated rapidly and intuitively, after which metacognitive processes evaluate whether the response

feels sufficiently plausible to retain or whether further analysis is warranted (Ackerman & Thompson, 2017; Thompson et al., 2011). Marcum's integrated model of clinical reasoning situates metacognitive monitoring as the mechanism that regulates transitions between non-analytic and analytic reasoning in professional judgement (Marcum, 2012). The distinction addressed in the present study – whether the learner can articulate the basis of confidence, rather than simply how strong that confidence feels – arises at this juncture.

Metamemory and the phenomenology of knowing

A distinct but closely related body of evidence concerns the phenomenological basis of knowing. Feeling-of-knowing research demonstrates that individuals may experience a strong sense that they know something even when they cannot retrieve or articulate the relevant knowledge (Hart, 1965; Koriat, 1993). Koriat's accessibility model is particularly pertinent: feeling-of-knowing judgements arise not from full access to target knowledge but from the accessibility of partial or related information (Koriat, 1993). Confidence is therefore not a direct readout of knowledge but an inferential judgement constructed from cues such as fluency, familiarity, coherence, retrieval ease, and partial evidence (Dinsmore & Parkinson, 2013; Koriat, 2012).

Critically, Koriat distinguishes between experience-based monitoring, where confidence derives from phenomenological cues such as processing fluency and familiarity, and information-based monitoring, where confidence derives from reportable content – a specific mechanism, reasoning chain, or evidence the learner can identify and state (Koriat, 1997). This distinction is central to reported metacognitive articulation. A conventional confidence scale records the magnitude of certainty, but does not distinguish confidence grounded primarily in familiarity or fluency from confidence supported by reportable content.

A further development of Koriat's account bears directly on the present study. The self-consistency model proposes that subjective confidence reflects the consistency of the cues, considerations, or representations a respondent samples when forming a judgement, rather than their objective correctness (Koriat, 2012). On this account, confidence is high when sampled considerations agree with one another, regardless of whether they are valid. This predicts, rather than merely accommodates, the existence of confident error: a learner who retrieves several mutually consistent but collectively mistaken considerations may experience strong and apparently justifiable confidence in an inaccurate response. Any instrument designed to classify the basis of confidence should therefore be expected to identify not only well-grounded justification but also internally coherent yet mistaken reasoning.

The remember/know/guess tradition provides the closest conceptual antecedent for capturing the phenomenological basis of a response rather than its confidence strength.

Remember/know paradigms ask participants to distinguish recollective experience from familiarity-based knowing, with extended formats adding a guess category (Gardiner, 1988; Tulving, 1985).

Although developed in recognition-memory research rather than higher-education assessment, these paradigms demonstrate that individuals can classify the experiential basis of a response, and that such classifications can carry information beyond confidence magnitude alone.

Thompson's feeling-of-rightness framework further proposes that individuals form an initial intuitive answer and then monitor whether that answer feels sufficiently correct to retain (Ackerman & Thompson, 2017; Thompson, 2009; Thompson et al., 2011). For higher-order problem-solving questions, students may experience a response as plausible before they can fully justify that response: they may have a sense that the response is right without yet having access to the reasoning that would support it. This phenomenological state – felt knowing without articulable grounding – is theoretically distinct from both guessing and explicitly justified confidence.

Dual-process reasoning and perceived processing basis

Dual-process theories distinguish fast, automatic, intuitive processing from slower, deliberate, effortful processing (Evans & Stanovich, 2013; Kahneman, 2003). This distinction is often expressed in terms of heuristic and analytic processing. Thompson's metacognitive account of dual-process reasoning is particularly relevant because it locates monitoring at the point where an initial intuitive response is evaluated for plausibility and possible analytic engagement (Ackerman & Thompson, 2017; Thompson, 2009; Thompson et al., 2011). This provides a theoretical bridge between the phenomenology of knowing and assessable self-report: learners may be able to report whether a response feels plausible but inarticulate, partially explainable, or explicitly justified.

Stanovich's tripartite account further distinguishes heuristic, algorithmic, and reflective processes (Stanovich, 2009, 2011). In this account, the algorithmic mind refers to the computational processes that sustain explicit reasoning, while the reflective mind regulates goals, beliefs, and the decision to engage such reasoning. The present paper uses *algorithmic* in this dual-process sense: not as an educational label for a taught procedure, but as a reference to explicit, rule-governed, or stepwise reasoning that can in principle be declared by the learner. This does not imply that intuition is inferior. Expert intuitive judgement may be highly accurate when grounded in well-organised knowledge, and analytic reasoning may fail when knowledge is incomplete or poorly structured (Norman et al., 2017).

Declarative and non-declarative knowledge

Cognitive neuroscience distinguishes between knowledge explicitly accessible to consciousness and reportable, and knowledge that guides performance without being available to conscious report (Cohen & Squire, 1980; Reber, 1989; Squire, 2004). Reber's work on implicit learning demonstrated that participants could acquire complex knowledge through exposure and experience the resulting judgements as a sense of rightness or familiarity, without being able to articulate what they had learned (Reber, 1989). This phenomenological state – performance guided by knowledge that cannot be stated –

is theoretically distinct from explicitly articulable understanding. Knowledge that can be articulated can be shared, examined, corrected, extended, and transferred. Knowledge guiding performance without explicit reportability may function reliably in familiar situations, but may be less available for explanation, correction, extension, or transfer when circumstances change.

The measurement gap and the DDMI

These four bodies of evidence converge on a measurement gap. The gap is not whether confidence can be recorded, but whether learners can report what their confidence is based on. Existing approaches address related but distinct assessment problems. Conventional confidence scales and MCQ certainty measures capture certainty magnitude and may provide useful information about learner self-monitoring (Curtis et al., 2013; Tweed et al., 2020; Tweed et al., 2024). However, on a five-point scale, the extremes are usually interpretable as low and high confidence, while movement between adjacent middle categories is often weakly specified: a shift from 3 to 4 may indicate greater perceived certainty, but the scale does not define what has changed functionally in the learner's judgement. This ambiguity becomes more pronounced as scales become more finely divided. Certainty-based and confidence-based marking go further by incorporating confidence into the score itself: learners receive greater credit for accurate responses endorsed with high confidence and may receive reduced credit or penalty when high confidence is attached to an inaccurate response (Caleon & Subramaniam, 2009; Curtis et al., 2013; Echternacht, 1972; Gardner-Medwin, 1995). These approaches are valuable for encouraging calibration, but their primary function is scoring or calibration feedback rather than classification of the experienced basis of confidence.

Multi-tier diagnostic tests come closest to this aim because they move beyond correctness alone by asking learners to provide or select a reason, explanation, or confidence judgement alongside their response, often to identify misconceptions or distinguish correct reasoning from correct guessing (Caleon & Subramaniam, 2009; Hasan et al., 1999). However, these instruments usually treat reasoning and confidence as separate response elements. They do not provide a single ordered item-level judgement through which learners classify how the basis of confidence is experienced: as guessed, felt but inarticulate, partially explainable, or explicitly justified.

The **Declarative Domain Metacognitive Instrument (DDMI)** was developed to address this gap by operationalising reported metacognitive articulation as an item-level judgement. The term *declarative* refers to the extent to which certainty is experienced as reportable, explainable, or justifiable. The word *domain* refers to the response space in which the learner locates the reported basis of confidence: a domain of declaration.

The DDMI comprises four defined metacognitive articulation states. DDMI 1 denotes guessing: the response is experienced as a complete or near-complete guess and no rationale or reliable cue can be stated. DDMI 2 denotes a felt sense of knowing: the response feels familiar, plausible, or likely to be correct, but the learner cannot state an articulable reason for it. This category aligns with experience-

based monitoring, feeling-of-rightness, and non-declarative knowledge. DDMI 3 denotes mixed partial articulation: the learner reports partial explicit reasoning that can be partially stated, but the account remains incomplete and is supplemented by heuristic, implicit, or familiarity-based cues; some information-based monitoring is available but remains partial. DDMI 4 denotes explicit perceived justification: the learner reports confidence as supported by a largely explicit analytic or algorithmic rationale, rule, mechanism, or robust recognised heuristic whose basis can be stated; the monitoring is predominantly information-based and the reasoning is believed to be declarable.

Furthermore, the DDMI can be understood as a construct-mapped assessment instrument. A construct map defines ordered positions in the level or form of a target construct and allows response categories to be interpreted within a theoretically specified progression (Wilson, 2005). In the DDMI, the construct is reported metacognitive articulation: the four categories are not arbitrary confidence labels but defined articulation states representing increasing reported articulability of confidence, from complete or near-complete guessing through explicit perceived justification. This construct-map logic explains why the DDMI is treated as ordinal and why the validity argument focuses on ordered category functioning rather than internal consistency. Cronbach's alpha is not treated as primary validity evidence here, not because alpha could not be computed, but because alpha addresses a different measurement question: the internal consistency of a multi-item trait measure. The DDMI is an item-level judgement applied repeatedly across questions that differ in topic, difficulty, and timing; apparent item-level covariance would reflect those differences as readily as the construct itself.

The DDMI also draws on epistemic cognition – the study of how individuals understand knowledge, justification, and the grounds on which claims are held to be known (Chinn et al., 2011; Hofer & Pintrich, 1997). When answering a question, learners implicitly adopt an epistemic stance toward their own response. The DDMI asks them to make that stance explicit: is the response guessed, familiarly known without articulable basis, partially justifiable, or explicitly grounded in a rationale that can be stated? The DDMI can therefore be understood as an item-level epistemic-metacognitive judgement that combines confidence monitoring with the reported articulability of the justificatory basis of confidence. To the author's knowledge, no previously described instrument has attempted to operationalise this progression within a single ordered item-level response scale in higher-education assessment.

The present study therefore examines exploratory validity evidence for the DDMI by testing whether the four articulation states used by learners, show ordered relations with accuracy, remain distinct from confidence magnitude, and identify educationally meaningful accuracy–articulation states. The study addressed six research questions. (1) Do learners use all four DDMI categories? (2) Do categories show an ordered relationship with item-level accuracy? (3) Does this relationship persist after adjustment for item difficulty and session? (4) Are DDMI ratings sensitive to item difficulty? (5) Does the DDMI relate to but remain distinct from a conventional percentage confidence scale? (6) Can DDMI responses classify accuracy–articulation states not visible from correctness alone?

Methods

Study design and cohort selection

The present study used a single-cohort observational design to examine exploratory validity evidence for the DDMI. It was framed as an assessment-validity investigation rather than as an evaluation of a teaching method. The empirical setting was undergraduate nursing pharmacology. The construct under examination, reported metacognitive articulation, is directly relevant to health professions education, where higher-order assessment routinely requires learners not merely to select correct responses but to justify clinical and mechanistic reasoning, and where the distinction between knowing and knowing why has clear implications for safe practice. While the construct is likely to generalise to higher-order assessment more broadly, the present validity argument is situated within the health professions education context in which the data were collected.

Data were collected across three successive cohort years. The 2020-21 cohort was designated as the primary cohort for all validity analyses. The 2021-22 and 2022-23 cohort data were available but were not used as formal replication evidence, for two empirically grounded reasons. Item-level text comparison between the 2021-22 and 2022-23 administrations identified 71.9% of questions as identical or closely similar in wording. Systematic analysis of difficulty index changes on these repeated items revealed that 19 of 37 items with difficulty data in both years (51.4%) showed increases exceeding 0.15 between administrations, and 13 (35.1%) showed increases exceeding 0.30, with individual items increasing by as much as 0.747 on identical question stems. Increases of this magnitude on unchanged questions are inconsistent with teaching effect alone and are consistent with prior item familiarity across cohort years. In open-book online assessment contexts, the informal retention and sharing of assessment materials between successive cohort groups is a recognised challenge that is difficult to prevent (Sindre & Vegendla, 2015), and assessments administered through general-purpose survey platforms such as Microsoft Forms do not permit the access controls available in dedicated assessment management systems. This finding provided the empirical basis for the conservative decision to anchor the present validity argument in the one cohort where this concern did not apply. The item-drift finding also has implications for assessment design more broadly, which are addressed in the Discussion.

A further cohort (2023-24) was excluded on separate grounds. Assessments were administered via Microsoft Forms in an open-book format, as in previous years. From 2023-24, despite an overhaul of the question bank, generative AI tools became widely and freely accessible to students, and the Microsoft Forms environment did not permit proctoring or AI-use detection. The DDMI construct assumes that a student's confidence judgement reflects their own epistemic access to the basis of their response. Where AI assistance cannot be excluded, this assumption cannot be maintained: a student's reported confidence in an AI-assisted answer does not reflect their own metacognitive articulation of the reasoning, and DDMI responses therefore become uninterpretable within the intended construct framework. Thus the 2023-24 cohort was excluded on grounds of construct incompatibility rather than protocol failure.

Setting and participants

The primary cohort comprised 76 third-year undergraduate nursing students enrolled in applied pharmacology modules – Clinical Pharmacology I (N301C) and Clinical Pharmacology II (N304C) – at the Royal College of Surgeons in Ireland – Medical University of Bahrain during the 2020-21 academic year. The assessment activities formed part of routine teaching and carried a low summative weighting. Scores were recorded for both the individual-test and group-test components, with the assessment contributing 10% of the course mark overall. This weighting was intended to encourage preparation and engagement while limiting high-stakes pressure. Ethical approval was obtained from RCSI-MUB research ethics committee, REC/2020/23/14-Jan-2020. Participation in the research component was voluntary and did not affect course progression or summative assessment outcomes. Data were pseudonymised prior to analysis.

Assessment context

The DDMI was embedded within a TBL-derived, same-item individual-group testing sequence based on the individual and group testing components of the TBL readiness-assurance process. Because the protocol did not include a separate team application exercise, it is not described here as full team-based learning, and the terms iRAT and gRAT/tRAT are not used. The two stages are referred to as the individual test and the group test. Application-level, problem-solving, and case-based questions were embedded directly within both stages. All sessions in the primary cohort were delivered online due to COVID-19 restrictions. Blackboard Collaborate was used as the synchronous online collaboration environment, while Microsoft Forms was used as the questionnaire and test platform for submitting individual-test and group-test responses. Students were randomly assigned to groups for each session with online breakout groups operational.

All assessments were administered in an open-book format. Students were permitted to consult lecture notes and module PowerPoint materials. The question sets were informally checked against standard web searches. This check indicated that web searching could only rarely provide information relevant to the response; the assessment should therefore not be presented as search-proof or as a secure test of unaided reasoning, but this was controlled to some degree. Generative AI tools were not available to students during the 2020-21 data-collection period in the form that later became relevant to open-book online assessment. Questions were designed at the application level, requiring clinical reasoning, mechanistic interpretation, and pharmacological judgement applied to scenarios rather than simple factual retrieval.

Prior to the first assessment, students completed a structured orientation to the DDMI at the beginning of the semester. Using mock questions and worked examples, the categories were introduced and students practised applying them to their own responses. Particular attention was given to the functional distinction between feeling certain of a response and being able to articulate the basis of that certainty – that is, between confidence magnitude and reported capacity to articulate a rationale. Implicit and

explicit reasoning were introduced only in basic practical terms, with examples showing how a response might feel familiar but remain difficult to explain, or alternatively be supported by a rationale that could be stated. Students were not asked to identify Type 1 or Type 2 processing, nor were these theoretical labels used as part of the response task. Students were asked to consider and discuss the source of their certainty for each practice item. This orientation was intended to establish a shared understanding of the intended category meanings before data collection began. Orientation materials are available from the corresponding author upon request.

Individual-test responses were the primary analytic dataset throughout the present paper. Post-discussion responses introduce additional mechanisms – peer explanation, confidence transfer, collaborative recalibration – that require separate methodological treatment and are reserved for a subsequent paper.

Instrument: Declarative Domain Metacognitive Instrument

After selecting a response to each item, students selected one DDMI response. Categories are scored 1 to 4, with higher scores representing greater reported metacognitive articulation. Table 1 presents the categories with their response options, construct labels, intended interpretations, and key caveats. For the accuracy-articulation state classification, categories may be combined operationally on the basis of a shared articulation property (DDMI 1 and DDMI 2 both lack reported explicit reasoning; DDMI 3 and DDMI 4 both include some reported explicit reasoning), but the four DDMI categories themselves remain distinct articulation states throughout all validity analyses.

Table 1

DDMI category structure, construct labels, intended interpretations, and key caveats. Categories are treated as ordinal positions within a construct map of reported metacognitive articulation, but each category also denotes a defined metacognitive articulation state. A level 4 response does not prove genuine understanding; it indicates that the student reports capacity to state a rationale, rule, mechanism, or robust recognised heuristic, which may itself be wrong.

Score	Response option	Construct label	Intended interpretation	Key caveat
1	It's a guess	Complete guess-like responding	Student reports no rationale or reliable cue that can be stated; the response is experienced as a complete or near-complete guess.	A guess may still be accurate by chance or weak implicit recognition.
2	I think that I know but I don't know how I know	Implicit/non-declarative felt knowing	Student experiences the response as familiar, plausible, or likely to be correct, but the basis remains implicit or non-declarative and cannot be articulated; theoretically aligned with feeling-of-knowing, feeling-of-rightness, heuristic monitoring, and non-declarative knowledge.	Should not be treated as direct evidence of Type 1 processing; it indicates only that the learner reports no articulable rationale.
3	I think that I know and I think that I know how I know	Mixed partial articulation	Student reports partial explicit reasoning, but the account remains incomplete and is supplemented by heuristic, implicit, or familiarity-based cues.	The rationale may be incomplete, post-hoc, misapplied, or inaccurate.
4	I know that I know and I know how I know	Explicit perceived justification	Student reports confidence as supported by a largely explicit analytic/algorithmic rationale, rule, mechanism, or robust recognised heuristic that can be stated.	Reported articulability is not verified understanding; a student may be confident-yet-inaccurate.

Comparator confidence measure

A six-category percentage confidence scale was administered at one planned comparator session (N304C TBL7). Students selected their estimated probability of being correct: 0%, 1–20%, 21–40%, 41–60%, 61–80%, or 81–100%, coded 1 to 6 for analysis, with higher values indicating greater confidence. Percentage confidence ratings are a conventional method for capturing confidence magnitude, but confidence judgements may draw on multiple cues, including fluency, familiarity, partial knowledge, and perceived evidence (Dinsmore & Parkinson, 2013; Koriat, 1997). A positive association between DDMI ratings and the percentage scale was therefore expected because both concern confidence in responses. Perfect correspondence was not expected because the percentage scale measures confidence magnitude, whereas the DDMI classifies the metacognitive articulation state of confidence.

Moderate-to-strong but incomplete associations were interpreted as evidence of convergence without redundancy.

Item scoring and difficulty

Items were scored dichotomously: correct = 1, incorrect = 0. Item difficulty indices (DI) were calculated from individual-phase responses only: $DI = \text{correct responses} / \text{total responses}$. Higher DI values indicate easier items. DI was included in adjusted regression models and used as a stratification variable in difficulty-stratified calibration analyses. A baseline calibration index ($CIBASE = \text{score} \times DDMI$, ranging from - 4 to + 4) and a difficulty-adjusted variant ($ADJCI = CIBASE \times DI$) were computed for descriptive purposes.

Items were scored dichotomously: correct = 1, incorrect = 0. Item difficulty indices (DI) were calculated from individual-phase responses only: $DI = \text{correct responses} / \text{total responses}$. Higher DI values indicate easier items. DI was included in adjusted regression models and used as a stratification variable in difficulty-stratified calibration analyses. A baseline calibration index ($CIBASE = \text{score} \times DDMI$, ranging from - 4 to + 4) and a difficulty-adjusted variant ($ADJCI = CIBASE \times DI$) were computed for descriptive purposes.

Accuracy-articulation state classification

Each item response was classified into one of four accuracy-articulation states by combining response accuracy with the DDMI articulation state reported by the student. Accurate responses at DDMI 3 or DDMI 4 were classified as accurate with reported articulation, because the learner reported reasoning that could be partially or fully stated. Accurate responses at DDMI 1 or DDMI 2 were classified as accurate without articulated rationale, because the learner selected the accurate response while reporting either guessing or felt knowing without knowing why. Inaccurate responses at DDMI 1 or DDMI 2 were classified as inaccurate without articulated rationale. Inaccurate responses at DDMI 3 or DDMI 4 were classified as inaccurate with reported articulation, because the learner reported partial or full articulable reasoning but the response was nonetheless inaccurate. This operational grouping was used solely to construct a formative accuracy-articulation typology based on the absence or presence of reported explicit reasoning. DDMI 1 and DDMI 2 are not equivalent articulation states, nor are DDMI 3 and DDMI 4 equivalent articulation states. Throughout all validity analyses, the four DDMI categories were treated as distinct ordered articulation states and were not collapsed into pairs.

Validity-evidence framework

Validity was treated as the extent to which evidence and theory support the intended interpretation and use of DDMI responses (American Educational Research Association et al., 2014; Cook & Beckman, 2006; Downing, 2003; Messick, 1995). The intended interpretation was that DDMI categories represent ordered, item-level self-reports of metacognitive articulation state, expressed as defined positions along

the construct map described above. Four sources of exploratory validity evidence were examined: content and theoretical evidence; internal category functioning; relation-to-other-variables evidence; and interpretive assessment utility through accuracy-articulation state classification. Response-process evidence was not collected. This is treated as a central limitation and the immediate next priority in the validation programme.

Statistical analyses

Analyses were conducted in IBM SPSS Statistics version 31.0 (IBM Corp., Armonk, NY) and R version 4.5.2 (R Core Team, 2025), with Python 3.13.5 (pandas 2.2.3, scipy 1.17.0, statsmodels 0.14.6) used for data preparation and exploratory analysis. The primary analytic dataset contained one row per student per item per session in the individual test. Chi-square tests with Cramer's V examined the association between DDMI category and correctness. Spearman correlations examined associations between DDMI, correctness, item difficulty, and the percentage confidence scale. Binary logistic regression examined the DDMI-accuracy relationship, with DDMI entered first as a categorical predictor (DDMI 1 as reference) and then as an ordinal predictor. The primary adjusted model specified $\text{logit}(\text{Correct})$ as a function of DDMI, item difficulty, and session, with the C-statistic and pseudo-R² reported. Within-student DDMI-accuracy correlations were calculated for students who used more than one DDMI category and contributed at least ten responses. Generalised estimating equation models clustered by student were used as clustered sensitivity analyses. An approximate crossed random-intercept binomial model with random intercepts for student and item was also fitted as a two-way clustering sensitivity analysis. Difficulty-stratified calibration grouped items into lower ($DI < 0.25$), middle ($DI 0.25\text{--}0.40$), and upper ($DI > 0.40$) strata. Comparator analyses examined whether DDMI was related to but not reducible to conventional confidence magnitude. Study-time and completion-time analyses were conducted at student-session level; additional GEE models tested whether completion time predicted responses at DDMI 3 (mixed partial articulation) or DDMI 4 (explicit perceived justification), and inaccurate responses with reported articulation, after adjustment for item difficulty and session.

Primary findings (the ordered gradient, within-student calibration, and clustered sensitivity analyses) were robust at $p < .001$; secondary analyses (the difficulty interaction, comparator-session associations, and completion-time models) are reported as exploratory and were not corrected for multiple comparisons, and are interpreted with corresponding caution. Exploratory factor analysis, Cronbach's alpha, and test-retest reliability were not used as primary evidence, for reasons grounded in the construct-map framework described above. The DDMI is not a multi-item trait questionnaire; it is an item-level judgement instrument whose variation across items is expected and theoretically informative rather than a source of measurement error.

The analytic strategy was developed collaboratively: the first author specified the construct questions and expected relationships to be tested, while the second author led the statistical implementation, including model selection, robustness checks, and verification of quantitative results.

Results

Participant and item characteristics

The analytic cohort, item set, response volume, and assessment context are summarised in Table 2

Table 2
Participant and item characteristics for the primary cohort (2020-21), individual phase.

Characteristic	Primary cohort (2020-21)
Students, n	76
Sessions, n	11 (N301C: 4; N304C: 7)
Distinct assessment items, n	78
Individual item-level responses, n	4,258
Overall accuracy (%)	34.5
Mean item difficulty (DI)	0.349
Item difficulty range	0.130–0.797
Missing DDMI ratings, n (%)	0 (0%)
Comparator confidence session	N304C TBL7 (280 responses; 40 students)
Delivery format	Online via Blackboard Collaborate; tests submitted through Microsoft Forms (COVID-19 restrictions)
Assessment format	Open-book; lecture notes and PowerPoint permitted; no generative AI available

DDMI category use and accuracy

All four DDMI categories were used across all sessions, indicating that the instrument did not collapse into a simple high/low confidence distinction. The relatively high proportion of DDMI-1 responses is consistent with the challenging question set (mean DI = 0.349). Accuracy increased monotonically across categories, from 26.4% at DDMI 1 to 53.9% at DDMI 4. The DDMI-1 accuracy rate (26.4%) did not differ significantly from the 25% expected by chance for four-option MCQ items (exact binomial $p = .18$), consistent with the intended interpretation of this category as a genuine guess. This was confirmed formally: category 1 accuracy did not differ significantly from chance (26.4%; exact binomial test against 0.25, $p = .18$), whereas category 2 accuracy (35.0%) was decisively above chance ($p < .001$), marking a qualitative discontinuity between the guess category and the first functioning category rather than a smooth increment. The DDMI-accuracy association was statistically significant: $\chi^2(3) = 129.22$, p

< .001. Because the chi-square statistic and Cramer's V (0.174) disregard the ordering of the categories, two ordinal association measures appropriate to an ordered-by-binary relationship are also reported: Kendall's tau-b = 0.154 ($p < .001$) and Goodman-Kruskal gamma = 0.268. The distribution of DDMI category use and corresponding accuracy rates is shown in Table 3. These confirm a modest but robust ordered association in which greater reported metacognitive articulation is associated with greater accuracy.

Table 3

DDMI category use and accuracy in the primary cohort (2020-21), individual phase. $N = 4,258$ item-level responses. Chi-square(3) = 129.22, $p < .001$, Cramer's V = 0.174.

DDMI	Response option	n	% of responses	% correct (95% CI)
1	It's a guess	1,676	39.4	26.4 (24.3–28.5)
2	I think that I know but I don't know how I know	1,375	32.3	35.0 (32.5–37.5)
3	I think that I know and I think that I know how I know	793	18.6	40.4 (36.9–43.8)
4	I know that I know and I know how I know	414	9.7	53.9 (49.1–58.7)

Logistic regression evidence

Binary logistic regression confirmed the ordered DDMI-accuracy pattern. Compared with DDMI 1, DDMI 2, DDMI 3, and DDMI 4 were each associated with progressively higher odds of correctness: OR = 1.50, 1.88, and 3.25 respectively. In the unadjusted ordinal model, each one-level increase in DDMI was associated with 1.438-fold higher odds of a correct response (OR = 1.438, 95% CI 1.348–1.534, $p < .001$). In the adjusted model including item difficulty and session, this reduced only slightly to OR = 1.403 (95% CI 1.310–1.502, $p < .001$; Nagelkerke $R^2 = 0.153$; Hosmer-Lemeshow chi-square = 7.37, $df = 8$, $p = .497$, indicating good model fit). The near-identical adjusted and unadjusted values confirm that the DDMI-accuracy relationship was not attributable to item difficulty alone. The logistic regression models supporting this ordered relationship are reported in Table 4. Considered on its own, the DDMI was a modest discriminator of correct from incorrect responses (area under the ROC curve = 0.596). The higher value for the full model (0.695) is attributable substantially to the item difficulty term, which is itself derived from response accuracy and is therefore reported as a covariate rather than as an independent predictor. The DDMI's contribution should be understood not as strong standalone prediction of correctness, but as the capacity to distinguish the perceived basis of responses in a manner that nonetheless tracks accuracy reliably, as the ordered gradient and within-student analyses demonstrate.

Table 4

Logistic regression models predicting correct response. Upper three rows: categorical model with DDMI 1 as reference. Adjusted ordinal model includes DDMI, item difficulty, and session. $N = 4,191$ responses with complete item difficulty data. Model AUC = 0.695; DDMI-alone AUC = 0.596. DI odds ratio is for the full 0–1 scale (per 0.1-unit OR = 1.57). Crossed model includes random intercepts for student and item as a sensitivity analysis.

Predictor	OR	95% CI	p
DDMI 2 vs 1 (unadjusted)	1.498	1.282–1.749	< .001
DDMI 3 vs 1 (unadjusted)	1.883	1.575–2.251	< .001
DDMI 4 vs 1 (unadjusted)	3.250	2.604–4.056	< .001
DDMI per category (adjusted)	1.403	1.310–1.502	< .001
Item difficulty (DI)	92.558	56.119–152.658	< .001
Session	1.007	0.985–1.029	.538
Crossed model: DDMI	1.349	1.310–1.390	< .001

Within-student calibration

The analyses reported above pool responses across students and therefore leave open an important alternative explanation: the ordered gradient could arise entirely between students, if higher-attaining students both selected higher DDMI categories and answered more items correctly, without any individual student distinguishing their own stronger from weaker responses. To address this, the DDMI-accuracy association was computed separately within each student who used more than one DDMI category and contributed at least ten responses (73 of 76 students). The within-student Spearman correlations were positive for 83.6% of students, with a mean of 0.117 (median 0.114, 95% CI 0.082 to 0.152) and a one-sample t-test against zero confirming the effect ($t = 6.52$, $p < .001$). The ordered DDMI-accuracy relationship is therefore a within-person phenomenon: individual students, judging their own responses, distinguished items they were more likely to answer correctly from those they were not. This is the strongest available evidence that the instrument captures a genuine within-person monitoring capability rather than a between-student ability artefact.

Accounting for clustering

Because each student contributed multiple responses (mean 56 per student) and each item was answered by multiple students, the assumption of independent observations underlying the chi-square and unadjusted logistic analyses is violated on two axes. To confirm that the DDMI-accuracy relationship was not an artefact of this clustering, a logistic generalised estimating equation model with an exchangeable working correlation structure, clustered by student, was estimated with item difficulty as a covariate. The DDMI effect held under this clustered specification (OR = 1.346 per category increase,

95% CI 1.255 to 1.444, $p < .001$), only marginally attenuated from the unadjusted estimate. A crossed random-intercept binomial sensitivity model produced a similar DDMI effect (OR = 1.349 per category increase; approximate 95% interval 1.310 to 1.390), with random-intercept standard deviations of 0.312 for student and 0.065 for item. The DDMI-accuracy association therefore persisted after accounting for clustering by student and was not materially altered when item-level clustering was also represented.

Sensitivity to item difficulty

At the item level, mean DDMI rating correlated positively with item difficulty index: Spearman $r_s = 0.308$, $p = .006$, $N = 77$ items. Easier items elicited greater reported metacognitive articulation. This finding is theoretically expected and should not be treated as confounding: it indicates that DDMI ratings varied appropriately with task challenge. Critically, DDMI remained associated with correctness after adjustment for item difficulty in the regression models, confirming that DDMI ratings carried information about the probability of correct responding beyond what item difficulty alone would predict.

Difficulty-stratified calibration

DDMI calibration differed systematically across item difficulty strata. In easier items (upper DI stratum), higher DDMI categories were more consistently associated with correct responses. In harder items (lower DI stratum), the gradient attenuated markedly and inaccurate responses with reported articulation were substantially more common. This pattern is consistent with the theoretical expectation that reported metacognitive articulation is more reliably grounded when questions are accessible, and more likely to be misplaced when questions are genuinely challenging. Difficulty-stratified accuracy and inaccurate-with-reported-articulation rates are presented in Table 5.

This fallibility is not confined to the hardest items. Across the full dataset, even responses at the highest category - DDMI 4, in which students reported that they knew that they knew and knew how they knew – were inaccurate 46.1% of the time. Far from undermining the instrument, this is among its most important findings. It demonstrates directly, in an authentic assessment setting, that strong reported capacity to articulate a rationale is no guarantee of an accurate response, and it quantifies the gap between felt and actual understanding. This is the illusion of explanatory depth (Rozenblit & Keil, 2002) rendered measurable at the item level. It is also consistent with the self-consistency model (Koriat, 2012), in which the considerations a learner samples on harder items may be mutually consistent yet collectively wrong, producing confident error precisely where the sampled cues are most misleading.

Table 5

Accuracy by DDMI category within item difficulty strata. Table includes 4,191 responses with complete item difficulty values. The DDMI by DI interaction was significant (OR = 2.055, $p = .004$).

Difficulty stratum	Total n	DDMI 1 accuracy %	DDMI 2 accuracy %	DDMI 3 accuracy %	DDMI 4 accuracy %	Inaccurate with reported articulation n (%)	Spearman's r_s (p)
Lower DI < 0.25, hardest	1,431	17.7	19.0	22.4	31.5	271 (18.9%)	0.077 (.004)
Middle DI 0.25–0.40	1,356	25.0	34.8	34.8	49.6	224 (16.5%)	0.137 (< .001)
Upper DI > 0.40, easiest	1,404	38.1	51.9	59.9	73.7	161 (11.5%)	0.232 (< .001)

Comparison with conventional percentage confidence scale

In the comparator session (N304C TBL7), 40 students each responded to seven items. Because item-level responses are clustered within students, associations were computed at the student level ($n = 40$ independent observations) to avoid overstating precision. DDMI ratings were positively associated with the percentage confidence scale ($r_s = 0.537$, $p < .001$), supporting convergence between the two instruments. Of percentage-scale responses, 62.5% were concentrated in the two lowest categories (0%: 16.4%; 1–20%: 46.1%), reflecting broadly low confidence on a challenging question set and indicating that the conventional scale had limited discriminative range in this context. The comparator-session associations are summarised in Table 6.

Despite this limited range, the DDMI retained differential association with accuracy at the student level: mean DDMI rating correlated with student accuracy ($r_s = 0.360$, $p = .023$), whereas mean percentage-scale confidence did not ($r_s = 0.016$, $p = .922$). This dissociation – a significant DDMI-accuracy association alongside a null scale-accuracy association in the same session, on the same items, from the same students – is the central comparator finding. These findings support the interpretation that DDMI ratings carry information about response accuracy beyond what conventional confidence magnitude captures, even when that magnitude is constrained by limited discriminative range. This comparator finding rests on a single session with 40 students and should be regarded as preliminary and exploratory; it requires replication through planned parallel administration of both instruments across multiple sessions.

Table 6

Comparison between DDMI and six-category percentage confidence scale, comparator session (N304C TBL7; 40 students, 280 item-level responses; all associations computed at the student level ($n = 40$) to avoid overstating precision given the clustered data structure).

Analysis	Estimate	p	n	Interpretation
DDMI vs percentage scale, Spearman r_s	0.537	< .001	40	Convergence – related but not identical (student-level)
DDMI vs accuracy (student-level)	0.360	.023	40	DDMI significantly tracks accuracy
Scale vs accuracy (student-level)	0.016	.922	40	Conventional scale does not
DDMI vs scale (student-level)	0.537	< .001	40	Convergence, not redundancy

Convergent behavioural evidence

Three convergent behavioural findings provide indirect support for the intended category interpretation. First, DDMI-1 responses ('It's a guess') produced an accuracy rate of 26.4%, which did not differ significantly from the 25% expected by chance for a four-option MCQ item (exact binomial $p = .18$), whereas category 2 was decisively above chance (35.0%, $p < .001$). Students selecting the guess category performed at a level indistinguishable from chance, consistent with the intended interpretation that this category indicates the absence of any reportable basis for the response – not merely a low level of confidence.

Second, self-reported study time predicted mean DDMI ratings (Spearman $r_s = 0.393$, $p < .001$, session-level) but was not associated with accuracy ($r_s = 0.069$, $p = .186$). Greater preparation increased students' reported metacognitive articulation without a corresponding improvement in actual accuracy. This supports the interpretation that DDMI ratings reflect something beyond preparation effort – specifically, the quality of students' reported articulation of the basis of their reasoning on individual items – and is consistent with the observation that familiarity and fluency built through preparation can inflate felt justification without guaranteeing correct reasoning.

Third, objective session completion time – derived from Microsoft Forms start and submission timestamps – showed a different pattern. At the session level, completion time was associated with both higher mean DDMI ratings ($r_s = 0.283$, $p < .001$) and higher accuracy ($r_s = 0.181$, $p < .001$). The contrast with study time is informative: pre-session preparation predicted DDMI without predicting accuracy, while within-session engagement predicted both. This is consistent with the interpretation that additional time-on-task may help learners access, reconstruct, or search for more explicit reasoning, although completion time is reported at the session level and should not be interpreted as an item-level process measure. The contrast is therefore offered as convergent behavioural evidence rather than as a direct construct validity claim.

These completion-time associations were not confined to accuracy and mean DDMI. Across 598 student-session observations, longer completion was also associated with reduced use of the guess category (proportion of DDMI 1 responses, $r_s = -0.264$, $p < .001$), greater combined use of DDMI 3 (mixed partial articulation) and DDMI 4 (explicit perceived justification) (proportion of responses at these articulation states, $r_s = 0.243$, $p < .001$), and a higher proportion of inaccurate responses with reported articulation ($r_s = 0.202$, $p < .001$). In adjusted GEE models, each additional 10 minutes of completion time was associated with higher odds of a response at DDMI 3 or DDMI 4 – the articulation states in which some explicit reasoning is reported (OR = 1.369, 95% CI 1.141 to 1.643, $p = .001$) and higher odds of an inaccurate response with reported articulation (OR = 1.335, 95% CI 1.140 to 1.564, $p < .001$). A tertile analysis clarifies the pattern: students in the fastest within-session tertile showed lower accuracy (29.4%), lower mean DDMI (1.71), and more guess responses (54.8%) than those in the slowest tertile (accuracy 38.4%, mean DDMI 2.22, guess responses 28.8%). The slower profile therefore shifted away from guessing and toward partial or explicit declarability rather than persisting in felt but inarticulate knowing. Crucially, however, inaccurate responses with reported articulation also rose across the tertiles, from 11.6% in the fastest to 20.5% in the slowest. Longer time-on-task was thus associated with both more rationale-oriented responding and more confident error, indicating that additional time supported productive reasoning in some cases and effortful but flawed justification in others.

Accuracy-articulation state classification

DDMI responses enabled classification of all 4,258 responses into four accuracy-articulation states. Of these, 543 (12.8%) were accurate with reported articulation, 924 (21.7%) were accurate without articulated rationale, 2,127 (50.0%) were inaccurate without articulated rationale, and 664 (15.6%) were inaccurate with reported articulation. The large proportion of inaccurate responses without articulated rationale reflects the challenging item set and indicates that students often recognised the absence of a statable basis when responding inaccurately. The inaccurate-with-reported-articulation category – clinically akin to confidently mistaken reasoning – identified students who reported capacity to articulate a rationale despite selecting an incorrect response, a state not identifiable from accuracy scores alone and of particular educational significance. The resulting accuracy–articulation state classification is presented in Table 7.

Table 7

Accuracy-articulation state classification. N = 4,258 item-level responses, primary cohort 2020-21, individual phase.

State	Definition	n	%	Assessment interpretation
Accurate with reported articulation	Correct at DDMI 3 or DDMI 4	543	12.8	Accurate response with reported rationale; may indicate relatively stable understanding
Accurate without articulated rationale	Correct at DDMI 1 or DDMI 2	924	21.7	Accurate response without articulated basis; possible guess, intuitive recognition, or fragile knowledge requiring consolidation
Inaccurate without articulated rationale	Incorrect at DDMI 1 or DDMI 2	2,127	50.0	Error without reported explicit reasoning; may be particularly receptive to corrective feedback
Inaccurate with reported articulation	Incorrect at DDMI 3 or DDMI 4	664	15.6	Error accompanied by a reported rationale; possible misconception, misplaced justification, poor calibration, or illusion of explanatory depth

Discussion

The present study examined reported metacognitive articulation as an item-level assessment construct and tested exploratory validity evidence for the DDMI as one operationalisation of that construct. DDMI categories were used across the full range, showed an ordered relationship with accuracy that persisted after adjustment for item difficulty and session, were sensitive to item difficulty in a theoretically expected direction, related to but were not reducible to conventional confidence magnitude, and enabled classification of accuracy-articulation states not identifiable from correctness alone. The dual-process interpretation sharpens, but does not alter, this validity argument: DDMI 1 is best interpreted as guess-like responding, DDMI 2 as Type 1-like felt knowing, DDMI 3 as mixed Type 1/Type 2 partial articulation, and DDMI 4 as predominantly Type 2-like explicit perceived justification. Each of these findings warrants discussion in relation to the theoretical framework and the broader assessment argument.

The ordered accuracy gradient and construct-map functioning

The monotonic increase in accuracy across DDMI categories, from near-chance at DDMI 1 to 53.9% at DDMI 4, is the primary evidence that the instrument's ordered categories function as the proposed construct map predicts. If the categories were arbitrary or simply reflected confidence magnitude in a repackaged form, accuracy would not necessarily increase monotonically across them, nor would this pattern persist after adjustment for item difficulty. Its persistence in the adjusted model (OR = 1.403 per unit increase) indicates that reported metacognitive articulation carried predictive information about response accuracy beyond what item difficulty alone predicted, even though the DDMI's standalone discrimination was modest (AUC = 0.596). This is consistent with the intended construct interpretation

and with the hypothesis that students who report a more explicit articulation state for their confidence are more likely to be accurate.

The near-chance accuracy of DDMI-1 responses (26.4% against 25% chance) is particularly informative. This finding would fail if students were using DDMI 1 simply as a low-confidence label while still drawing on partial knowledge or implicit recognition: such students would perform above chance even when selecting the guess category. Its near-chance rate supports the interpretation that students used DDMI 1 as intended – to indicate the absence of any reportable basis – and constitutes convergent behavioural evidence for the construct interpretation of the lowest category.

From confidence magnitude to basis of confidence

The comparison with the six-category percentage confidence scale is the most directly relevant finding for this article's central claim. Both instruments measure something related to confidence, and both were administered simultaneously to the same students on the same items in the comparator session. Yet they showed different relationships with performance. At the student level ($n = 40$), DDMI was significantly associated with accuracy ($r_s = 0.360$, $p = .023$), whereas the percentage scale was not ($r_s = 0.016$, $p = .922$), even though the two instruments correlated substantially with each other ($r_s = 0.537$).

This pattern is consistent with the theoretical distinction drawn in the Introduction. Conventional confidence magnitude captures how certain the learner feels, which in a challenging open-book applied pharmacology context may not discriminate well between accurate and inaccurate responses, particularly when, as was the case here, 62.5% of responses clustered in the two lowest percentage categories. The DDMI's question is not simply how strong the certainty is, but which metacognitive articulation state the learner reports for that certainty: complete or near-complete guessing, Type 1-like implicit/non-declarative felt knowing, mixed Type 1/Type 2 partial articulation, or predominantly Type 2-like explicit perceived justification. This appears to carry predictive information that magnitude alone does not. The scale differences between instruments are modest in absolute terms, and this should be noted. The argument for the DDMI's distinctiveness rests on the specificity and direction of this difference, consistently observed, rather than on large effect sizes.

Fallible metacognitive articulation and the illusion of explanatory depth

A central finding that should not be minimised is that reported metacognitive articulation is fallible. The difficulty-stratified calibration analysis makes this clear: in the hardest items ($DI < 0.25$), the accuracy gradient was attenuated – 17.7%, 19.0%, 22.4%, and 31.5% across DDMI categories 1 to 4 – and 18.9% of responses in this stratum were inaccurate with reported articulation. Students reported capacity to articulate a rationale while responding inaccurately at substantial rates on the most difficult questions.

This observation reflects the illusion of explanatory depth – the tendency to believe one understands a mechanism in more detail than one can actually produce on demand (Rozenblit & Keil, 2002). In the

present study, DDMI 4 should therefore be interpreted as a learner's report that a response feels explicitly justifiable, not as evidence that the justification is complete, valid, or correctly applied. The accuracy data show that this reported justification was often fallible, especially on difficult items. Whether DDMI 4 responses are accompanied by coherent explanations, incomplete mechanisms, post-hoc rationales, or misapplied rules remains a response-process question for future work.

Inaccurate responses with reported articulation should therefore be interpreted as meaningful assessment signals rather than as measurement error. They identify cases in which learners report access to a rationale while selecting an incorrect answer, suggesting that the error is not recognised as an error. These responses are educationally consequential because they may require targeted feedback directed at the learner's underlying reasoning, rather than simple correction of the selected response. These responses were not uniformly distributed across items: The most error-prone decile of question items (the 10% of items contributing most frequently to inaccurate responses with reported articulation) accounted for 17.9% of all such responses, compared with approximately 10% expected under an even distribution, consistent with the self-consistency account in which particular items reliably elicit internally coherent but mistaken reasoning. In clinical and professional education contexts, confident misconceptions carry the greatest risk of consequence.

Metacognitive monitoring at the interface of intuition and analysis

The findings can be interpreted through the meta-reasoning framework introduced earlier. The DDMI does not observe the cognitive process that generated a response; rather, it records the learner's reported monitoring judgement after a response has been selected. An initial answer may arise from pattern recognition, partial retrieval, or more systematic reasoning, after which monitoring evaluates whether the response feels plausible and whether its basis can be articulated (Thompson, 2009; Thompson et al., 2011). Marcum (2012), in his integrated model similarly places metacognitive monitoring at the transition between non-analytic and analytic reasoning in clinical judgement, making this framework relevant to higher-order pharmacology assessment in health professions education. The present data support this interpretation by showing that these reported monitoring states were ordered with accuracy, while also demonstrating that reported articulation remained fallible.

Preparation, familiarity, and fallible articulation

The contrast between study time and completion time is theoretically informative. Greater self-reported study time was associated with higher DDMI ratings but not with higher accuracy. This suggests that preparation increased learners' reported metacognitive articulation without reliably improving whether the articulated basis was correct. In educational terms, this is a consequential dissociation: learners may study enough to recognise terminology, retrieve fragments of explanation, or generate a plausible rationale, yet still apply that knowledge incorrectly. Preparation may therefore increase fluency,

familiarity, and the availability of reasons without guaranteeing valid understanding (Dinsmore & Parkinson, 2013; Koriat, 1997; Koriat, 2012).

Completion time showed a different pattern. Longer within-session time was associated with both higher DDMI ratings and higher accuracy, suggesting that time-on-task may allow learners to search, reconstruct, or test a rationale while responding. However, longer completion time also predicted more inaccurate responses with reported articulation. This indicates that effortful reasoning was not uniformly productive: in some cases it may have supported accurate reasoning, while in others it may have produced coherent but mistaken justification. This pattern is consistent with meta-reasoning accounts in which a response may feel right, and may be supported by accessible considerations, before the validity of those considerations has been fully established (Ackerman & Thompson, 2017; Thompson et al., 2011).

A compositional interpretation: heuristic and algorithmic content

The dual-process distinction between fast, intuitive (Type 1) and slow, deliberate (Type 2) processing offers a natural but ultimately too coarse scaffold for interpreting the DDMI categories. Melnikoff and Bargh (2018) caution against treating dual-process theory as evidence for two discrete cognitive systems, arguing that features such as automaticity, consciousness, speed, control, and effort do not always cluster neatly into a simple binary distinction. This critique is consistent with the present interpretation of the DDMI categories as defined metacognitive articulation states that may contain plausible combinations of implicit and explicit elements, rather than as fixed markers of separate cognitive systems. The relevance of dual-process theory here is specific: dual-process theory helps frame the distinction between responses experienced as fluent, familiar, or intuitively right and responses for which the learner believes they can articulate an analytic or algorithmic rationale, rule, mechanism, or robust recognised heuristic. A more precise reading treats the categories as defined metacognitive articulation states. DDMI 1 represents complete or near-complete guessing, where no reliable cue or rationale is reportable. DDMI 2 represents Type 1-like implicit or non-declarative felt knowing: the learner has some sense of the correct response, but cannot articulate how that sense was produced. DDMI 3 represents a mixed Type 1/Type 2 state in which partial explicit reasoning is available but is completed or stabilised by implicit, heuristic, or familiarity-based cues. DDMI 4 represents predominantly Type 2-like explicit perceived justification: the learner believes the response rests on a largely explicit procedure, derivation, mechanism, analytic/algorithmic rationale, or robust recognised heuristic whose basis can be stated.

Two features of the data support this compositional reading more directly than a simple Type 1 versus Type 2 account would. First, the boundary between categories 1 and 2 was a genuine discontinuity rather than a smooth step. Category 1 accuracy did not differ from four-option chance (26.4%, exact binomial $p = .18$), whereas category 2 accuracy was decisively above chance (35.0%, $p < .001$). This is the qualitative signature expected if category 1 reflects complete or near-complete guessing, performing no better than chance, while category 2 reflects an implicit or non-declarative response structure that tracks

something real about the problem despite being unavailable to articulation. The lowest point of the scale is thus not maximal Type 1 processing but the absence of any usable reported process, thus functioning Type 1-like implicit or heuristic monitoring begins at category 2.

Second, the relationship between DDMI category and accuracy was moderated by item difficulty in a theoretically informative direction. The significant DDMI by difficulty interaction ($OR = 2.055$, $p = .004$) indicated that higher DDMI categories benefited disproportionately from easier items and lost the most accuracy as item difficulty increased. This might initially suggest that the higher categories are more fragile. A more coherent interpretation is that higher categories make stronger claims about explicitness and articulability, and item difficulty provides the more stringent test of whether those claims are warranted.

This interpretation is reinforced by the behaviour of the hardest items considered alone. Within the lower difficulty stratum, accuracy across categories 1 to 4 was 17.7%, 19.0%, 22.4%, and 31.5% respectively – a gradient that is preserved in direction but severely compressed, and in which even the highest category, representing believed-complete predominantly Type 2-like articulation, yielded only about a third correct. Reported metacognitive articulation is therefore least reliable as an indicator of accuracy on the most demanding material, where the consequences of confident error are greatest. In clinical and professional reasoning contexts, this is the combination that most warrants attention: strong perceived justification attached to difficult problems is the configuration in which confident error is most likely.

These process-level readings are necessarily interpretive. The DDMI records a self-reported category and an outcome; it does not directly observe whether an algorithm, a heuristic, or an implicit/non-declarative process generated the response. The compositional account is therefore offered as a theoretically disciplined interpretation of the observed accuracy structure, not as a measurement of cognitive process. It nonetheless yields a specific, falsifiable prediction for the planned response-process work: when written rationales are collected, incorrect category 4 responses should disproportionately reveal incomplete mechanisms, misapplied algorithms, or robust but misleading heuristics rather than complete valid derivations, and category 3 responses should reveal explanations in which explicit reasoning is present but mixed with implicit or heuristic assumptions. Confirmation of this pattern would convert the present interpretation into measured evidence.

The item-drift finding and its implications for assessment design

The item-drift analysis conducted to evaluate later cohorts produced a finding of broader relevance for assessment practice. Of 37 repeated questions with difficulty data in both years, 51.4% showed DI increases exceeding 0.15 between administrations, with one item increasing by 0.747 on an identical question stem. These magnitudes substantially exceed what would be expected from teaching effect alone and are consistent with prior item familiarity – the informal retention and sharing of assessment materials between successive cohort groups, a recognised challenge in open-book online assessment (Sindre & Vegdla, 2015).

This finding illustrates a structural vulnerability of open-book online assessments that use a fixed question bank across cohort years, particularly when administered through general-purpose platforms without access control. Even questions designed to require genuine application rather than recall are not immune to this effect when prior familiarity with the specific item exists. The practical implication is that systematic item refreshment between cohort years is necessary when open-book online assessment formats are used, and that item-drift analysis of the kind conducted here provides a practical empirical method for detecting when this has not been achieved.

The compositional interpretation advanced above also casts the excluded cohorts in a specific light. Question leakage does not merely add noise; it converts items that previously required a derivation into items solvable merely by recognition of a remembered question. In the terms used here, leakage manufactures a recognition heuristic where the instrument was designed to detect an algorithm, systematically inflating reported metacognitive articulation while severing it from genuine derivation. This predicts a particular distortion: the accuracy gradient should compress, as recognition lifts responses that would otherwise have been reported at DDMI 1 (guessing) or DDMI 2 (implicit felt knowing), while DDMI 3 (mixed partial articulation) and DDMI 4 (explicit perceived justification) cannot rise much further. The later cohorts showed precisely this compression relative to the primary cohort. Although these cohorts were excluded from the validity analysis for the reasons given, the direction of their distortion is itself consistent with the interpretation that DDMI 3 and DDMI 4 index perceived algorithmic articulation rather than confidence magnitude alone – since it is exactly recognition-based responding that the construct should, and apparently did, fail to distinguish from genuine justification when leakage made the two phenomenologically identical.

Implications for health professions and problem-space assessment

The DDMI has particular relevance for health professions education because clinical and professional learning often requires judgement under uncertainty. In these contexts, the educational problem is not only whether a learner selects the correct diagnosis, explanation, mechanism, or management option, but whether the learner recognises the basis and limits of that judgement. Accuracy scores identify correctness, and conventional confidence ratings can identify certainty magnitude, but neither directly distinguishes uncertainty, intuitive recognition, partial reasoning, or reported justification (Curtis et al., 2013; Dinsmore & Parkinson, 2013; Tweed et al., 2020). The DDMI therefore adds formative value by making visible accuracy–articulation states that have different implications for feedback, remediation, and the analysis of judgement.

This distinction is especially important where learners are wrong but report partial or explicit articulation. Such responses suggest that the learner is not merely uncertain or guessing, but has attached a rationale, rule, mechanism, or recognised heuristic to an inaccurate judgement. In health professions education, this state aligns with concerns about misplaced certainty, overconfidence, and error in clinical reasoning, where learners may hold an inaccurate judgement as though it were justified (Berner & Graber,

2008; Norman et al., 2017). Clinical judgement is not only an outcome, but a metacognitive state: a decision may be guessed, intuitively plausible, partially reasoned, or explicitly justified from the perspective of the person making it. In case-based teaching, simulation, workplace-based discussion, and clinical reasoning assessment, DDMI-style responses could help supervisors and learners distinguish uncertainty, intuitive pattern recognition, partial reasoning, and confidently held but erroneous judgement. Such use would require further validation, but the present findings suggest that the accuracy–articulation distinction is educationally meaningful and potentially transferable to clinical decision analysis.

This logic may become increasingly relevant in AI-supported clinical decision-making. As AI-based clinical decision support systems become more prominent in health professions education and practice, educators will need ways to examine not only whether learners accept, reject, or modify an AI suggestion, but how they understand the basis of their own judgement in relation to that suggestion. A DDMI-style judgement could help distinguish clinicians who defer to an AI output without an articulated rationale from those who use the output to support, challenge, or refine an explicitly reasoned clinical judgement. In this context, the DDMI could provide a structured way to analyse clinician judgement, automation bias, misplaced certainty, and the extent to which AI-supported decisions remain explainable to the learner or clinician (Greengrass, 2026).

This framework is not limited to pharmacology or to health professions education. Its broader relevance is to assessments that require learners to operate within problem spaces: tasks in which they must interpret information, select among plausible alternatives, apply principles, justify decisions, or transfer knowledge to unfamiliar situations. In such contexts, knowing whether a learner guessed, recognised a familiar pattern, reasoned partially, or believed they could justify their response may be educationally consequential. The DDMI could therefore be adapted for disciplines such as law, engineering, ethics, business, mathematics, and the sciences, wherever assessment aims to examine applied judgement rather than recall alone.

Limitations

Several limitations should be acknowledged directly. The evidence base is a single cohort from one institution in one disciplinary context. This is perhaps sufficient for exploratory validity evidence but not for claims about the instrument's behaviour across higher education more broadly. The DDMI is a self-report instrument: it captures students' reported metacognitive articulation state, not the cognitive processes that generated their responses, and reported articulation may not fully correspond to what occurred cognitively.

Formal response-process evidence was not collected, and this remains the most important limitation. The present findings show that DDMI categories behaved in theoretically expected ways, but they do not, on their own, establish how students understood and interpreted the response options. This limitation is mitigated, though not removed, by the structured orientation described in the Methods: students were

introduced to the categories, practised applying them, and explicitly discussed the distinction between confidence magnitude and reported capacity to articulate a rationale before data collection began. This increases confidence that students approached the categories with an interpretation consistent with their intended meaning. It does not, however, establish that those interpretations were retained consistently across all eleven sessions or applied uniformly to every item, nor that each student's internal interpretation matched the intended construct at the moment of response. A student might still select category 4 because they are highly confident rather than because they have consciously accessed an explicit rationale, or select category 2 because an answer feels familiar without their being able to verbalise why. These possibilities require cognitive interviews, think-aloud protocols, and coded written rationales to address, as set out below.

The comparator analysis was restricted to one session (40 students) in which the percentage scale showed limited discriminative range, which may have reduced its apparent association with accuracy; future studies should administer both instruments in parallel across a wider range of items.

Finally, group-phase and post-discussion data were not analysed in the present paper. These data are available and will be examined in a subsequent paper focused on peer influence, collaborative recalibration, and group metacognitive dynamics in the same teaching sequence.

Further Work

The immediate next step is response-process evidence. Cognitive interviews should examine how students interpret each DDMI category and, in particular, whether the boundary between categories 2 and 3 – felt but inarticulate knowing versus partial articulability – is interpreted as intended. Think-aloud protocols during problem-solving would allow DDMI selections to be compared with concurrently verbalised reasoning. A particularly informative design would pair DDMI responses with immediate written explanations: if DDMI 4 selections are consistently accompanied by coherent mechanistic explanations, this would support the construct interpretation; if they are frequently accompanied by brief, vague, or post-hoc justifications, this would suggest the illusion of explanatory depth (Rozenblit & Keil, 2002) operates more widely than the accuracy data alone would indicate. Written rationales also offer a pragmatic classroom-based alternative to laboratory methods.

A further, more speculative possibility is that reported metacognitive articulation has a developmental trajectory. A relationship first learned as an explicit derivation may, through repetition and consolidation, become more compressed, automatic, or semanticised: retained as confident knowledge while becoming less tied to the original contextual or derivational pathway (Dudai et al., 2015; Winocur et al., 2010). This possibility is consistent with accounts of memory transformation from context-rich episodic representation toward more decontextualised semantic knowledge, with cognitive skill theories of proceduralisation, and with medical expertise accounts of knowledge encapsulation and illness-script formation (Schmidt & Rikers, 2007; Tulving, 1985). Longitudinal DDMI work could test whether movement across articulation states represents increasing expertise, loss of accessible rationale, or

both, extending recent work showing that aspects of learner self-monitoring can be tracked across medical training using MCQ response certainty (Tweed et al., 2024).

Conclusion

The present study introduces reported metacognitive articulation as an item-level assessment construct and provides exploratory validity evidence for the DDMI as one operationalisation of that construct within a construct-map framework. DDMI categories showed: (1) ordered relations with accuracy consistent with the proposed construct progression; (2) sensitivity to item difficulty; (3) convergence with, but non-redundancy relative to, a conventional percentage confidence scale; and (4) the capacity to classify accuracy–articulation states not identifiable from correctness alone. Interpreting the categories through a dual-process lens sharpened the meaning of this progression: from guess-like responding, through Type 1-like felt knowing and mixed Type 1/Type 2 partial articulation, to predominantly Type 2-like explicit perceived justification. Convergent behavioural evidence supported the intended interpretation of the guess category, while the dissociation between reported study time and accuracy indicated that DDMI ratings were not reducible to general preparation effort.

The DDMI should not be interpreted as a completed psychometric scale or as a direct measure of cognitive processing systems. Its value lies in distinguishing how confident a learner is from the metacognitive articulation state reported for that confidence. This distinction has immediate formative relevance because it makes visible accurate responses without articulated rationale and, perhaps more importantly, inaccurate responses accompanied by reported justification. These latter responses identify cases where learners may hold erroneous judgements as though they were reasoned, making them especially important targets for feedback and conceptual remediation. Response-process validation is therefore the immediate next step: future work must establish how learners interpret the categories and whether reported articulation corresponds to written or verbalised reasoning. The item-drift finding also identifies a practical assessment design challenge: item exposure across cohort years in open-book online contexts. The present paper provides the empirical foundation for the next stages of this validation programme and for wider exploration of the DDMI in health professions and problem-space assessment.

Declarations

Funding: No external funding was received for this study.

Competing interests. The authors declare no competing financial or non-financial interests directly or indirectly related to this work.

Ethics approval. Ethical approval for this study was obtained from the RCSI-MUB Research Ethics Committee (REC/2020/23/14-Jan-2020). The assessment activities formed part of routine teaching; participation in the research component was voluntary and did not affect course progression. Data were pseudonymised before analysis.

Consent to participate. Informed consent was obtained from all participants for use of their pseudonymised assessment data for research purposes.

Authors contributions

Colin Greengrass conceptualised the study, developed the DDMI, designed and implemented the educational protocol, collected and curated the data, specified the construct-validity argument, identified the relationships to be tested, interpreted the findings, and drafted the manuscript. Kianna Thelen led the statistical analysis, advised on model selection and robustness checks, contributed to interpretation of the quantitative findings, and critically reviewed and revised the manuscript. Both authors approved the final version of the manuscript.

Acknowledgements

The authors thank the participating students for their engagement with the assessment activities and for consenting to the use of their pseudonymised assessment data for research purposes. The authors also thank Professor Steve Safrany, formerly of RCSI-MUB and now at Keele University, for his invaluable support with the implementation and delivery of the learning sessions in which the DDMI was embedded.

Data availability

The pseudonymised data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to institutional ethical approval and participant confidentiality requirements.

References

1. Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607–617. <https://doi.org/10.1016/j.tics.2017.05.004>
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
3. Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, 121(5), S2–S23. <https://doi.org/10.1016/j.amjmed.2008.01.001>
4. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
5. Caleon, I. S., & Subramaniam, R. (2009). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32(7), 939–961. <https://doi.org/10.1080/09500690902890130>

6. Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition. *Educational Psychologist*, 46(3), 141–167.
<https://doi.org/10.1080/00461520.2011.587722>
7. Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia. *Science*, 210(4466), 207–210. <https://doi.org/10.1126/science.74143>
8. Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2), 166e167–166e116.
<https://doi.org/10.1016/j.amjmed.2005.10.036>
9. Curtis, D. A., Lind, S. L., Boscardin, C. K., & Dellenges, M. (2013). Does student confidence on multiple-choice question assessments provide useful information? *Medical Education*, 47(6), 578–584. <https://doi.org/10.1111/medu.12130>
10. Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, 24, 4–14. <https://doi.org/10.1016/j.learninstruc.2012.06.001>
11. Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830–837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>
12. Dudai, Y., Karni, A., & Born, J. (2015). The Consolidation and Transformation of Memory. *Neuron*, 88(1), 20–32. <https://doi.org/10.1016/j.neuron.2015.09.004>
13. Echternacht, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research*, 42(2), 217–236. <https://doi.org/10.2307/1170018>
14. Eva, K., & Regehr, G. (2005). Self-assessment In The Health Professions: A Reformulation And Research Agenda. *Academic medicine: journal of the Association of American Medical Colleges*, 80, S46–54. <https://doi.org/10.1097/00001888-200510001-00015>
15. Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241.
<https://doi.org/10.1177/1745691612460685>
16. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10), 906. <https://doi.org/10.1037/0003-066X.34.10.906>
17. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition [Review]. *Frontiers in Human Neuroscience*, Volume 8–2014. <https://doi.org/10.3389/fnhum.2014.00443>
18. Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16(4), 309–313. <https://doi.org/10.3758/BF03197041>
19. Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology*, 3(1), 80–85. <https://doi.org/10.1080/0968776950030113>
20. Greengrass, C. J. (2026). Transforming clinical reasoning—the role of AI in supporting human cognitive limitations [Review]. *Frontiers in Digital Health*, Volume 7–2025.
<https://doi.org/10.3389/fdgth.2025.1715440>

21. Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56(4), 208–216. <https://doi.org/10.1037/h0022263>
22. Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, 34(5), 294–299. <https://doi.org/10.1088/0031-9120/34/5/304>
23. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
24. Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67(1), 88–140. <https://doi.org/10.3102/00346543067001088>
25. Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American psychologist*, 58(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
26. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedem.12000>
27. Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4), 609–639. <https://doi.org/10.1037/0033-295X.100.4.609>
28. Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
29. Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119(1), 80–113. <https://doi.org/10.1037/a0025648>
30. Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20(2), 159–183. [https://doi.org/10.1016/0030-5073\(77\)90001-0](https://doi.org/10.1016/0030-5073(77)90001-0)
31. Marcum, J. A. (2012). An integrated model of clinical reasoning: dual-process theory of cognition and metacognition. *Journal of evaluation in clinical practice*, 18 5, 954–961. <https://doi.org/10.1111/j.1365-2753.2012.01900.x>
32. Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, 22(4), 280–293. <https://doi.org/10.1016/j.tics.2018.02.001>
33. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
34. Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *The psychology of learning and motivation* (Vol. 26, pp. 125–173). [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
35. Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, 92(1), 23–30. <https://doi.org/10.1097/ACM.0000000000001421>

36. Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118(3), 219–235. <https://doi.org/10.1037/0096-3445.118.3.219>
37. Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1
38. Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>
39. Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12), 1133–1139. <https://doi.org/10.1111/j.1365-2923.2007.02915.x>
40. Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 415–429). Routledge. <https://doi.org/10.4324/9780203876428>
41. Sindre, G., & Vegendla, A. (2015). E-exams versus paper exams: A comparative analysis of cheating-related security threats and countermeasures. Norwegian Information Security Conference (NISK), Norway.
42. Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82(3), 171–177. <https://doi.org/10.1016/j.nlm.2004.06.005>
43. Stanovich, K. E. (2009). *What Intelligence Tests Miss: The Psychology of Rational Thought*. Yale University Press.
44. Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. Oxford University Press.
45. Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans, & K. Frankish (Eds.), *Two Minds: Dual Processes and Beyond* (pp. 171–195). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199230167.003.0008>
46. Thompson, V. A., Turner, P., J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3), 107–140. <https://doi.org/10.1016/j.cogpsych.2011.06.001>
47. Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1), 1–12. <https://doi.org/10.1037/h0080017>
48. Tweed, M., Purdie, G., & Wilkinson, T. (2017). Low performing students have insightfulness when they reflect-in-action. *Medical Education*, 51(3), 316–323. <https://doi.org/10.1111/medu.13206>
49. Tweed, M., Purdie, G., & Wilkinson, T. (2020). Defining and tracking medical student self-monitoring using multiple-choice question item certainty. *Bmc Medical Education*, 20(1), 344. <https://doi.org/10.1186/s12909-020-02250-x>
50. Tweed, M., Willink, R., & Wilkinson, T. J. (2024). Using MCQ response certainty to determine how aspects of self-monitoring develop through a medical course. *Medical Education*, 58(5), 535–543. <https://doi.org/10.1111/medu.15253>
51. Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410611697>

52. Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, 48(8), 2339–2356.
<https://doi.org/10.1016/j.neuropsychologia.2010.04.016>