

# Rethinking Validity in Educational Assessment When AI Co-Produces Performance

**Xiaoran Li**  
Cognality Learning LLC

**Tanesia Beverly**  
Independent Researcher

## Abstract

Generative AI complicates a core assumption of assessment that observed performance reflects an individual's own cognition. Using StudyChat, we show AI supply only moderately tracks student intent, and assignment scores are largely insensitive to either. With the disruption of validity warrant, response process validity needs reconceptualizing when AI co-produces performance.

## 1 Introduction

Artificial intelligence (AI) is deeply integrated into educational settings, shaping how learners produce answers across all areas of learning. Prior research has examined the validity of AI-generated scores (Ferrara & Qunbar, 2022; Hannah et al, 2023) and automatically generated assessment items (Gierl & Haladyna, 2013). However, most of this work assumes that the underlying basis for validity remains the same. Traditional assessment frameworks assume that observed performance mainly reflects individual cognition or cognitive ability. However, AI changes this dynamic, as the solutions are co-produced through interactions between AI and learners. The link between observed performance and underlying competence has changed consequentially, and raises important questions about what qualifies as evidence of competence. Additionally, it impacts how validity arguments should be developed in AI-mediated contexts (Messick, 1995; Mislevy, 2018; Li & Beverly, 2026).

Response process evidence has long been regarded as a critical source of validity evidence because it provides insight into how learners engage in the cognitive processes intended by the construct being assessed (Ercikan & Oliveri, 2016; Sireci, 2007). Historically, such evidence has been gathered through think-aloud protocols, cognitive interviews (Leighton, 2004; Padilla & Benitez, 2014), and digital process data such as response times, clickstreams, navigation patterns, and eye-tracking measures (Bergner & von Davier, 2019; Goldhammer & Zehner, 2017; Li, Banerjee,

& Zumbo, 2017). A related body of work has examined aberrant behaviors that threaten the response process validity, such as answer-changing (Farr, Pritchard, & Smitten, 1990), test-wiseness and test anxiety (Millman, Bishop, & Ebel, 1965; Geiger, 1997), and speeded or rapid-guessing response patterns (Liu, Li & Xin, 2019; Kahraman, Cuddy, & Clauser, 2013; Schnipke & Scrams, 1999; Wise, 2017). Although these approaches vary in focus and methodology, they share a common assumption, that the observed processes originate primarily from the learner. Generative AI challenges this assumption by participating directly in the co-production of response processes themselves by providing ideas, hints, and even completely finished work. Consequently, the question is no longer simply whether learners engaged in the intended cognitive processes that are required by the task, but which aspects of human-AI interaction provide interpretable evidence that those cognitive processes occurred. We argue that this shift necessitates a reconceptualization of response process validity and reimagining assessment design.

This paper located the threat to response process validity within ECD's own evidentiary chain. Specifically, the task models of traditional assessment do not represent generative AI as a participant, yet its contribution enters the work product regardless; and the evaluation model has no mechanism for detecting how much of that product it supplied. Using StudyChat data ((McNichols et al., 2026), we decomposed the observed human-AI interaction logs into a

learner-attributed component (i.e., prompts) and an AI-attributed component (i.e., AI responses), then tested 1) whether an AI's response reflects what a student actually asked for, and 2) whether an assignment score reflects how the response was produced.

## **2 Theoretical Framework: Evidence-Centered Design and the Evidential Chain**

### **2.1 The ECD Evidential Chain**

Evidence-Centered Design conceives of assessment as an evidentiary argument: a task model specifies the situations that will elicit performance; a work product is the artifact a student produces in that situation; an evidence model identifies observable features of the work product and accumulates them into scores; and a student model supports the interpretive claim about competence that the score is meant to warrant (Mislevy et al., 2003). Validity, in this framing, depends on each link in the chain preserving a defensible connection to the one before it: the work product must be attributable to the student the task model assumes is producing it, the evidence identified from that product must reflect the construct of interest, and the accumulated score must remain sensitive to variation in that evidence. (AERA, APA, & NCME, 2014).

### **2.2 The Socio-Cognitive Layer**

The theoretical warrant for why this chain can break comes from Mislevy's later work on ECD into a socio-cognitive perspective (Mislevy, 2018, 2024). He argued that competence is never purely in the head, it is enacted through tools, language, and interaction. The measurement has traditionally rested on an individualist view of competence that was only defensible when the tools available to examinees were fixed, known, and roughly uniform (e.g., a calculator, Mislevy, 2018). Generative AI breaks this idealization because its contribution is not fixed or uniform, it has its own response policy, participatory and variable across turns and students. This is the theoretical reason the evidentiary chain breaks the way we show it breaking below, rather than merely adding noise that averages out.

### **2.3 Where AI Mediation Breaks the Chain**

AI participation enters the ECD pipeline at precisely the two links current task and evidence models do not represent. First, at the work-product link: when an AI system co-produces a response, the work product is no longer attributable to the student alone, yet conventional task models specify a single human performer and have no slot for a second, variable-policy contributor. Second, at the evidence link: even where an interaction log is retained, current evidence models were built to accumulate evidence of individual cognition, not to separate a student's contribution from an AI's before scoring. An unmodeled actor entering at either link is not a measurement imperfection that averages out across a group of examinees; it is a structural gap in the evidentiary argument itself, because the actor's contribution is neither random noise nor a stable offset.

## **3 Research Questions**

To empirically test whether AI mediation disrupts this evidentiary chain, we posed two research questions:

**RQ1.** Does AI supply reflect student intent/request in AI-mediated interactions?

**RQ2.** Does the assessment score reflect process information (intent/request and supply) in AI-mediated environments?

## **4 Method**

### **4.1 Data**

We use StudyChat (McNichols et al., 2026), a corpus of real student–AI interactions collected from an introductory, university-level artificial intelligence course in which students used an LLM-powered chatbot (GPT-4o-mini) while completing programming assignments. We restrict analysis to 65 students across six assignments from fall 2024 who gave permission to use their scores, excluding one assignment consisting of a generic written reflection with no problem-solving content, yielding 4,066 student–AI turns (a student prompt paired with the AI's response) in total.

### **4.2 Recoding Procedure**

To better quantify student intent (i.e., how much of the target performance the student's prompt asked for) and AI supply (i.e., how much of the target performance the AI's response actually delivered), each turn was coded on two four-point ordinal

scales. Intent/Request (R) captures the level of assistance a student's prompt asked for: 1=a request for a principle, definition, or method for conceptual understanding (e.g., what a tree model means); 2=a request for execute the students' own idea with specific steps, or check/diagnose something they made; 3=a request for strategic or partial worked help to produce the actual graded answers; and 4=a request for a complete solution. Supply (S) captures the level of assistance actually provided by the AI's, coded on the mirrored four anchor points independently of what was requested for each turn: 1= Explained an idea or method in general; 2=Provided a small specific piece of work with detailed instruction; 3=Finished part of the assignment that would count toward the final grade; 4=Provided the completely finished work.

A stratified 200-turn sample was independently double-coded by a human rater and by an LLM-based rater, blind to each other's codes, to establish agreement before the full corpus was coded (reported in Section 5.1). Following acceptable agreement, the remaining turns were coded by the LLM rater.

#### 4.3 Analytic Approach to RQ1

RQ1 asks whether AI supply reflects student intent. The turn-level mismatch between what was requested and what was supplied is defined as

$$M_i = S_i - R_i \quad (1)$$

for turn  $i$ , where positive values indicate the AI supplied more assistance than the student's turn requested. We report the distribution of  $M_i$  (mismatch), the Spearman rank correlation between  $R$  and  $S$ , and a concordance table cross-tabulating  $S$  against  $R$ , to characterize both the overall tendency and the turn-level variability in the  $R$ - $S$  relationship.

#### 4.4 Analytic Approach to RQ2

RQ2 asks whether the assignment score reflects this process information.  $R$  and  $S$  were aggregated to the student $\times$ assignment level as the mean over that student's turns within the assignment,

$$\bar{R}_s = (1/T_s) \sum_i R_i \quad (2)$$

$$\bar{S}_s = (1/T_s) \sum_i S_i \quad (3)$$

where  $T_s$  is the number of turns students produced on that assignment. Assignment-level scores were

then related to mean  $R$  and mean  $S$  in the linear model

$$\text{score}_s = \beta_0 + \varepsilon_s \quad (4)$$

$$\text{score}_s = \beta_0 + \beta_1 \bar{R}_s + \varepsilon_s \quad (5)$$

$$\text{score}_s = \beta_0 + \beta_2 \bar{S}_s + \varepsilon_s \quad (6)$$

$$\text{score}_s = \beta_0 + \beta_1 \bar{R}_s + \beta_2 \bar{S}_s + \varepsilon_s \quad (7)$$

We report Spearman correlations between assignment-level mean  $R$ , mean  $S$ , and assignment score, and a nested sequence of linear models: an intercept-only model,  $R$ -only,  $S$ -only, and the joint model (Equation 4-7) to assess how much variance in score is explained by process information, individually and jointly.

## 5 Results

### 5.1 Recoding Reliability

On the 200-turn double-coded sample, human-LLM agreement for Request was 86.3% exact and 98.5% within one scale point; agreement for Supply was 74.5% exact and 95.5% within one scale point. Both scales met the threshold for proceeding to LLM coding of the full corpus, with Supply showing somewhat lower exact-match reliability than Intent/Request, consistent with Supply requiring a judgment about the completeness of a generated solution, a harder discrimination than classifying a student's stated request.

### 5.2 RQ1: Does AI Supply Reflect Student Intent/Request?

Table 1: Distribution of coded Intent/Request (R) and Supply (S) levels across 4,066 turns.

| Level | R (n) | S (n) |
|-------|-------|-------|
| 1     | 971   | 1,120 |
| 2     | 1,522 | 826   |
| 3     | 1,426 | 1,443 |
| 4     | 147   | 677   |

Descriptively, intent/requests cluster in the moderate range (72.5% of turns were level 2 or level 3), while full-solution requests (level 4) are rare (3.6%). Supply is more heavily weighted toward complete or near-complete responses (levels 3-4 account for 52.1% of turns) than Intent/Request is toward complete-solution asks,

an asymmetry that foreshadows the mismatch result below.

The turn-level mismatch (Equation 1) has a mean of +0.237 and only 56.9% of turns show exact R–S agreement. The positive mean was consistent across all six assignment topics, ranging from +0.13 to +0.37, indicating a stable tendency for the AI to over-supply relative to what was requested rather than an artifact of any single assignment.

The Spearman rank correlation between R and S was  $\rho = .61$  ( $p < .001$ ), indicating a positive but only moderate rank association. request level explains part, but far from all, of the variation in what the AI supplied. The concordance table below shows this more directly: at the low end, most level 1 requests (73.1%) received level 1 supply, but a substantial minority (24.1%) received level 3 supply. At the high end, the pattern is starker: only 51.0% of level 4 (full-solution) requests received level 4 supply, while low-R turns overall still generated levels 3–4 responses often enough that supply cannot be predicted from request alone with confidence.

Table 2: R–S concordance (row percentages).

| R \ S | S=1   | S=2   | S=3   | S=4   |
|-------|-------|-------|-------|-------|
| R=1   | 73.1% | 2.5%  | 24.1% | 0.3%  |
| R=2   | 21.4% | 46.5% | 22.2% | 9.9%  |
| R=3   | 5.8%  | 5.8%  | 57.0% | 31.4% |
| R=4   | 1.4%  | 8.2%  | 39.5% | 51.0% |

### 5.3 RQ2: Does Assignment Score Contain Process Information?

At the student×assignment level ( $n = 286$ ), Spearman correlations between process measures and assignment\_score were weak and non-significant for Intent/Request ( $\rho = .07$ ,  $p = .27$ ) and marginal for Supply ( $\rho = .11$ ,  $p = .06$ ). Figure 1 shows both relationships, scores cluster tightly near ceiling across the full range of mean Request and mean Supply, with a thin tail of lower-scoring students showing no systematic association with how much assistance they requested or received.

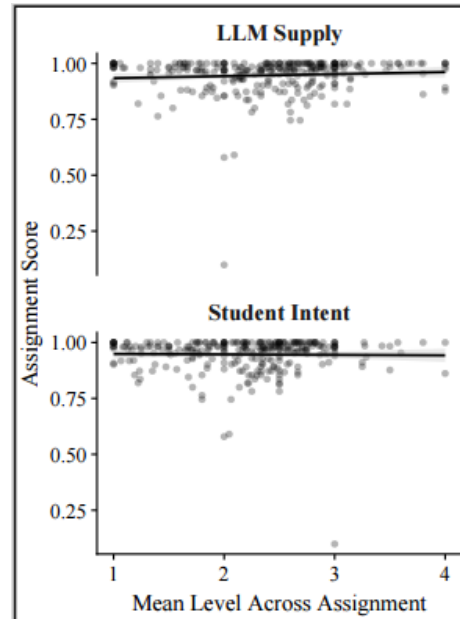


Figure 1: Assignment score against mean Request (a) and mean Supply (b) at the student×assignment level ( $n = 286$ ).

The intercept-only model shows assignment scores clustering near ceiling ( $M = 0.947$ , residual  $SE = 0.081$  on a 0–1 scale), leaving little room for any predictor to explain variance. Table 3 reports the nested regression sequence.

Table 3: Nested linear models predicting assignment\_score.

| Model          | R <sup>2</sup> | Adj. R <sup>2</sup> | F    | p    |
|----------------|----------------|---------------------|------|------|
| Intercept only | —              | —                   | —    | —    |
| Request        | .000           | –.003               | 0.07 | .787 |
| Supply         | .006           | .002                | 1.65 | .200 |
| Request+Supply | .020           | .013                | 2.87 | .058 |

Neither Intent/Request nor Supply alone explains a meaningful share of score variance. In the joint model, the two coefficients diverge in sign — mean R is negative and marginally significant ( $\beta = -.025$ ,  $p = .045$ ) while mean S is positive and marginally significant ( $\beta = .026$ ,  $p = .018$ ). This pattern is consistent with statistical suppression: each predictor captures a negligible effect once the shared variance with the other is partialled out, and the joint model's overall F remains only marginal ( $p = .058$ ). We report this pattern rather than over-interpreting it as evidence that requesting help lowers scores while receiving it raises them; with  $R^2 = .020$ , the joint model explains roughly 2% of

score variance regardless of how the individual coefficients are read.

## 6 Discussion

### 6.1 Where the Evidential Chain Breaks

Read together, RQ1 and RQ2 locate two distinct breaks in the ECD evidentiary chain described in Section 2. The first break is at the work-product link: Section 5.2 shows that what an AI supplies is only moderately predictable from what a student requested ( $\rho = .61\%$ ), and that the AI systematically over-supplies relative to request. A transcript of a student working with an AI is a record of two participants, and a meaningful share of what appears in it originates in the AI's own response tendencies rather than the student's intent. Reading such a transcript as a window onto a single student's effort or understanding, is therefore unsupported without first accounting for how much the AI contributed on its own initiative.

The second break is at the evidence link. If assignment scores account for how a response was produced, the first break would be self-correcting. For instance, heavily AI-supplied work would simply score lower than lightly-supplied work, and the score would remain an interpretable summary. Section 5.3 shows this does not happen: score is nearly blind to both intent/request and supply (joint  $R^2 \approx .02$ ), and this blindness is compounded by a ceiling effect, since scores compressed near 0.95 leave little variance for any process measure to explain even where a true relationship exists. A grade is meant to stand in for "this student understands the material," and that substitution is only defensible if performance and grade move together. Two students can arrive at the identical grade after taking substantively different paths, for example, one doing the reasoning themselves, one letting the AI do most of the work, but the grade cannot tell them apart.

### 6.2 Implications

Current institutional responses to AI-mediated assessment such as going back to bluebook, increasing oral defense, using AI-detection tools, and outright banning the use of AI largely treat the problem identified above as one of containment. Prohibition and detection enforce the individualist view temporarily, but they do not restore the interpretability of the evidence itself, and they

forgo whatever pedagogical value AI-mediated interaction could otherwise provide.

We argue instead that AI participation is not a contamination problem to be managed, but a design problem to be solved (Beverly & Li, 2026). That's said, assessment interactions should be engineered so that the evidence they generate remains interpretable regardless of whether, or how much, AI participates in producing the response. This reframes the guiding question from whether a student's answer is authentically their own to whether the questions asked, and the moves that follow them, actually reveal a student's understanding in the first place. The results above suggest that, at present, they largely do not in traditional assessments. Interactions in StudyChat are structured around task completion rather than around eliciting and documenting the reasoning process itself, and the resulting scores do not recover what the interaction concealed.

### 6.3 Limitations and Future Work

Several limitations qualify these findings. The R and S scales, while showing acceptable human–LLM agreement, rely on a single course's assignments and a single underlying model (GPT-4o-mini); supply behavior may differ across models, domains, and system prompts that explicitly instruct pedagogical restraint. The joint regression's suppression pattern (Section 5.3) also warrants replication with a larger sample before its individual coefficients are treated as substantive.

If the evidence identified here is not currently interpretable, the natural next question is what would make it so. Specifically, what kind of assessment questions can elicit evidence-bearing responses under AI mediation? Answering this question will provide a foundation for designing AI-mediated assessments that elicit more informative evidence regardless of AI participation, rather than simply restricting access to the AI tool.

## 7 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

- Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6), 706–732.  
<https://doi.org/10.3102/1076998618784700>
- Beverly, T., & Li, X. (2026). From task design to interaction design: A unified theory of validity for AI-mediated assessment. Paper presented at the *IMPS 35th conference*. Seoul, South Korea.
- Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education*, 29(4), 310–318.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee changes an answer. *Journal of Educational Measurement*, 27(3), 209–226.
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3), 288–313.
- Geiger, M. A. (1997). The effects of test-wiseness and test-taking anxiety on college student performance. *Journal of Experimental Education*, 66(1), 75–88.
- Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge. — includes Bejar, I. I., "Item generation: Implications for a validity argument" (Ch. 4), which is the direct validity-theory anchor for this literature.
- Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, 15(3–4), 128–132.  
<https://doi.org/10.1080/15366367.2017.1411651>
- Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, 20(4–5), 399–420.
- Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, 37(5), 343–360.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23(4), 6–15.
- Li, X., & Beverly, T. (2026). From Huamn-Only to Hybrid Performance: Rethinking Validity in AI-Mediated Assessment. Paper presented at the *International Test Commission Biennial Conference*. Auckland, New Zealand.
- Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and if not what would it take to do so? In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 159–178). Springer.
- Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives*, 17(3), 133–142.
- McNichols, H., Ikram, F., & Lan, A. (2026). The StudyChat dataset: Analyzing student dialogues with ChatGPT in an artificial intelligence course. In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026)*. ACM.  
<https://doi.org/10.1145/3785022.3785029>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25(3), 707–726.  
<https://doi.org/10.1177/001316446502500304>
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.  
<https://doi.org/10.4324/9781315871691>
- Mislevy, R. J. (2024). Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment. *Psychometrika*, 89(1), 64–83.  
<https://doi.org/10.1007/s11336-024-09966-5>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational

- assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–67.
- Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144.
- Raskin, J. D. (2002). Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism. *American communication journal*, 5(3), 1-25.
- Schnipke, D. L., & Scrams, D. J. (1999). Modeling item response times with a two-state mixture model: A new approach to measuring speededness (LSAC Research Report Series No. 96-02). Law School Admission Council.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477–481.  
<https://doi.org/10.3102/0013189X07311609>
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52–61.