

Applying Evidence-Centered Design to Automated Evals of AI-Powered Assessment Systems

Kristen DiCerbo

Khan Academy
kdicerbo@cox.net

Britte Haugan Cheng

Menlo Education Research
bcheng@menloedu.org

John Whitmer

Learning Data Insights
john@ld-insights.com

Abstract

This paper demonstrates an application of Evidence-Centered Design (ECD) as a principled approach to design automated evaluations of AI-powered assessment outputs. We demonstrate this application through Khan Academy’s “Explain Your Thinking” conversational agent for mathematics items, showing how ECD’s layered models can be translated into rigorous and interpretable systems to demonstrate validity, reliability and fairness in AI systems to diverse audiences.

1 Introduction

The emergence of generative AI in educational assessment has expanded what is possible: real-time conversational probing of student reasoning, automated criterion-by-criterion scoring, and dynamic supervision of AI-generated instructional dialogue. These capabilities bring genuine potential — and a genuine validation problem. As AI systems take on roles previously reserved for human scorers and interviewers, generating follow-up questions, interpreting student responses, and assigning scores in real time, the field needs rigorous frameworks for designing and validating the performance of these systems.

Without such frameworks, automated evaluations of generative AI outputs, commonly called “evals”, risk becoming ad hoc checklists that test for surface-level accuracy without grounding claims in a defensible theory of evidence. This is not merely a theoretical concern. A growing body of work on automated scoring and LLM-based evaluation has demonstrated that agreement with human raters, while necessary, is insufficient as a standalone validity criterion: what counts as adequate evidence, and for what claims, must be specified before evaluation begins (Kane, 2006; Mislevy et al., 2003b).

Evidence-Centered Design (ECD), developed by Mislevy and colleagues, was designed precisely to

build validity arguments for claims (Mislevy et al., 2003a,b). ECD frames assessment as “an exercise in evidentiary reasoning.” Its layered architecture explicitly connects the claims being made about a student, the evidence required to support those claims, and the task conditions under which that evidence is gathered. Crucially, ECD builds validity into design from the outset rather than treating it as a post-hoc audit.

These properties make ECD a natural scaffold not only for designing assessments, but also for designing the evaluations that determine whether AI components within those assessments are functioning as intended. Yet despite ECD’s established role in assessment design, its application as a framework specifically for AI component evaluation, for structuring the claims, evidence, and tasks involved in evaluating AI scorer and conversationalist components, has not been systematically explored. This paper addresses that gap.

2 Background

2.1 Evidence-Centered Design

ECD is a comprehensive framework for assessment design introduced by Mislevy, Steinberg, and Almond (2003). Its conceptual roots draw on developments in Bayesian inference, cognitive science, and legal evidentiary reasoning, including Wigmore’s and Toulmin’s frameworks for structuring arguments from evidence to claim (Mislevy et al., 2003b; Toulmin, 1958).

The core of ECD’s Conceptual Assessment Framework (CAF) consists of four interrelated models:

- The Student Model specifies what claims are being made about the examinee — the knowledge, skills, and abilities (KSAs) the assessment is designed to measure. These are expressed as variables that can take on different

values and that can, in principle, be updated by evidence.

- The Evidence Model connects observations to claims. It has two sub-components: the evaluation component, which describes what features of a student's work product are relevant to each KSA, and the measurement model, which specifies how those features are translated into updated beliefs about the student model variables.
- The Task Model specifies the conditions under which evidence is elicited: the characteristic features of tasks that are relevant to the targeted KSAs, and the variable features that can be manipulated.

By requiring explicit articulation of each model and the connections between them, it forces designers to specify the full inferential chain from task performance to score-based claim, making the reasoning transparent, auditable, and extensible/improvable (Mislevy and Riconscente, 2006; Mislevy et al., 2012).

2.2 The Problem of AI Eval Design

Generative AI is increasingly being used to create new assessment items and experiences. The proliferation of AI components in educational assessment has created a new class of evaluation problems. An AI scoring application is not a test, it is a prediction of a score that must itself be evaluated. An AI conversation agent is not a student, it is an algorithm that predicts likely responses based on observed patterns whose output quality must be assessed against specified criteria. Compounding this factor, generative AI has built-in variation as a feature to create more dynamic, human-sounding language. The field has not yet developed principled, reusable frameworks for this class of evaluation problem.

Current practice in AI eval design tends toward one of two options. First, benchmark-based evaluation assess AI components against standardized datasets without specifying the inferential chain linking benchmark performance to real-world claims. Most benchmarks are built around predicting test responses, not identifying learning-related behaviors nor competencies. Second, ad hoc validation creates bespoke evaluation datasets and metrics, but often without a principled framework to ensure that the evidence gathered is sufficient for

the claims being made. Too often, these evaluations choose what is easy to measure using existing algorithmic techniques, instead of selecting the most important behavior to measure to demonstrate a claim.

Both approaches lack the explicit links between claims and evidence. ECD's framework addresses this antecedent question directly.

3 Context: Khan Academy's Explain Your Thinking System

Khan Academy's Explain Your Thinking (EYT) feature provides the empirical context for this paper. EYT engages grade 3–12 students in multi-turn AI-guided conversations about their mathematical reasoning after they solve a problem. The system's purpose is to elicit and assess the conceptual understanding underlying a student's correct or incorrect answer — a diagnostic function that goes beyond what a single response can reveal. Currently used in high school Algebra and Geometry with approximately 10,000 students, EYT extends well-established research on conversation-based assessment (Jackson and Zapata-Rivera, 2016; Liu et al., 2016) by using generative AI to operationalize conversations at scale — replacing the brittle, pre-scripted dialogue trees that constrained earlier systems with flexible prompt-based control (DiCerbo et al., 2026).

3.1 System Architecture

EYT employs a modular architecture with three distinct AI components:

- The AI Scorer evaluates each student response against a set of analytic, dichotomously-scored criteria authored by content experts. Each criterion represents a specific aspect of conceptual understanding targeted by the item. The Scorer's output, a binary judgment for each criterion, drives the conversation.
- The Conversation Agent generates targeted follow-up questions based on the criteria that the Scorer has identified as unmet. Its role is to probe the student's reasoning in a way that mirrors what a skilled teacher might do: asking questions that are targeted, non-leading, and appropriate to the student's demonstrated level of understanding.
- The Supervisor Agent reviews each question generated by the Conversation Agent before it

is presented to the student. Its role is to ensure that questions do not inadvertently provide the answer, violate content policies, or otherwise compromise the integrity of the assessment.

Together, these three components implement a complete assessment cycle: scoring, targeted questioning, and quality assurance. Each component makes consequential decisions that depend on the outputs of the others, creating a tightly coupled system in which validation failures in one component propagate through the whole.

3.2 EYT Validation Practice

Current validation practice for EYT has involved constructing synthetic student response datasets, establishing human-labeled ground truth through expert annotation, and iterating on scoring rubrics through pre-pilot validation cycles before advancing to pilot testing with authentic student responses collected from live use with approximately 10,000 high school mathematics students.

To make these practices concrete, a human content creator writes an item asking students to select which triangle is similar to a given triangle. They then write criteria for the EYT conversation about similarity, including, “Student understands that the two triangles are similar by AA similarity criteria (i.e., two of the angle pairs are the same). Similarity must be stated and not merely implied, and the angles must also be mentioned as a basis for similarity.” Following this, synthetic conversations responding to the item are developed. Humans then label these conversations as meeting or not meeting the criteria. In places where there is significant human disagreement, the criteria are revised and the process started again. When there is human agreement, AI scoring prompts are developed and iterated on until they reach sufficient reliability.

Human agreement on criterion judgments is measured using Matthews Correlation Coefficient (MCC) and Cohen’s kappa to establish a reliability ceiling before AI agreement is assessed. This sequencing, establishing human-human reliability before measuring human-AI agreement, reflects an important methodological principle that ECD makes explicit.

4 Methods: Mapping ECD to AI Eval Design

We conduct a structured conceptual mapping between ECD’s framework and the practice of AI eval

design, grounded in the EYT case. For each ECD model, we identify the corresponding eval design question and demonstrate how EYT’s validation approach addresses each question, or would benefit from explicitly addressing it. The mapping is summarized in Table 1 (below).

4.1 The Student Model

In an AI eval context, the Student Model becomes what we term the AI Component Model: an explicit specification of the claims being made about the AI component’s behavior. This is the first discipline ECD imposes, and it is frequently skipped in practice.

For EYT’s AI Scorer, the claim is that its binary criterion judgments correspond to those an expert human rater would make when applying the same rubric to the same response. This claim has an important internal structure: it is made per criterion, not globally. A scorer that performs well on high-frequency, clear-cut criteria but poorly on low-frequency, ambiguous ones has a very different validity profile than one that performs uniformly, a distinction that aggregate accuracy statistics obscure. ECD’s student model architecture, which represents claims as a structured set of variables rather than a single score, provides the representational vocabulary to make this distinction explicit.

There are currently two claims for the Conversation Agent: (a) targeted at one or more unmet criteria, (b) not leading or outright giving the answer. Each of these sub-claims requires its own evidence model. In the future, other criteria may be explored, such as aligning to the student’s demonstrated level of reasoning, requiring the development of additional evidence models, or other items..

4.2 The Evidence Model

ECD’s evidence model has two sub-components that map directly onto the two central decisions in eval dataset construction: what to label, and how to aggregate those labels into a measurement.

The evaluation component maps onto the design of the labeling rubric and the annotation task. In EYT’s pre-pilot validation, human expert labels serve as Observable Variables in ECD’s terminology: observable expressions of the targeted construct (criterion mastery) that provide the evidentiary basis for evaluating the AI component’s claims. The quality of these labels and therefore the quality of the entire evaluation depends on the quality of the annotation rubric and the expertise

ECD Model	AI Eval Analogue	Core Design Question	EYT Instantiation
Student Model	AI Component Model	What claims are being made about the component's behavior?	Scorer: binary criterion judgments match expert human rater; Conversation Agent: questions are aligned to criteria and non-leading
Evidence Model (Evaluation Component)	Labeled Dataset Design	What observable evidence bears on which claims?	Human-labeled criterion judgments as ground truth; rubric-governed annotation
Evidence Model (Measurement Model)	Metric Selection & Sequencing	How is evidence translated into updated claims?	MCC and Cohen's kappa; human IRR established before AI agreement assessed
Task Model	Eval Dataset Design	What response conditions must be represented to support inferences?	Synthetic datasets with balanced categories, challenge sets, edge cases

Table 1: ECD Models and Their AI Eval Analogues

of the annotators. ECD's framework makes this dependency explicit and provides vocabulary for discussing it.

The measurement model maps onto the selection and interpretation of agreement statistics. ECD distinguishes between the evaluation component (the rubric that produces labels) and the measurement model (the statistical procedure that aggregates those labels into evidence about the claim). This distinction clarifies why human inter-rater reliability must be established before AI agreement is assessed: if humans applying the rubric cannot agree, the labels are not a reliable ground truth, and AI-human agreement statistics are uninterpretable. The field does not always observe this sequencing. ECD provides both the vocabulary and the logical structure to make the sequencing requirement explicit.

In EYT's context, the measurement model also requires distinguishing error types. Because false positive errors by the Scorer terminate conversation prematurely, they carry a higher cost than false negative errors, which merely prolong the conversation unnecessarily. An ECD-informed evidence model specifies this asymmetry in advance — it is a property of the measurement model, not a post-hoc discovery from pilot data.

4.3 The Task Model

ECD's task model specifies the characteristic and variable features of assessment tasks that are rele-

vant to the targeted KSAs. In an eval context, the task model translates to the design of the evaluation dataset itself: what response types must the dataset include to support valid inference about the AI component's behavior?

EYT's use of synthetic response datasets — constructed to include balanced score categories, challenge sets targeting edge cases, and prevalence-weighted distributions — reflects implicit task modeling. ECD provides the vocabulary to make this implicit modeling explicit and replicable. In particular, ECD's distinction between characteristic features (features that make a task an instance of the relevant task type) and variable features (features whose manipulation generates evidence about specific aspects of component behavior) maps cleanly onto how datasets are constructed. The characteristic features of the dataset include back and forth interactions between the AI tutor and student, AI tutor probing student understanding, and student responses. Variable features include the length of student response, the correctness of student responses, and completeness of individual student responses.

5 Anticipated Results

This paper reports work in progress. The full conceptual mapping, including populated examples from EYT's validation materials, will be completed in advance of the conference. Anticipated results include:

- A complete translation table mapping each

ECD construct to its AI eval analogue, populated with concrete EYT examples

- Identification of at least two cases where ECD would have prescribed a different validation sequence than was initially followed, and an assessment of the consequences
- A preliminary assessment of the generalizability of the mapping to other AI-powered assessment components beyond EYT

6 Discussion and Anticipated Contributions

6.1 ECD Improves Validity by Forcing Claim Specification

The most fundamental contribution of applying ECD to AI eval design is the discipline it imposes on claim specification. Ad hoc evals frequently conflate what is being claimed about an AI component, what evidence would support that claim, and what task conditions are required to elicit that evidence. ECD's layered architecture separates these questions and requires each to be answered before the next can be addressed.

This has direct practical consequences. An AI scorer evaluated only against a convenience sample of student responses cannot support claims about its performance on low-frequency but educationally significant response types. An AI Conversation Agent evaluated only on the fluency of its questions cannot support claims about their non-leading quality. ECD's framework makes these gaps visible before evaluation begins — not after results fail to generalize.

6.2 ECD Supports Modular Eval Design

EYT's modular component architecture — a Scorer, a Conversation Agent, and a Supervisor, each with distinct responsibilities — creates both a design opportunity and an evaluation challenge. Each component can, in principle, be evaluated independently. But the components are also coupled: the Conversation Agent's behavior is conditioned on the Scorer's output, and the Supervisor's behavior is conditioned on the Conversation Agent's output. Evaluation designs that treat the components as fully independent may miss failure modes that only emerge in the full system.

ECD's framework supports modular eval design in a way that is sensitive to this coupling. By specifying the claims for each component explicitly,

and by tracing the inferential chain from each component's output to the downstream claims it supports, an ECD-informed eval design can identify the points at which component-level validity failures propagate to system-level validity failures — and prioritize evaluation effort accordingly.

6.3 ECD as a Reusable Architecture for a Field That Needs One

Beyond the EYT case, ECD offers a reusable architectural vocabulary for AI eval design that does not need to be reconstructed from first principles for each new system. The translation table in Section 4 (Table 1) is intended as a working template that developers of other AI-powered assessment systems can adapt to their own contexts.

The field of AI evaluation in education is currently characterized by proliferating evaluation frameworks and benchmarks with inconsistent theoretical grounding (cf. recent work on LLM-as-judge methodology and its reliability challenges (Schroeder and Wood-Doughty, 2025)). ECD does not resolve all of these challenges, but it provides a principled starting point: a mature framework with a 20-year track record in high-stakes assessment design, whose core logic is directly applicable to the AI eval problem and whose vocabulary is already shared across the measurement science community.

6.4 Open Questions and Limitations

Several open questions remain. First, ECD was designed for assessments of human examinees; adapting its vocabulary for AI components requires conceptual work that this paper initiates but does not complete. Second, the translation from ECD's static, design-time framework to the dynamic, iterative context of AI development, where components are continuously updated and eval results feed back into model training, raises questions about how ECD's validity argument should be revised as the system evolves. Third, the relationship between ECD-informed eval design and the broader project of AI safety and alignment evaluation remains to be worked out.

Extensions to be explored in future work include: using ECD as a framework for other AI-centered design activities beyond eval; applying ECD's evidentiary reasoning structure to the automated construction and updating of evals; and examining how ECD-informed eval design interacts with emerging regulatory and legal frameworks for high-stakes AI systems.

7 Conclusion

This paper argues that ECD provides a principled, reusable architecture for designing evaluations of AI-powered assessment components, and demonstrates this argument through a structured conceptual mapping applied to Khan Academy’s Explain Your Thinking system. The mapping reveals both 1) the practical value of ECD for AI eval design, forcing claim specification, supporting modular evaluation, providing reusable architectural vocabulary and 2) the conceptual work required to address the open questions that arise when a framework designed for static human assessments is applied to dynamic, interdependent AI systems.

As AI systems take on increasingly central roles in educational assessment, the field needs evaluation frameworks that are not merely technically adequate but theoretically grounded and defensible. ECD, with its roots in evidentiary reasoning and its 20-year track record in high-stakes assessment design, offers a strong foundation for that work.

References

- Kristen DiCerbo, Lisa Deters, Tyler Burleigh, J. Chen, and S. Frohn. 2026. [Conversation-based assessment powered by generative ai](#). In M. Mayrath, J. T. Behrens, and D. H. Robinson, editors, *Handbook of Generative AI in Education*. Springer, Cham.
- G Tanner Jackson and Diego Zapata-Rivera. 2016. Conversation-Based assessment. In *The Handbook of Cognition and Assessment*, pages 563–579. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Michael T. Kane. 2006. Validation. In R. Brennan, editor, *Educational measurement*, 4 edition, pages 17–64. American Council on Education and Praeger.
- Lei Liu, Jonathan Steinberg, Farah Qureshi, Isaac Bejar, and Fred Yan. 2016. [Conversation-based assessments: An innovative approach to measure scientific reasoning](#). *Bulletin of the IEEE Technical Committee on Learning Technology*, 18(1):10–13.
- Robert J. Mislevy, Russell G. Almond, and Janice F. Lukas. 2003a. [A brief introduction to evidence-centered design](#). *ETS Research Report Series*.
- Robert J. Mislevy, John T. Behrens, Kristen E. DiCerbo, and Roy Levy. 2012. Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4:11–48.
- Robert J. Mislevy and Michelle M. Riconscente. 2006. Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing and T. Haladyna, editors, *Handbook of test development*, pages 61–90. Erlbaum.
- Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. 2003b. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1):3–67.
- Kai Schroeder and Zach Wood-Doughty. 2025. [Can you trust llm judgments? reliability of llm-as-a-judge](#). *Preprint*, arXiv:2412.12509.
- Stephen E. Toulmin. 1958. *The uses of argument*. Cambridge University Press.