
Cytross Nursing: Advancing the Nursing Education Frontiers

Cytross Nursing Team

Technical Report · August 26, 2026

An intelligent tutoring system at the intersection of nursing education, adaptive learning, and psychometrics.

Abstract

Preparation for nursing licensure sits at an uncomfortable junction: the examination it anticipates is an adaptive, criterion-referenced measurement instrument grounded in modern psychometrics, while most preparation tools remain static banks of practice questions scored by raw percent-correct. This report describes the methodology behind **Cytross Nursing**, an intelligent tutoring system (ITS) whose core is **Dora**, an adaptive, personalized learning agent — equivalently, the decision policy that governs each interaction — and whose pedagogical complement is the **Deep Review** library, a searchable corpus of textbook-depth topic lectures coupled to the item bank. We position the system in the tri-disciplinary space of nursing education, adaptive learning, and psychometrics, and we formulate the problems it is designed to solve: continuous latent-ability estimation under joint learner–item uncertainty; clinically safety-aware scoring of polytomous responses, in which *which* wrong answer was chosen carries measurement value; adaptive item selection under dual practice and assessment regimes with an information-based, measurement-precision stopping criterion; session-level equating of safety signals across nonequivalent forms; fair, anti-memorization peer duels; and curiosity-adaptive personalization that responds to a learner’s own reading behavior without contaminating ability estimates. We state the design requirements and guarantees for each problem, situate our choices in the peer-reviewed literature, and summarize an internal ablation program whose discipline is to change nothing that measurement evidence does not justify. Deliberately, this report specifies problem formulations, objectives, and invariants rather than reproducible parameterizations: it is a statement of method and intent, not an implementation recipe.

Keywords: intelligent tutoring systems; computerized adaptive testing; item response theory; nursing education; clinical judgment; mastery learning; knowledge tracing; test equating; personalization.

Contents

1	Introduction	4
1.1	Scope and disclosure posture of this report	4
1.2	Contributions	5
2	Background and Related Work	5
2.1	Item response theory and latent-trait measurement	5
2.2	Computerized adaptive testing	5
2.3	On-the-fly calibration and rating-system approaches	6
2.4	Knowledge tracing and learner modeling	6
2.5	Intelligent tutoring systems and mastery learning	6
2.6	Large language models and generative teaching agents	7
2.7	Retrieval practice, spacing, and desirable difficulties	7
2.8	Nursing licensure and the measurement of clinical judgment	7
3	System Overview: Dora as an Adaptive Learning Agent	8
3.1	The learner model	8
3.2	Two channels, one agent	8
3.3	The Deep Review library	8
4	Methodology: Problem Formulations and Design Principles	9
4.1	Continuous latent-ability estimation under joint uncertainty	9
4.2	Clinical-safety-aware scoring	10
4.3	Adaptive selection and measurement-precision stopping	11
4.4	Session-level equating of the safety signal	12
4.5	Fair peer duels	12
4.6	Curiosity-adaptive personalization	13
4.7	Deep Review: pedagogy against item-fitting	13
5	Ablations and Empirical Analysis	14
5.1	Plausibility-weighted difficulty: a refinement declined	15
5.2	Anchoring the correct response on the severity scale	15
5.3	Guarantee testing for curiosity-adaptive personalization	15
6	Discussion	16
6.1	What the tri-disciplinary framing buys	16
6.2	The agent stance	16
6.3	Honest positioning	16
7	Limitations and Ethical Considerations	17
8	Conclusion	17

1 Introduction

Every year, hundreds of thousands of nursing candidates prepare for the NCLEX-RN¹ examination, an adaptive licensure test whose modern form is explicitly grounded in a measurement model of clinical judgment [Dickison et al., 2019, Betts et al., 2019]. The examination they face is a psychometrically sophisticated instrument: it estimates a latent ability online, selects items adaptively, and terminates on a confidence criterion rather than a fixed count [Wainer, 2000, van der Linden and Glas, 2010]. The preparation tools most candidates use are not. The dominant paradigm remains a large, static question bank, uniform for every user, scored by percent-correct — a statistic that conflates item difficulty with learner ability, ignores response process entirely, and treats a catastrophic clinical error and a near-miss as the same “wrong.”

This mismatch is not merely aesthetic. It produces three concrete failure modes that we take as the founding problems of Cytross Nursing:

1. **Measurement failure.** Percent-correct over a self-selected subset of a bank is neither comparable across learners nor across sessions, and it supports no principled readiness decision. A learner deserves an estimate with known precision, on a stable scale, that improves with evidence — the core promise of item response theory (IRT) [Rasch, 1960, Birnbaum, 1968, Lord, 1980].
2. **Pedagogical failure.** A static bank drills the learner where the bank is dense, not where the learner is weak, and provides no mechanism for mastery-referenced progression. Decades of evidence favor tutoring architectures that adapt to the individual state of the learner [Bloom, 1984, Anderson et al., 1995, VanLehn, 2011].
3. **Clinical-safety failure.** In nursing, wrongness is not one-dimensional. Selecting an option that would harm a patient is a categorically different event from omitting a secondary intervention, yet dichotomous scoring erases the distinction that the profession itself considers primary [Dickison et al., 2019].

Cytross Nursing is an intelligent tutoring system (ITS) built to close this gap. Its core is Dora, an adaptive, personalized learning agent that maintains a multidimensional model of each learner, decides what the learner should encounter next, scores what happened in a clinically safety-aware way, and reports mastery and readiness in calibrated probabilistic terms. Its pedagogical complement is the *Deep Review* library: a searchable corpus of textbook-depth topic lectures linked to the item bank, designed to combat the failure mode we call *item-fitting* — learning the answer to a question rather than the concept beneath it — and to promote transfer to novel presentations [Barnett and Ceci, 2002].

1.1 Scope and disclosure posture of this report

This is a methodology white paper, not a product tour and not an implementation disclosure. Our aim is to state, with the precision the learning-science and psychometrics communities expect, *which problems* the system solves, *why those problems are hard*, *what requirements and invariants* our solutions satisfy, and *how the design decisions are situated* in the peer-reviewed literature. Where mathematics appears, it is restricted to standard, already-public forms — the logistic

¹NCLEX-RN® and NCSBN® are registered trademarks of the National Council of State Boards of Nursing, Inc. (NCSBN). Cytross Nursing is an independent product and is not affiliated with, endorsed by, or sponsored by NCSBN.

response function, Fisher information, standard-error-based stopping, inverse document frequency — presented as the families of methods we build upon. Dora’s specific estimators, update rules, schedules, and parameterizations are proprietary and are deliberately described at the level of their objectives and guarantees rather than their instantiation. Nothing in this report is fabricated; everything asserted is true of the deployed system. What is omitted is omitted by design.

1.2 Contributions

1. A tri-disciplinary problem formulation for nursing-licensure preparation, unifying adaptive measurement, mastery pedagogy, and clinical safety into a single agent architecture (§3).
2. A set of six methodological problem statements with explicit design requirements and invariants: continuous ability estimation (§4.1), safety-aware scoring (§4.2), dual-regime adaptive selection with precision-based stopping (§4.3), session-level safety equating (§4.4), fair peer duels (§4.5), and curiosity-adaptive personalization (§4.6).
3. A pedagogical account of the Deep Review library as an anti-item-fitting instrument grounded in retrieval practice, spacing, and transfer research (§4.7).
4. A description of our internal ablation discipline and the qualitative findings that have already shaped the system (§5).

2 Background and Related Work

2.1 Item response theory and latent-trait measurement

IRT models the probability of a correct response as a function of a latent learner trait θ and item parameters. In the simplest and most robust member of the family, the Rasch / one-parameter logistic model [Rasch, 1960], an item is characterized by a difficulty b and

$$\Pr(X = 1 \mid \theta, b) = \sigma(\theta - b) = \frac{1}{1 + e^{-(\theta - b)}}, \quad (1)$$

with well-known generalizations adding discrimination and guessing parameters [Birnbaum, 1968, Lord, 1980] and polytomous extensions grading partial credit across ordered categories [Samejima, 1969]. Two properties of this family matter for our purposes. First, ability and difficulty live on a common interval scale, so “how hard is this item for this learner” is a single subtraction. Second, the model is a measurement model: it supports standard errors, information functions, and equating, which percent-correct does not [Kolen and Brennan, 2014].

2.2 Computerized adaptive testing

Given a calibrated bank, computerized adaptive testing (CAT) selects each next item to be maximally informative about the current ability estimate and stops when the estimate is precise enough for the decision at hand [Wainer, 2000, van der Linden and Glas, 2010]. For a logistic item, the Fisher information contributed at success probability P is

$$I(\theta) = P(\theta) (1 - P(\theta)), \quad (2)$$

maximized where $P = \frac{1}{2}$ — the formal reason adaptive tests probe the frontier of a candidate’s ability rather than its comfortable interior. The running estimate $\hat{\theta}$ after administering a set J of items carries the standard error

$$\text{SE}(\hat{\theta}) = \left(\sum_{j \in J} I_j(\hat{\theta}) \right)^{-1/2}, \quad (3)$$

and a classification CAT against a passing standard θ_{pass} terminates when the confidence interval $\hat{\theta} \pm z_{\alpha/2} \text{SE}(\hat{\theta})$ falls entirely on one side of the standard [Wainer, 2000]. Operational CATs additionally constrain selection for content balance and exposure control so that maximal information does not degenerate into over-drilling a single seam of the blueprint [Chang and Ying, 1999, van der Linden and Glas, 2010].

2.3 On-the-fly calibration and rating-system approaches

Classical IRT calibrates items offline on large pretest samples. A second research line — originating in the Elo rating system for chess [Elo, 1978] and its Bayesian refinement [Glickman, 1999] — treats each response as a match between learner and item and updates both ratings online. Klinkenberg et al. [2011] demonstrated that an Elo-family model with an explicit scoring rule supports large-scale adaptive *practice* with on-the-fly ability and difficulty estimation, and Pelánek [2016] surveys the design space of such systems in education. The generic update of this family,

$$R \leftarrow R + K (S - \mathbb{E}[S]), \quad (4)$$

moves a rating R by the surprise of an observed score S relative to its expectation, at a responsiveness K . The critical design questions — how K should depend on the maturity of each side’s estimate, how outcomes should be defined when responses are polytomous, and how the practice channel relates to the measurement channel — are where systems in this family differ, and where Dora makes its proprietary choices (§4.1).

2.4 Knowledge tracing and learner modeling

Orthogonal to IRT’s static-trait view, knowledge tracing models the learner as a dynamical system: Bayesian knowledge tracing posits latent skill acquisition with learning and slip/guess parameters [Corbett and Anderson, 1994], and deep knowledge tracing replaces the parametric state with a learned recurrent one [Piech et al., 2015]. Our reading of this literature is architectural rather than doctrinal: a preparation system must model *change* (learning, forgetting, staleness of evidence) as a first-class citizen, not as noise around a fixed trait. Dora accordingly maintains explicit uncertainty about every quantity it estimates and lets that uncertainty grow with disuse — forgetting is modeled as decaying *confidence* in stale evidence rather than as punitive score decay.

2.5 Intelligent tutoring systems and mastery learning

Bloom’s two-sigma results framed the standing challenge: tutored, mastery-referenced instruction dramatically outperforms conventional group instruction, and the research program of the field is to approximate the tutor at scale [Bloom, 1984]. Cognitive tutors demonstrated that model-based tutoring systems achieve large gains in real curricula [Anderson et al., 1995], and meta-analytic work places well-designed ITSs near the effectiveness of human tutors [VanLehn, 2011]. VanLehn’s analysis of tutoring behavior — outer loops that choose tasks, inner loops that respond to steps

[VanLehn, 2006] — is a useful lens on Dora: the agent’s outer loop chooses what a learner encounters next (practice item, duel, assessment, or reading), and its inner loop scores, diagnoses, and remediates what happened.

2.6 Large language models and generative teaching agents

The newest line of work in this space reframes adaptive instruction as the province of *generative agents* built on large language models — conversational tutors, automated feedback at scale, open-ended dialog tutoring — a design space that surveys are mapping faster than deployments can validate [Kasneci et al., 2023, Wang et al., 2024]. The empirical record tempers the enthusiasm. Benchmarked against human teachers in educational dialogues, large generative models are fluent but pedagogically shallow [Tack and Piech, 2022]; dialogue-tutoring corpora built with practicing teachers show how much scaffolding it takes to make such a model teach rather than tell [Macina et al., 2023]; case studies of automated feedback find its quality real but uneven [Dai et al., 2023]; and systematic reviews catalog reliability, cost, and equity challenges that grow with deployment scale [Yan et al., 2024]. Two things in this literature matter for the present report. First, it has fixed the contemporary meaning of an adaptive *learning agent*: a system that perceives a learner’s state and commits to pedagogical action in response — which is, in decision-theoretic terms, precisely a *policy*, a mapping from learner state to instructional action, and we use the two words interchangeably for Dora from here on (§3). Second, consistent with our disclosure posture (§1.1), we characterize Dora by its decisions and the invariants they respect, and we caution the reader — here as in the Discussion (§6.2) — against inferring its computational substrate from any of the method families this section surveys.

2.7 Retrieval practice, spacing, and desirable difficulties

Testing is not merely assessment: retrieval practice produces more durable learning than restudy [Roediger and Karpicke, 2006], distributed practice outperforms massed practice with regularities that reward deliberate scheduling [Cepeda et al., 2006], and conditions that make practice harder in the short term often improve long-term retention and transfer — Bjork’s “desirable difficulties” [Bjork, 1994]. These results shape two of Dora’s commitments: practice is pitched to be challenging but winnable, in the spirit of the zone of proximal development [Vygotsky, 1978], and re-encounters with mastered material are scheduled as spaced checks rather than immediate repetition.

2.8 Nursing licensure and the measurement of clinical judgment

The Next Generation NCLEX program re-grounded the licensure examination in an explicit clinical judgment measurement model [Dickison et al., 2019] and in item formats engineered to elicit judgment rather than recall [Betts et al., 2019]. Two consequences follow for preparation systems. First, polytomous, partial-credit item formats are now the norm, so scoring must be defined for graded outcomes, not just dichotomies [Samejima, 1969]. Second, the profession’s own measurement model treats recognizing cues, analyzing them, and *acting* safely as distinct competencies — a distinction a preparation system should represent structurally rather than average away.

3 System Overview: Dora as an Adaptive Learning Agent

Dora is the decision-making core of Cytross Nursing: an adaptive, personalized learning agent that mediates every interaction between a learner and the content corpus. Throughout this report we use *agent* and *policy* interchangeably for Dora — an agent in that it perceives a learner state and acts on it, a policy in the decision-theoretic sense of a mapping from learner state to action. The two words name the same object from two vantage points, the behavioral and the mathematical, and we alternate between them as emphasis demands. We describe the architecture at the conceptual level.

3.1 The learner model

Dora maintains, for each learner, a multidimensional latent-ability state organized along three crossed facets of the nursing domain:

- **Content axes** — the clinical subject areas of the curriculum (e.g., adult cardiopulmonary care, pediatrics, maternal health, mental health, foundations), with items permitted to load on several axes at once so that multi-system presentations are represented as such;
- **Cognitive stance** — whether an item demands *recognition and analysis* of a clinical situation or a committed *action* within it, mirroring the structure of contemporary clinical judgment models [Dickison et al., 2019];
- **Competency family** — the professional competency the item exercises (e.g., safe medication practice, disease management, prioritization).

Every estimated quantity in this state carries an explicit uncertainty that shrinks with evidence and grows with disuse. Cold-start estimates are stabilized by hierarchical pooling: with little evidence, a cell of the model borrows strength from related cells in an empirical-Bayes fashion, and earns its independent value as evidence accumulates. Where a flattering estimate could be clinically consequential — the action-oriented, safety-critical region of the model — pooling is deliberately conservative: when related evidence sources disagree, the less favorable reading governs.

3.2 Two channels, one agent

A single agent serves two purposes that most products conflate: practice (learning) and assessment (measurement). Dora keeps the regimes distinct — different selection objectives, different stopping behavior, different reporting — while sharing one learner model, so that everything the learner does informs everything the agent decides (Table 1). Alongside these, a social channel (friend-to-friend duels, §4.5) and a curiosity channel (library-driven personalization, §4.6) feed the same model under strict non-contamination rules described below.

3.3 The Deep Review library

The item bank is coupled to a searchable library of textbook-depth topic lectures. Items link to the lecture that treats their underlying concept; lectures open only after an item is answered, so they never leak answers; and the library is independently browsable and searchable. Pedagogically, the library is the system’s instrument against item-fitting (§4.7); architecturally, it is also a *sensor*: what a learner chooses to read and search is treated as a signal of curiosity that the agent may respond to (§4.6).

	Practice regime	Assessment regime
Objective	Learning: durable acquisition and retention	Measurement: a defensible readiness decision
Difficulty policy	Challenging but winnable; stochastic around a motivating success band	Maximum information near the ability frontier [Wainer, 2000]
Content policy	Mastery-referenced coverage with spaced re-checks	Blueprint-balanced sampling; exposure-aware
Stopping	Learner-paced	Precision-based: stops when the ability interval clears the standard (cf. Eq. 3)
Reporting	Mastery, diagnostics, remediation	Calibrated pass-band estimate with stated uncertainty

Table 1: The two regimes Dora operates, sharing one learner model.

4 Methodology: Problem Formulations and Design Principles

We now state the six methodological problems the agent solves. For each we give the formulation, the requirements a solution must satisfy, the family of methods we build on, and the invariants our design guarantees. Consistent with §1.1, we characterize our solutions by their properties, not their parameterizations.

4.1 Continuous latent-ability estimation under joint uncertainty

Problem. Estimate, online and per response, a multidimensional learner ability and a per-item difficulty on a common logistic scale (Eq. 1), when (i) both sides of every interaction are uncertain, (ii) the learner is genuinely non-stationary (they are here to change), and (iii) the system must behave sensibly from the very first response, with no pretest calibration window for new items and no placement test for new learners.

Why it is hard. Offline IRT calibration assumes a static trait and a large calibration sample; neither holds. Pure rating-system updates [Elo, 1978, Klinkenberg et al., 2011] handle non-stationarity but face a bias-variance schedule problem: an estimator that moves fast enough to track a fast-improving novice is too jumpy for a well-measured veteran, and an estimator gentle enough for the veteran anchors the novice to their first bad week. Multidimensionality sharpens the problem: credit for a multi-system item must be apportioned across the axes it loads on, and early apportionment errors compound.

Our approach (properties). Dora’s estimator belongs to the on-the-fly, rating-family lineage of Eq. 4 [Klinkenberg et al., 2011, Pelánek, 2016], with proprietary departures that we characterize as follows:

- **Evidence-weighted responsiveness.** The magnitude of every update is governed by the *relative* uncertainty of the two parties: an uncertain learner moves quickly against a well-calibrated item, a well-measured learner moves slowly against a poorly known item, and the item’s own calibration mirrors the same principle in reverse. No fixed K exists anywhere in the system; responsiveness is a function of the state of knowledge, and it anneals as evidence accrues.
- **A smooth estimator schedule, not a switch.** The estimator interpolates continuously between a fast, bias-tolerant regime appropriate to first contact and a low-bias regime appro-

priate to a mature evidence base, as a smooth function of accumulated evidence — avoiding the discontinuities (and gaming opportunities) of hard mode switches.

- **Structured credit assignment.** Updates respect the crossed facet structure of the learner model: multi-axis items apportion credit by their loadings, shared components are updated coherently rather than independently, and the safety-critical interaction structure of the model is trained under deliberately conservative conditions so that noise cannot masquerade as safety evidence.
- **Forgetting as confidence decay.** Idle time increases the uncertainty attached to an estimate; it never directly reduces the estimate. A returning learner is re-measured efficiently rather than punished, and demonstrated, extensively-evidenced mastery is protected from being fully reabsorbed into the prior.
- **A human-legible scale.** Internally the model lives on the logistic scale of Eq. 1; externally, abilities and difficulties are presented on a familiar Elo-style rating scale [Elo, 1978, Pelánek, 2016], an affine choice that changes nothing about the measurement and everything about interpretability.

Invariants. Ability estimates are monotone in evidence in the following sense: no single response, however extreme, can move an estimate past the point that the totality of evidence supports; and the influence of any fixed response decays as later evidence accumulates.

4.2 Clinical-safety-aware scoring

Problem. Score polytomous nursing items so that the measurement system sees not only *whether* a response was right but *how it was wrong* — while preserving the integrity of item calibration.

Why it is hard. In clinical education, distractors are not exchangeable. One wrong option represents a scope-of-practice or failure-to-rescue event; another represents a defensible near-miss that a strong candidate might select for subtle reasons. Dichotomous scoring erases this; naive severity-weighted scoring creates two new failure modes: (i) *double punishment*, where a dangerous error is charged against both the learner’s knowledge estimate and their safety profile until the two conflate; and (ii) *ruler corruption*, where safety adjustments leak into item-difficulty calibration, so that an item acquires a difficulty that reflects how frightening its distractors are rather than how much knowledge it demands.

Our approach (properties). Every distractor in the corpus is annotated, at authoring time, on a multi-level ordinal severity taxonomy that spans the range from catastrophic error (scope or legal breach, failure to rescue) down to near-miss cognitive distortions; a fully correct response occupies the top of the scale by construction — the absence of error is the best possible error profile, a seemingly small anchoring decision whose measurable consequences we discuss in §5. Scoring then obeys four rules:

1. **Commission dominates.** A response’s severity is that of the *most dangerous* wrong option it selected: choosing one lethal distractor is not excused by also choosing three harmless ones.
2. **Omission is the mildest failure.** Missing a key without selecting anything wrong sits at the near-miss end of the scale: nothing unsafe was chosen.
3. **The learner channel blends; the item channel does not.** The outcome that drives the *learner* update tempers correctness with a bounded, monotone function of error severity, so

a catastrophic error and a near-miss are distinguishable events to the learner model. Item difficulty, by contrast, calibrates on pure correctness alone. Difficulty is a property of an item’s knowledge demand; the safety blend moves the learner, never the ruler.

4. **Severity feeds pedagogy, not just measurement.** The severity history of a learner’s errors — recency-weighted, so an old lapse fades — prioritizes remediation: between two equally weak areas, the one whose errors are dangerous outranks the one whose errors are benign, and remediated safety-critical material must be re-confirmed under stricter criteria than anything else in the system.

Invariant. The severity of an error a learner actually made is never excused, discounted, or re-weighted by any property of the item that invited it. Contextual quantities (§4.4) adjust *expectations*, never the record of what happened.

4.3 Adaptive selection and measurement-precision stopping

Problem. Decide, at every step, which item a learner encounters next — under two regimes with different objectives (Table 1) — and, in the assessment regime, decide when to stop.

Practice regime. The objective is learning, and the difficulty policy follows the desirable-difficulty and ZPD literature [Bjork, 1994, Vygotsky, 1978]: the policy selects items so that predicted success lands in a motivating band above chance — hard enough to carry information and force retrieval [Roediger and Karpicke, 2006], never so hard that practice teaches discouragement. The target within that band is deliberately stochastic, so that the served sequence does not collapse into a deterministic rut; and selection is subject to recency and exposure constraints so that no item is milled into familiarity. Beyond difficulty, the practice outer loop is mastery-referenced [Bloom, 1984]: it balances *exploiting* the frontier where learning is fastest, *exploring* cells of the model with little evidence, and *maintaining* previously mastered cells whose confidence has gone stale — the last implementing spaced re-checks in the spirit of the distributed-practice literature [Cepeda et al., 2006].

Assessment regime. The objective is measurement, and the design follows classical CAT [Wainer, 2000, van der Linden and Glas, 2010]: the policy targets maximal Fisher information (Eq. 2), i.e., items whose outcome is least predictable for this learner; a running scalar readiness ability is maintained with its standard error (Eq. 3); and the session terminates the moment the confidence interval around the estimate lies entirely on one side of the passing standard, bounded below by a minimum length so that a lucky streak cannot end a test early, and above by a maximum length so that an undecided candidate still receives a decision. Selection is additionally nudged toward blueprint categories that are under-represented relative to their target share, so that maximal information never degenerates into drilling one seam [Chang and Ying, 1999] — diversity shapes the test without ever blocking it. Case-based, multi-step items are interleaved for their formative value but deliberately excluded from the ability estimate: they are not calibrated on the standalone scale, and mixing scales would corrupt the measurement [Kolen and Brennan, 2014].

Invariant. The two regimes never blur: practice difficulty policy never leaks into assessment selection, and assessment sessions are scored under the same response rules as practice, so “correct” means one thing everywhere in the system.

4.4 Session-level equating of the safety signal

Problem. Make the safety profile of a session comparable across sessions, when sessions are nonequivalent forms: one session’s items may offer mostly catastrophic wrong options while another’s offer mostly benign ones. A raw “mean severity of errors made” punishes the same clinical judgment more harshly on the first form than the second.

Approach (properties). This is a test-equating problem [Kolen and Brennan, 2014], and we solve it in the spirit of equating: each completed session is characterized by the severity profile of the errors it *made available* — an expected-error baseline for a hypothetical average miss on that form — and the learner’s realized error profile is read *against* that baseline, as an alignment ratio centered at unity. A reading near one says the learner’s errors were about as safe as the form allowed on average; readings on either side say their errors gravitated toward the more dangerous or the safer end of what was offered. The same session also reports the raw severity profile on a fixed absolute scale, so that the equated (relative) and absolute readings can be compared, and degenerate forms (e.g., sessions offering no error surface at all) are handled by construction rather than by exception. Consistent with the invariant of §4.2, equating adjusts the *expectation* against which a learner is read — never the severity of any error they actually made. We considered and rejected a composite that would multiply correctness by the alignment ratio: it conflates two channels that report different things, and we surface both channels separately instead.

4.5 Fair peer duels

Problem. Let friends compete head-to-head on nursing items in a way that is (i) fair by construction rather than by apology, (ii) impossible to farm by memorizing the bank, and (iii) motivationally sound — speed should matter, but guessing fast must lose.

Approach (properties). Three design commitments define the duel:

- **Fairness by construction.** The policy composes the challenge set *before* the first item is served, from a snapshot of both players’ abilities frozen at kickoff, and drawn exclusively from items *neither* player has ever answered — the entirety of the anti-memorization control: grinding the bank buys familiarity only with items the match will never serve. The set is stratified so that each player receives an equal allocation of their own “home turf” (items where they hold a clear predicted advantage), alongside symmetric strata where the players are closely matched — including the shared frontier where both are likely to fail, which is where a match is genuinely decided — and a coverage stratum that probes territory neither player’s history has reached. Predicted advantages are computed in the paired-comparison tradition [Elo, 1978, Glickman, 1999]; because both players face common item difficulties, difficulty cancels in the comparison and can never bias who is favored.
- **Speed scored symmetrically.** Per-response scoring weights correctness by response speed in a signed fashion, in the spirit of the speed-accuracy scoring-rule literature [Maris and van der Maas, 2012]: answering quickly amplifies a correct answer and a wrong answer *equally*, and an answer at the buzzer is worth little either way. Fast guessing is thereby a strictly losing strategy, and slow certainty is a viable one. A match reports both the speed-weighted total that decides it and raw accuracy beside it, because the two can disagree and hiding the disagreement is what makes a verdict feel arbitrary.
- **A separate competitive ladder.** Duels move a dedicated competitive rating, kept deliberately separate from the ability model — the separation that blunts sandbagging — and updated

in the rating-with-uncertainty tradition [Glickman, 1999]: margins are graded rather than binarized, normalized across match lengths, and newcomers’ ratings move faster than veterans’. Competition can still *teach*: in a designated practice mode, the match log is afterwards replayed through the ordinary measurement channel, under exactly the same scoring rules as solo practice. Clinical severity (§4.2) is surfaced in the match report when touched, but never decides the winner — a rule that let safety override the scoreboard would make the scoreboard unreadable without teaching the lesson, which the post-match reveal does.

4.6 Curiosity-adaptive personalization

Problem. A learner who spends an evening reading Deep Review lectures on a topic is expressing epistemic curiosity — a powerful, well-documented lever on engagement and learning [Loewenstein, 1994]. Can the agent respond to that signal by steering some practice toward the learner’s demonstrated interests, without (i) contaminating ability measurement, (ii) letting the mechanism be gamed, or (iii) letting interest crowd out the blueprint?

Approach (properties). The agent derives from library telemetry — lecture reads and searches — a bounded, recency-decayed *interest intensity*, and the batch-composition policy reserves a correspondingly sized minority slice of each practice batch for items connected to what the learner has been reading. The design guarantees are the substance:

- **Selection-only, never scoring.** The interest signal influences *which* items are chosen; it never touches ability or difficulty estimation. Reading is evidence of curiosity, never of competence.
- **Bounded by construction.** The interest slice is confined to a narrow band between a small floor and a modest ceiling; the mapping from telemetry to slice size is saturating and smooth, so no volume of reading can push personalization past its ceiling, and the remainder of every batch is composed by the ordinary mastery-referenced policy at its usual difficulty targeting.
- **Gaming-resistant at the definition.** Repeated reads of the same lecture within a day are deduplicated at the source; the intensity map saturates; and an anti-avoidance gate ties the personalization increment to recent *answering* volume, so a pure reader who has stopped practicing receives no enlarged slice — reading cannot substitute for retrieval practice [Roediger and Karpicke, 2006].
- **Legible relevance.** Search-driven interest is matched to items through keyword salience weighted by standard inverse document frequency [Spärck Jones, 1972],

$$\text{idf}(t) = \log \frac{N}{n_t}, \quad (5)$$

where N is the corpus size and n_t the number of items bearing term t — so that distinctive clinical terms, not ubiquitous ones, drive the match, and every personalized item can display the reason it was chosen.

- **Fails closed.** If any telemetry input is unavailable, batch composition degrades byte-for-byte to the non-personalized policy.

4.7 Deep Review: pedagogy against item-fitting

Problem. Question banks invite a quiet failure mode we call *item-fitting*: the learner converges on the mapping from item surface features to correct options rather than on the clinical model that

Ablation question	Finding (direction)	Decision
Should difficulty baselines weight distractors by authored plausibility?	Orderings essentially unchanged; virtually no session-level readouts moved	Keep the simpler estimator
Should they weight by <i>empirical</i> selection behavior?	Materially different on a meaningful minority of items; current behavioral data too sparse to trust	Defer; pre-registered revisit once live data de-noises
Does mis-anchoring the correct response on the severity scale matter?	Yes: systematically depressed ability trajectories and readiness decisions in simulation	Anchor corrected; correct = top of scale
Does the curiosity slice violate its guarantees?	No: bounds, non-interference, gating, determinism, and fail-closed behavior all held under structural test	Ship floor conservatively; ceiling gated on live evidence

Table 2: The internal ablation program, summarized qualitatively.

generates the correct option. Item-fitting inflates within-bank performance while leaving transfer to novel presentations — which is what both the licensure examination and clinical practice demand — unimproved [Barnett and Ceci, 2002].

Approach. Deep Review attacks the problem structurally:

1. **Concept-level treatment, item-level anchor.** Each lecture is a textbook-depth treatment of a whole topic — the mechanism, the connected concepts around it, and the specific traps examinations set when testing it — linked from the items that instantiate it. The learner who understands why potassium does what it does to a cardiac cell can answer a hyperkalemia item they have never seen, phrased in a way no rationale prepared them for.
2. **Post-answer availability.** Lectures open only after the item is answered, preserving the retrieval event that drives durable learning [Roediger and Karpicke, 2006] — the lecture elaborates a memory trace that the attempt has just reactivated, rather than pre-empting the attempt.
3. **A browsable, searchable library.** Between sessions the library is an independent study surface, and the curiosity it elicits is itself a signal the agent can respond to (§4.6) — closing a loop between self-directed reading and directed practice.
4. **Alignment with spaced re-encounter.** Because the agent schedules spaced re-checks of mastered material (§4.3), concepts revisited through the library are also re-tested at spacings the distributed-practice literature supports [Cepeda et al., 2006].

5 Ablations and Empirical Analysis

Our evaluation discipline is conservative by policy: no change to the measurement machinery ships without an ablation that isolates its effect, and plausible-sounding refinements are rejected when the evidence says they do not pay for their complexity. We summarize three studies at the level of their questions and findings; consistent with §1.1, we report directions and decisions rather than the underlying parameterizations (Table 2).

5.1 Plausibility-weighted difficulty: a refinement declined

The severity taxonomy of §4.2 induces, for every item, an expected-error baseline used by session equating (§4.4). A natural refinement holds that this baseline should weight each distractor by how *tempting* it is, not merely average over the distractor field: a devastatingly plausible near-miss and a throwaway near-miss are not the same opportunity for error. The distinction between an option’s danger and its plausibility is psychometrically sound — distractor functioning is a recognized driver of item difficulty — so we tested it, over the full live corpus, in two forms: weighting by the *authored* plausibility annotation each distractor already carries, and weighting by the *empirical* frequency with which learners actually select each distractor.

The authored variant failed to justify itself: it reordered item difficulty baselines nearly identically to the flat estimator, moved almost no item meaningfully, and left session-level equated readouts effectively unchanged — the authored annotation is too coarse, and too correlated with severity itself, to carry independent information. The empirical variant was a different story: it disagreed with the flat estimator on a meaningful minority of items, in exactly the way the hypothesis predicts (items whose dangerous options are rarely chosen are, empirically, safer forms than they appear on paper). But at our current scale the behavioral selection data is sparse enough that a substantial part of that disagreement is noise. The decision was therefore *measure twice*: the flat estimator ships; the empirical variant is deferred behind a pre-registered re-evaluation that triggers once live selection data accumulates; and the authored annotation is retained as a cold-start prior for that future instrument. An appealing idea was thereby declined *twice* — once for lack of effect, once for lack of data — which is precisely the discipline the ablation program exists to enforce.

5.2 Anchoring the correct response on the severity scale

The severity taxonomy assigns every *response* a position on the ordinal scale, and a fully correct response must sit at its top — no error at all is the best possible error profile. An earlier variant of the scoring channel anchored correct responses below the top of the scale, which is subtly but systematically wrong: it treats every correct answer as carrying a residue of error, and because the learner channel blends correctness with error quality (§4.2), the mis-anchor acts as a persistent drag applied to correct responses only. Simulation with the production update path, run identically under both anchors across learners of varying ability and a range of passing standards, showed the expected signature: ability trajectories under the mis-anchor rise more slowly and asymptote lower for otherwise identical response streams, and readiness decisions under the precision-stopping rule of §4.3 shift measurably toward failure across the whole range of standards tested. The effect compounds with evidence — the more consistently correct the learner, the larger the accumulated understatement. The anchor was corrected, and the episode is retained in our test suite as a regression: it is a clean example of a one-line semantic choice with a measurable psychometric consequence, exactly the class of defect that only end-to-end ablation catches.

5.3 Guarantee testing for curiosity-adaptive personalization

Because the curiosity channel (§4.6) is a deliberate intervention in selection, we test its *guarantees* rather than its hoped-for benefits: that the personalized slice respects its floor and ceiling for every profile; that batch difficulty targeting is statistically indistinguishable with the channel on and off; that a learner who only reads but does not practice receives no enlarged slice; that identical inputs produce identical batches; and that telemetry failure degrades exactly to the non-

personalized composition. All held under structural evaluation across seeded learner profiles. We are explicit about what has *not* been demonstrated: that a larger curiosity slice improves outcomes. The motivational literature predicts it [Loewenstein, 1994], but prediction is not evidence; the ceiling therefore ships at a conservative value, and raising it is gated behind a pre-registered live comparison relating realized exploration share to subsequent mastery trajectories and retention.

6 Discussion

6.1 What the tri-disciplinary framing buys

Each parent discipline supplies a discipline the others lack. Psychometrics contributes the measurement conscience: scales with standard errors, equating across nonequivalent forms [Kolen and Brennan, 2014], stopping rules with stated confidence [Wainer, 2000]. Adaptive-learning research contributes the dynamics: learners change, forget, and differ, and the system must be built for trajectories rather than snapshots [Corbett and Anderson, 1994, Klinkenberg et al., 2011]. Nursing education contributes the values: the primacy of clinical judgment [Dickison et al., 2019], the non-exchangeability of errors, and the insistence that acting safely is a distinct competency from knowing. The methodological contributions of this report — severity-aware dual-channel scoring, safety-equated sessions, measurement-preserving personalization, fairness-by-construction duels — each exist precisely at a pairwise seam of these disciplines, which is our answer to why an adequate system could not be assembled from any one of them alone.

6.2 The agent stance

We describe Dora as an agent advisedly — and, per the equivalence drawn in §3, we could everywhere have written *policy*, since neither word constrains the substrate that computes the state-action mapping both of them name. The system’s intelligence is expressed in its *decisions* — what to serve, when to stop, what to remediate, when to say “ready” — and in the invariants those decisions provably respect, rather than in any particular computational substrate. We are deliberately silent in this report on Dora’s internal mechanism, and we caution against inferring it from the method families we cite: the cited literature locates the problems, not our solutions. What we do assert is behavioral: every decision the agent makes is deterministic given its state (or reproducibly stochastic where stochasticity is a design goal), every number it reports has a definition a psychometrician can audit, and no decision path depends on quantities a learner cannot in principle be shown.

6.3 Honest positioning

Cytross Nursing is a young system, and this report is written at the beginning of its empirical life, not the end. Our simulation and structural evidence (§5) supports the internal coherence of the measurement machinery; it does not yet constitute evidence of superior licensure outcomes, and we make no such claim here. The field’s history — where “guaranteed pass” marketing is everywhere and evidence rarely follows — is exactly why we consider the discipline of this section a feature of the methodology rather than a disclaimer appended to it.

7 Limitations and Ethical Considerations

Constants are engineering judgment pending data. The thresholds, bands, decay horizons, and ceilings that instantiate the designs of §4 were chosen by analysis and worked-example calibration, not fit to outcome data. Where a choice is consequential and testable, it is explicitly gated behind pre-registered live evaluation (§5); until those gates open, the honest description of every such constant is *defensible, not demonstrated*.

No outcome claims. We claim internal validity (the system measures what it defines, stably and reproducibly), not external efficacy (that preparation with the system improves examination or clinical outcomes). External efficacy is an empirical question we intend to answer with appropriate designs, and nothing in this report should be read as pre-empting it.

The severity taxonomy embeds judgment. Distractor severity is authored by clinical educators applying a rubric. It inherits their perspective, and miscalibrated annotations propagate into remediation priorities (though never into correctness, and never into item difficulty — the channel separation of §4.2 bounds the blast radius). Auditing severity annotations against empirical response behavior is an ongoing obligation, not a solved problem.

Formative, not predictive, safety signals. The safety profile (§4.2–4.4) is a formative pedagogical signal. It is not a validated predictor of clinical incident risk, and it must not be used for gatekeeping, employment, or academic sanction. We surface it to the learner it describes, in service of their own remediation.

Privacy. Personalization consumes behavioral telemetry (reading, searching, answering). All such telemetry is held pseudonymously, is used under the selection-only invariant of §4.6, is visible to the learner it belongs to, and is deletable by them. A learner who wants no curiosity adaptation receives, by construction, the standard composition.

Fairness and access. Adaptive systems can entrench early disadvantage if cold-start estimates harden. Dora’s uncertainty-first design is a partial mitigation — early estimates are held loosely by construction and pooling never penalizes below the evidence — but systematic fairness auditing across demographic and educational backgrounds remains future work we regard as owed, not optional.

Competitive integrity. Duels are engineered to be fair and farm-resistant (§4.5), but any competitive surface can shape behavior in unintended ways. Our design already refuses to let safety errors be laundered by a winning scoreboard (severity is surfaced in the report), and we monitor for strategies that would trade clinical prudence for speed.

8 Conclusion

We have presented the methodology of Cytross Nursing: an intelligent tutoring system whose core policy — the adaptive learning agent Dora — unifies continuous latent-ability estimation, clinically safety-aware scoring, dual-regime adaptive selection with precision-based stopping, session-level safety equating, fair peer competition, and curiosity-adaptive personalization over a single multidimensional learner model — and whose Deep Review library attacks item-fitting at its pedagogical root. The report’s contribution is the problem formulation and the invariants: the places where nursing education, adaptive learning, and psychometrics constrain one another, and the design commitments a system must make to honor all three at once. Our ablation discipline — change nothing the evidence does not justify, and pre-register the evidence that would — is, we believe, as much a part of the methodology as any estimator. The empirical program this report opens — live evaluation of the personalization ceiling, data-gated refinement of the equating

baseline, and outcome studies worthy of the claims we have declined to make — is where this work continues.

Correspondence: Cytross Nursing Team. This report describes methodology at a level of abstraction chosen to be verifiable in its claims and non-reproducible in its specifics; requests for collaboration or audit access are welcome.

References

- John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207, 1995. doi: 10.1207/s15327809jls0402_2.
- Susan M. Barnett and Stephen J. Ceci. When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637, 2002. doi: 10.1037/0033-2909.128.4.612.
- Joe Betts, William Muntean, Doyoung Kim, Natalie Jorion, and Philip Dickison. Building a method for writing clinical judgment items for entry-level nursing exams. *Journal of Applied Testing Technology*, 20(S2):21–36, 2019.
- Allan Birnbaum. Some latent trait models and their use in inferring an examinee’s ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Addison-Wesley, Reading, MA, 1968.
- Robert A. Bjork. Memory and metamemory considerations in the training of human beings. In Janet Metcalfe and Arthur P. Shimamura, editors, *Metacognition: Knowing about Knowing*, pages 185–205. MIT Press, Cambridge, MA, 1994.
- Benjamin S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984. doi: 10.3102/0013189X013006004.
- Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354–380, 2006. doi: 10.1037/0033-2909.132.3.354.
- Hua-Hua Chang and Zhiliang Ying. a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3):211–222, 1999. doi: 10.1177/01466219922031338.
- Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994. doi: 10.1007/BF01099821.
- Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325, 2023. doi: 10.1109/ICALT58122.2023.00100.
- Philip Dickison, Katie A. Haerling, and Kathie Lasater. Integrating the National Council of State Boards of Nursing Clinical Judgment Model into nursing educational frameworks. *Journal of Nursing Education*, 58(2):72–78, 2019. doi: 10.3928/01484834-20190122-03.

- Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, 1978.
- Mark E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999. doi: 10.1111/1467-9876.00159.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eva Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oliver Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. doi: 10.1016/j.lindif.2023.102274.
- Sharon Klinkenberg, Marthe Straatemeier, and Han L. J. van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011. doi: 10.1016/j.compedu.2011.02.003.
- Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking: Methods and Practices*. Springer, New York, 3rd edition, 2014. doi: 10.1007/978-1-4939-0317-7.
- George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75–98, 1994. doi: 10.1037/0033-2909.116.1.75.
- Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, 2023. doi: 10.18653/v1/2023.findings-emnlp.372.
- Gunter Maris and Han van der Maas. Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4):615–633, 2012. doi: 10.1007/s11336-012-9288-y.
- Radek Pelánek. Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179, 2016. doi: 10.1016/j.compedu.2016.03.017.
- Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 505–513, 2015.
- Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960. Expanded edition reprinted 1980, University of Chicago Press.
- Henry L. Roediger and Jeffrey D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006. doi: 10.1111/j.1467-9280.2006.01693.x.

- Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2 (Monograph No. 17)), 1969.
- Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.
- Anaïs Tack and Chris Piech. The AI Teacher Test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, 2022. arXiv:2205.07540.
- Wim J. van der Linden and Cees A. W. Glas, editors. *Elements of Adaptive Testing*. Springer, New York, 2010. doi: 10.1007/978-0-387-85461-8.
- Kurt VanLehn. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265, 2006.
- Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011. doi: 10.1080/00461520.2011.611369.
- Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.
- Howard Wainer, editor. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2000.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large language models for education: A survey and outlook, 2024. arXiv preprint arXiv:2403.18105.
- Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024. doi: 10.1111/bjet.13370.