Conceptual Regression Depth (CReD): A Framework for Psychometric-Integrated Tutoring Systems that Preserve Critical Thinking

Syed A. Hadi
SyedH@workmail.com

Abstract

Current Intelligent Tutoring Systems (ITS) face a critical paradox: while they can improve immediate performance, they often promote cognitive off-loading that fosters dependency rather than independent learning. This paper introduces Conceptual Regression Depth (CReD), a framework that leverages personalized and statistically validated prerequisite learning paths to guide remediation as an alternative to cognitive off-loading. CReD functions by parsing educator supplied instructional content into discrete concept units broken down by Bloom's taxonomy cognitive levels representing nodes in a directed acyclic graph. The resulting knowledge tree guides learning, which is pedagogically aligned with the classroom, through the ITS. As students interact with the ITS, proficiency for each node is calculated through mastery probability estimation, which is represented as a temporal sparse matrix. This data is suitable for Item Response Theory (IRT) modeling which makes detailed psychometric analysis possible, offering several benefits; teachers obtain student level diagnostics, learning and remediation pathways can be validated, and educators can effectively design targeted intervention. CReD contributes a systematic method for bridging conversational AI with well-established psychometric methods, extracting insights that facilitate the responsible integration of generative AI in education.

1. Introduction and Background

1.1 The Promise and Peril of Intelligent Tutoring Systems

Current Intelligent Tutoring Systems (ITS) excel at content personalization but require careful guardrails when applied in educational contexts. Research has raised significant concerns regarding the long-term impact of over reliance on these systems, revealing a troubling pattern: while AI tutoring can improve immediate performance, it may undermine the very skills education seeks to develop.

A recent study by Michael Gerlich at SBS Swiss Business School demonstrated that increased reliance on artificial intelligence (AI) was associated with the erosion of critical thinking skills, largely due to cognitive off-loading, a process in which individuals reduce their mental effort by depending on AI tools (Gerlich, 2025). This finding was reinforced by controlled research at the University of Pennsylvania, which found that students performed better while using an AI tutor, yet once the tutor was removed, those same students performed worse compared with peers who had never relied on an AI system (Bastani et al., 2024). These findings suggest that students may come to use AI as a crutch, thereby missing opportunities to engage in independent thinking and problem-solving.

1.2 Limitations of Current Approaches

The effective implementation of ITS in education requires frameworks that not only support learning but also promote critical thinking and align with classroom pedagogy (Favero et al., 2025). However, existing ITS approaches reveal important limitations across multiple dimensions. Systems that assess student knowledge through binary mastery classifications,

indicating only whether a concept is understood or not, may provide surface-level diagnostics and immediate remediation, but they do not offer systematic guidance on the depth of remediation needed (Ostrow et al., 2015). Research has shown that generative AI systems face systematic challenges in logical reasoning, limiting their capacity to diagnose foundational learning deficiencies without a structured instructional framework (Mirzadeh et al., 2025).

Framework-based AI tutors that impose externally designed taxonomic structures risk misalignment with classroom pedagogies (Díaz & Nussbaum, 2024). Additionally, external taxonomic structures may risk *Simplicity Bias* caused by large language models (LLM) through more complex terminology than necessary (Kuribayashi et al., 2024). Importantly, current educational technologies more broadly lack integration with robust measurement and psychometric modeling (Chang, 2024).

1.3 The Conceptual Regression Depth Framework

To address these challenges, this paper introduces Conceptual Regression Depth (CReD), a framework that quantifies prerequisite learning distances while simultaneously generating psychometric profiles from conversational learning data. This work was influenced by Hu (2024) who stated:

"While the journey towards intelligent adaptive learning systems is complex and uncharted, the destination is one worth pursuing. By thoughtfully leveraging AI to enhance psychometric assessments and personalized support, we have the potential to revolutionize education, enabling every student - regardless of their learning difficulties - to thrive and reach their full potential. It

is a grand challenge, but one that promises profound benefits for individual learners, educational equity and society at large."

Unlike traditional applications of psychometric approaches that rely on formal assessments, CReD-integrated ITS transforms patterns of student interaction, including mistakes, questions, and problem-solving attempts, into Item Response Theory parameters that capture individual ability estimates and concept difficulty calibrations. The proposed system draws topic progression, content guidelines and item banks from teacher-supplied artifacts that align with a student's classroom learning environment.

This approach provides teachers with precise diagnostic information about student progress and enables evidence-based refinement of learning trajectories and remediation pathways. In doing so, the framework strengthens foundational concepts by providing attentive remediation and promoting critical thinking through the ITS, rather than serving as a crutch.

1.4 Addressing Educational Workforce Challenges

Beyond technological considerations, the framework also responds to pressing challenges in the educational workforce. Public schools in the United States entered the 2024–25 academic year with an average of six teaching vacancies, and seventy-four percent reported difficulty filling one or more of those positions with fully certified teachers (NCES, 2024). A review conducted by the Learning Policy Institute in July 2025 reported that one in eight teaching positions was either unfilled or staffed by under-qualified teachers (LPI, 2025). Simultaneously, educational outcomes reveal urgent needs for individualized support. The Nation's Report Card in 2024 revealed that fewer than one-third of students nationwide were performing at the National

Assessment of Educational Progress (NAEP) Proficient level in reading at grades four and twelve (NAGB, 2025).

By introducing a human-in-the-loop design, the CReD framework augments rather than replaces teachers, offering guided support that mirrors classroom pedagogy while equipping educators with diagnostics that would otherwise require dramatically smaller class sizes. Our framework extends the reach of quality instruction beyond traditional classroom constraints while preserving the central role of educators in the learning process.

2. Methodology

2.1 System Overview

We propose a computational framework for Conceptual Regression Depth (CReD), a system that leverages prerequisite relationships in a structured concept hierarchy and external knowledge deficiencies that contribute to learning gaps to guide remediation. Our approach constructs the concept graph and progression rules directly from teacher-supplied curricular artifacts, including textbook chapters, lesson plans, and instructional notes. By grounding the prerequisite structure in instructional materials familiar to the learner, the system ensures alignment with classroom pedagogy and facilitates contextually relevant remediation.

The framework operates through six sequential and interdependent stages that transform curricular content into diagnostic and psychometric profiles. Beginning with automated parsing and semantic structuring of teacher-uploaded instructional content into discrete concept units, the system extracts inter-concept dependency relations and represents them as a directed acyclic graph (DAG) to preserve logical learning order. The system identifies and categorizes critical

external knowledge requirements that fall outside the primary learning sequence but remain essential for concept mastery, followed by systematic decomposition of each concept node into hierarchically ordered sub-nodes corresponding to Bloom's revised taxonomy cognitive levels (Sudirtha et al., 2022). The system subsequently computes the primary CReD score as the shortest path distance from failed nodes to identified root gaps within the prerequisite hierarchy, while maintaining a categorical inventory of external touchpoints across predefined conceptual domains. Finally, the framework transforms interaction patterns into sparse matrices compatible with Item Response Theory (IRT) models, enabling standardized ability estimation, concept difficulty calibration, and at-risk student identification through established psychometric methods.

2.2 Curricular Resource Ingestion

The CReD framework begins by transforming teacher-provided instructional resources into a machine-readable representation of the course's conceptual structure. These resources may include digitized textbooks, lesson plans, and structured syllabus outlines. Textual content is first normalized and converted to a uniform internal markup format to preserve headings, lists, and table structures.

Concept extraction proceeds via hierarchical text segmentation. Section and subsection headings are parsed to identify candidate concept boundaries, supplemented by semantic paragraph clustering using transformer-based sentence embedding models to group content describing the same instructional objective. This process yields discrete concept units, each corresponding to a node in the eventual prerequisite graph.

2.3 Prerequisite Graph Construction

Following concept extraction, the identified concept units are organized into a directed acyclic graph where each vertex corresponds to a distinct instructional concept and each directed edge denotes that one concept must be mastered before another can be effectively learned, forming a prerequisite graph (ACE, 2024).

Prerequisite relationships are inferred through multiple complementary mechanisms. First, explicit ordering cues present in the instructional materials, such as textbook chapter sequencing and syllabus topic order, are converted into directed edges under the assumption that earlier-presented concepts serve as potential prerequisites for later ones. Second, semantic analysis of prerequisite indicators within the text detects linguistic cues such as "requires understanding of," "builds upon," and "before attempting" to identify/validate dependency relationships.

Additional sources may include concept maps derived from educational standards, established taxonomies within specific domains, or combinations of these approaches. Different subjects require different techniques for prerequisite detection, and the framework accommodates various methodologies to classify concept relationships as prerequisite, co-requisite, or independent based on domain-specific characteristics and available resources.

To ensure topological validity, the resulting graph undergoes automated cycle detection with any identified cycles flagged for resolution to preserve acyclicity (Bender et al., 2015). A hierarchical compression step prevents over-fragmentation by aggregating concept units sharing identical prerequisite sets and overlapping instructional objectives into composite nodes (Zhang et al.,

2025). This reduces graph density and improves interpretability for subsequent Bloom's-level decomposition and CReD computation.

Teachers retain oversight through a streamlined review interface that presents the generated prerequisite map for validation and adjustment. Rather than requiring extensive manual curation, the system highlights automatically detected relationships that may warrant attention, such as potential cycles or unexpected dependencies. This human-in-the-loop validation ensures pedagogical accuracy while leveraging automated processing to reduce teacher workload (ACE, 2024). As students interact with the system over time, performance data through psychometric analysis provides additional validation of prerequisite relationships, creating a feedback mechanism that can refine the conceptual structure without requiring ongoing manual intervention (Xu & Mostow, 2013).

2.4 Bloom's Taxonomy Integration

To capture intra-concept cognitive progression, each concept node in the prerequisite graph is decomposed into six sequential sub-nodes corresponding to Bloom's revised taxonomy: Remember, Understand, Apply, Analyze, Evaluate, and Create (Anderson & Krathwohl, 2001). These sub-nodes are connected by directed vertical edges, forming a hierarchical chain within the same concept, where mastery of level L_k typically serves as a prerequisite for attempting L_{k+1}. No horizontal edges are created between Bloom levels of different concepts; inter-concept prerequisites exist only at the parent concept node level.

The decomposition process uses automated classification to assign learning statements and associated activities to appropriate Bloom levels (Banujan et al., 2023). This is followed by

teacher verification to ensure pedagogical accuracy. Li et al. (2022) demonstrated automated decomposition of learning material into Bloom levels with a human-in-the-loop annotation/ validation mechanism. This hierarchical structure enables the CReD computation to distinguish between different cognitive depths of understanding within individual concepts while maintaining clear prerequisite relationships across the broader conceptual framework.

2.5 Mastery Probability Estimation

Mastery probabilities $p_{s,c,L}$ for student s, concept c, and Bloom cognitive level L are dynamically estimated using in-system assessments derived from the ingested curricular materials. The system operates with dual probability tracking: internal mastery probabilities for concepts within the structured curriculum, and external mastery probabilities for knowledge requirements that fall outside the primary learning sequence.

2.5.1 External Knowledge Detection and Evaluation

Majnik et al. (2013) has proposed a systematic approach to knowledge gap detection in automated learning systems. The system monitors student interactions for external knowledge gaps through incorrect responses suggesting missing foundational skills, student queries requesting clarification on assumed knowledge, and semantic analysis revealing conceptual blind spots. When detected, the system creates external nodes representing these knowledge requirements and employs a criticality evaluation mechanism to distinguish between consequential and non-consequential gaps (Schmidt, 2020). External nodes are classified as critical if they represent fundamental cognitive or academic skills necessary for concept mastery (e.g., basic arithmetic operations, reading comprehension), versus contextual nodes involving

domain-specific cultural knowledge that can be bypassed without compromising learning objectives (e.g., unfamiliarity with specific sports terminology used in a word problem). Only critical external nodes undergo formal mastery probability estimation through targeted micro-assessments to evaluate $p_{s,e,L}$ for student s, external node e, and cognitive level L.

2.5.2 Mastery Probability Calculation

During interaction, the system evaluates mastery of preceding lower levels within the same concept while assessing any critical external dependencies. Mastery probability $p_{s,c,L}$ is calculated using Bayesian estimation with flexible priors that adapt to the available evidence:

$$p_{s,c,L} = (\text{correct responses} + \alpha) / (\text{total responses} + \alpha + \beta)$$

Where the prior parameters adjust based on question availability: for single-question assessments, $\alpha=0.5$ and $\beta=0.5$ provide weak priors that allow higher probability estimates; for two-question assessments, $\alpha=1$ and $\beta=1$ establish uniform priors; and for three or more questions, $\alpha=2$ and $\beta=1$ create slightly optimistic priors that allow the data to dominate the estimate. Similar to the way Sapountzi et al. (2021) demonstrated adaptive Bayesian updating for monitoring learner knowledge states, this dynamic prior adjustment balances evidence with uncertainty and supports more reliable mastery estimation across varying assessment conditions. Students are classified as demonstrating mastery when their estimated probability exceeds 0.75 for internal concept nodes and 0.70 for critical external nodes.

2.6 CReD Computation

The system computes a Conceptual Regression Depth (CReD) score whenever a student fails an assessment at a given concept-Bloom level pair (c,L). The computation operates through two parallel processes: internal prerequisite path analysis for the primary CReD score, and external touchpoint identification for supplementary profiling.

2.6.1 Internal Prerequisite Analysis

The prerequisite path is analyzed through a CReD score computation which measures the depth of remediation required. While this score does not directly support student analytics, it evaluates knowledge states of interconnected concepts to diagnose learning gaps. The CReD computation begins with a backtracking search through the prerequisite graph. The search first moves vertically within the same concept, traversing down the Bloom hierarchy from level L toward "Remember." If mastery gaps are detected at lower levels within the same concept, the traversal continues horizontally to prerequisite concepts identified in the DAG.

At each visited node (c′,L′), the system evaluates the mastery probability $p_{s,c',L'}$ against the established thresholds (0.75 for internal nodes, 0.70 for critical external nodes). Contextual cues can allow skipping nodes. The stopping condition is met when the first node with mastery probability below threshold is found. This node is designated as the internal root gap, representing the earliest cognitive or conceptual deficiency within the prerequisite structure that explains the observed failure at (c,L).

The primary CReD score is defined as the length of the shortest path from the failed node (c,L) to the internal root gap (c′,L′). Liang et al. (2015) has demonstrated the success of a single metric *reference distance* approach in measuring prerequisite relationships.

2.6.2 External Touchpoint Analysis

Simultaneously, the system identifies critical external knowledge deficiencies that may contribute to the failure at (c,L). External touchpoints are categorized into predefined high-level conceptual domains established during system configuration, such as Arithmetic (Addition, Subtraction), Reading Comprehension (Inference, Vocabulary), or Logical Reasoning (Conditional Statements, Pattern Recognition). Schmidt (2020) demonstrates broad classification of knowledge gaps and their identification. Each identified external touchpoint is recorded with its categorical classification and assessed mastery level, providing a complementary diagnostic profile that informs intervention beyond the structured curriculum sequence.

2.6.3 The CReD Score

The system uses CReD measurements in three complementary dimensions. The internal distance dimension quantifies the number of prerequisite hops from the failure point to the identified root gap within the concept hierarchy. The cognitive complexity dimension specifies the Bloom level L′ of the internal root gap, indicating the cognitive complexity at which remediation should begin. The external profile provides a categorized inventory of external touchpoints with their respective mastery assessments, enabling comprehensive intervention planning that addresses curricular and foundational skill gaps.

2.7 Managing Item Sparsity

While the system predominantly leverages educator supplied material to extract evaluation items for each concept node and sub-nodes, we acknowledge that conversational learning requires repeated and non-redundant evaluation which may not always be adequately supplied by curricular artifacts. Russell-Lasalandra et al. (2025) demonstrate the success and utility of generative AI in developing items for underrepresented constructs. By integrating large language models with network psychometrics, their approach generates scales with structural validity comparable to traditional expert-developed measures (Russell-Lasalandra et al., 2025). Although this reduces reliance on manual intervention, our focus is on validating item–concept pairs and expanding the pool of evaluation resources available to educators and the ITS.

3. Psychometric Modeling via Sparse Matrix Construction

3.1 Rationale and Approach

Psychometric evaluations are commonly employed to recognize students' unique learning profiles and customize instructional strategies to meet their needs (Fletcher & Vaughn, 2009). The individual mastery probability calculations provide a foundation for psychometric analysis through Item Response Theory (IRT) modeling. By transforming individual student-concept interaction patterns into sparse matrices, we apply established psychometric methods to extract latent ability parameters and concept difficulty estimates.

Our approach leverages the 2-Parameter Logistic (2PL) IRT model, which balances analytical tractability with meaningful parameter interpretation. The 2PL model estimates both concept difficulty (β) and discrimination (α) parameters while providing student ability (θ) estimates,

making it well-suited for educational applications where both student proficiency and concept characteristics matter.

3.2 Sparse Matrix Construction

For each student s and concept-Bloom level pair (c,L), we construct a sparse interaction matrix M where rows represent students, columns represent concept-Bloom level pairs, and cell values M[s,(c,L)] encode interaction outcomes. The encoding scheme assigns values of 1 for mastery achieved (probability ≥ threshold), 0 for failure requiring regression (CReD > 0), and missing entries for unrecorded interactions.

To optimize matrix density, we implement temporal aggregation within rolling time windows, focus initially on core cognitive levels (Remember, Understand, Apply), and apply minimum interaction thresholds for reliable parameter estimation. Zhang et al. (2025) proposed a similar framework to address data sparsity in multidimensional learning performance datasets.

3.3 IRT Model Implementation

We implement the 2PL IRT model:

$$P\left(X_{s,(c,L)} = 1 \mid \theta_s, \alpha_{(c,L)}, \beta_{(c,L)}\right) = \frac{\exp\left(\alpha_{(c,L)} \cdot \left(\theta_s - \beta_{(c,L)}\right)\right)}{1 + \exp\left(\alpha_{(c,L)} \cdot \left(\theta_s - \beta_{(c,L)}\right)\right)}$$

Where:

Where θ_s represents student ability, $\beta_{(c,L)}$ represents concept-Bloom difficulty, and $\alpha_{(c,L)}$ represents concept-Bloom discrimination. Parameter estimation employs Marginal Maximum

Likelihood Estimation (MMLE) using E-M algorithms for sparse data and Expected A Posteriori (EAP) for individual student ability estimates (Dempster et al., 1977).

3.4 Educational Applications

3.4.1 Student Profiling and Risk Identification

The IRT model generates standardized ability estimates (θ) that indicate proficiency relative to peer groups. Longitudinal tracking of θ changes identifies patterns of learning acceleration or deceleration, enabling early identification of students requiring intervention. At-risk classification flags students with θ estimates below -1.0 for intensive support consideration.

3.4.2 Concept Difficulty Calibration

Difficulty parameters (β) enable empirical validation of prerequisite sequences by comparing statistically-derived difficulty estimates with in-system learning progressions and remediation pathways. Discrimination parameters (α) identify concepts with low values ($\alpha < 0.5$) that fail to effectively differentiate between student ability levels, indicating potential assessment or instructional design issues.

3.5 Implementation Validation

Psychometric data extracted from the system can be leveraged to validate prerequisite paths and candidate remediation routes, ensuring that the conceptual dependencies and suggested learning interventions are empirically supported. This provides critical reinforcement, enabling the system to adaptively self-refine. Prior work demonstrates the value of such an approach; Chen and

Chang (2018) highlight the success of psychometric methods in guiding learning trajectory development and topic recommendation, underscoring that data-driven psychometric integration into instructional design is well-established and effective. Teacher oversight remains essential as the final arbiter, ensuring alignment with pedagogical goals and contextual classroom needs.

4. Discussion

The CReD framework addresses a fundamental gap in educational technology by providing systematic quantification of learning remediation requirements. Our approach adaptively constructs remediation pathways by addressing knowledge gaps therefore promoting critical thinking instead of serving as a crutch. The integration with psychometric modeling through IRT creates a comprehensive diagnostic system that combines immediate actionable insights with longitudinal ability tracking.

The framework's reliance on teacher-provided curricular materials ensures pedagogical alignment while reducing implementation barriers compared to systems requiring extensive historical data collection. By constructing prerequisite graphs directly from instructional content, the system maintains fidelity to classroom learning progressions rather than imposing external taxonomic structures that may not reflect actual teaching practices.

The dual-metric approach distinguishes between internal prerequisite deficiencies within established learning sequences and external knowledge gaps that fall outside the primary curriculum. This distinction proves particularly valuable for identifying students whose difficulties stem from foundational skills rather than sequential concept mastery, enabling more targeted support allocation.

Several limitations warrant consideration. The framework's effectiveness depends on the quality of prerequisite relationship detection and teacher validation of concept dependencies. Complex domains with highly interconnected concepts may produce prerequisite graphs that oversimplify actual learning relationships. A teacher in the loop mechanism helps mitigate this. Additionally, the psychometric modeling component requires sufficient interaction density to generate reliable parameter estimates, potentially limiting applicability in specialized or low-enrollment contexts.

The computational requirements for real-time CReD computation and IRT parameter estimation may present scalability challenges in large educational deployments. However, the modular architecture enables selective implementation of framework components based on available resources and institutional priorities.

5. Case Study: Algebraic Equations

To illustrate the framework's operation, consider a middle school algebra curriculum focused on solving linear equations. The curricular ingestion process identifies discrete concepts including "Variables and Constants," "Basic Operations," "Equation Setup," "Isolation Techniques," and "Solution Verification." Prerequisite graph construction establishes dependencies where "Variables and Constants" precedes "Equation Setup," which precedes "Isolation Techniques."

Each concept undergoes Bloom's taxonomy decomposition. For "Isolation Techniques," the system creates sub-nodes: Remember (recall isolation rules), Understand (explain why isolation works), Apply (solve standard equations), Analyze (identify efficient solution paths), Evaluate (verify solution correctness), and Create (formulate equations from word problems).

Consider a student who fails at the Apply level for "Isolation Techniques" with mastery probability 0.60. The CReD computation initiates backtracking search, first checking lower Bloom levels within the same concept. Finding adequate mastery at Remember (0.85) and Understand (0.80) levels, the search moves horizontally to prerequisite concepts. At "Equation Setup," the student demonstrates insufficient mastery at the Apply level (0.65), triggering continued regression to "Basic Operations" where the Apply level shows mastery probability 0.45.

Simultaneously, external touchpoint analysis identifies critical gaps in "Arithmetic Operations" (specifically integer operations with negative numbers) and "Reading Comprehension" (parsing mathematical word problems). These external deficiencies are categorized within their respective domains and assessed through targeted micro-evaluations.

The resulting CReD output specifies: internal distance of 2 hops (from "Isolation Techniques Apply" to "Basic Operations Apply"), cognitive complexity at the Apply level, and external profile indicating deficiencies in arithmetic operations and reading comprehension. This analysis suggests that effective remediation requires addressing foundational arithmetic skills and comprehension strategies before returning to algebraic isolation techniques.

The psychometric modeling component transforms this interaction pattern into sparse matrix entries, contributing to IRT parameter estimation. Over time, the student's ability estimate ($\theta = -0.8$) indicates below-average mathematical proficiency, while concept difficulty parameters reveal that "Basic Operations Apply" ($\beta = -0.2$) is empirically easier than "Isolation Techniques Apply" ($\beta = 0.4$), validating the prerequisite structure.

For educational practitioners, this analysis provides concrete intervention guidance: prioritize arithmetic operation practice and reading comprehension support before advancing to algebraic techniques. The quantified regression depth indicates moderate remediation requirements, while the external profile highlights the need for cross-curricular support beyond mathematics instruction.

This systematic approach contrasts with traditional tutoring systems that might simply re-present algebraic isolation problems, potentially reinforcing student frustration without addressing underlying deficiencies. The CReD framework's diagnostic precision enables targeted resource allocation and realistic expectations for remediation timeline, supporting more effective educational intervention.

References

ACE: AI-Assisted Construction of Educational Knowledge Graphs with Prerequisite Relations. (2024). *Journal of Educational Data Mining*, 16(2), 85-114. https://doi.org/10.5281/zenodo.14250896

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Pearson Education Group.

Banujan, K., Kumara, S., Prasanth, S., & Ravikumar, N. (2023). Revolutionising educational assessment: Automated question classification using Bloom's taxonomy and deep learning techniques – A case study on undergraduate examination questions. *International Journal of Education and Development using Information and Communication Technology*, *19*(3), 259–278. https://files.eric.ed.gov/fulltext/EJ1413430.pdf

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). *Generative AI can harm learning* (Working Paper No. 4895486). SSRN. https://doi.org/10.2139/ssrn.4895486

Bender, M. A., Fineman, J. T., Gilbert, S., & Tarjan, R. E. (2016). A new approach to incremental cycle detection and related problems. *ACM Transactions on Algorithms*, *12*(2), Article 2. https://doi.org/10.1145/2756553

Chang, H.-H. (2024). Harnessing AI for educational measurement: Standards and emerging frontiers. *Journal of Educational and Behavioral Statistics*, *49*(5), 702–708. https://doi.org/10.3102/10769986241264033

Chen, Y., & Chang, H. H. (2018). Psychometrics help learning: From assessment to learning. *Applied Psychological Measurement*, *42*(1), 3–4. https://doi.org/10.1177/0146621617730393

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Díaz, B., & Nussbaum, M. (2024). Artificial intelligence for teaching and learning in schools: The need for pedagogical intelligence. *Computers & Education*, *217*, Article 105071. https://doi.org/10.1016/j.compedu.2024.105071

Favero, L., Pérez-Ortiz, J.-A., Käser, T., & Oliver, N. (2025). *Do AI tutors empower or enslave learners? Toward a critical use of AI in education* (arXiv:2507.06878). arXiv. https://arxiv.org/abs/2507.06878

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*(1), 30–37. https://doi.org/10.1111/j.1750-8606.2008.00072.x

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), Article 6. https://doi.org/10.3390/soc15010006

Hu, A. (2024). *Developing an AI-based psychometric system for assessing learning difficulties and adaptive system to overcome: A qualitative and conceptual framework* (arXiv:2403.06284). arXiv. https://arxiv.org/abs/2403.06284

Kuribayashi, T., Oseki, Y., & Baldwin, T. (2024). *Psychometric predictive power of large language models* (arXiv:2311.07484). arXiv. https://arxiv.org/abs/2311.07484

Learning Policy Institute. (2025, July 16). *An overview of teacher shortages: 2025*. https://learningpolicyinstitute.org/product/overview-teacher-shortages-2025-factsheet

Li, Y., Rakovic, M., Poh, B. X., Gasevic, D., & Chen, G. (2022, July). Automatic classification of learning objectives based on Bloom's taxonomy. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 530–537). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853191

Liang, C., Wu, Z., Huang, W., & Giles, C. L. (2015). Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1668–1674). Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1193

Majnik, M., Kristan, M., & Skočaj, D. (2013, February). *Knowledge gap detection for interactive learning of categorical knowledge*. Paper presented at the 18th Computer Vision Winter Workshop, Hernstein, Austria. https://core.ac.uk/download/pdf/11679313.pdf

Mirzadeh, S. I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2025). GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *Proceedings of the Thirteenth International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=AjXkRZIvjB

National Assessment Governing Board. (2025, January 29). *The Nation's Report Card shows declines in reading, some progress in 4th grade math*. News and Events. Retrieved July 17, 2025, from https://www.nagb.gov/news-and-events/news-releases/2025/nations-report-card-decline-in-reading-progress-in-math.html

National Center for Education Statistics. (2024, October 17). *Most U.S. public elementary and secondary schools faced hiring challenges for the start of the 2024–25 academic year* [Press release]. https://nces.ed.gov/whatsnew/press_releases/10_17_2024.asp

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving student modeling through partial credit and problem difficulty. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 11–20). Association for Computing Machinery. https://doi.org/10.1145/2724660.2724667

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2025, August 29). Generative Psychometrics via AI-GENIE: Automatic Item Generation with Network-Integrated Evaluation. https://doi.org/10.17605/OSF.IO/ZCYTB

Sapountzi, A., Bhulai, S., Cornelisz, I., & van Klaveren, C. (2021). *Personalized stopping rules in Bayesian adaptive mastery assessment* (arXiv:2103.03766). arXiv. https://arxiv.org/abs/2103.03766

Schmidt, D. P. (2020). *Identifying knowledge gaps using a graph-based knowledge representation* [Doctoral dissertation, Wright State University]. CORE Scholar. https://corescholar.libraries.wright.edu/etd_all/2313

Sudirtha, I., Widiana, I., & Adijaya, M. (2022). The effectiveness of using revised Bloom's taxonomy-oriented learning activities to improve students' metacognitive abilities. *Journal of Education and e-Learning Research*, *9*(2), 56–61. https://doi.org/10.20448/jeelr.v9i2.3804

Xu, Y., & Mostow, J. (2013, July). Using item response theory to refine knowledge tracing. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 358–365). International Educational Data Mining Society. https://educationaldatamining.org/EDM2013/proceedings/paper_120.pdf

Zhang, L., Lin, J., Sabatini, J., Borchers, C., Weitekamp, D., Cao, M., Hollander, J., Hu, X., & Graesser, A. C. (2025). *Data augmentation for sparse multidimensional learning performance data using generative AI* (arXiv:2409.15631). arXiv. https://arxiv.org/abs/2409.15631

Zhang, Y., Wu, R., Cai, P., Wang, X., Yan, G., Mao, S., Wang, D., & Shi, B. (2025). *LeanRAG: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval* (arXiv:2508.10391). arXiv. https://arxiv.org/abs/2508.10391