ERICA J. YOON, PhD

erica.jiye.yoon@gmail.com · 650-924-5675 · linkedin.com/in/ericajyoon · Palo Alto, CA

Research Portfolio

I study how learning systems produce valid signals of understanding, and how to design both the learning experience and the measurement architecture to make those signals reliable, actionable, and honest.

My background is in developmental and cognitive psychology (Stanford PhD), but the questions I have been working on for the past decade are fundamentally applied: What do learners actually need, and how do I find out when they won't tell me directly? Does this assessment measure what I think it measures? Does early engagement predict outcomes reliably enough to act on? When AI enters the system, what breaks, and how do you fix it? These are user research questions as much as instructional design questions, and answering them requires the same toolkit: careful observation, clear hypotheses, behavioral data, and a willingness to be wrong about what you thought you knew.

What that has looked like in practice: building learning programs from zero with no predecessor and no playbook, which meant uncovering learner needs from the ground up through direct observation, structured one-on-one conversations, behavioral data, and iterative course corrections. Translating those findings into design decisions. Measuring whether the decisions worked. Bringing the evidence back to senior faculty and program leadership to inform strategy. And then doing it again. The pieces in this portfolio document that cycle across three contexts: a dual-modality course design study, an AI feedback pipeline built and debugged in production, and a construct validity analysis of project-based assessment. Together they show how I move from observation to insight to design decision to measurement, and back to observation again.

Portfolio Contents

Case Study 1 *Same Objectives, Different Worlds*	How I designed Cross-Cultural Psychology simultaneously for two structurally different learner populations: dual-enrollment high school students and online adult learners. With no predecessor and no playbook for either context, I uncovered learner needs from scratch through direct observation, structured one-on-one conversations embedded into the course design, and behavioral data from Canvas analytics. Documents the research, design decisions, and outcome validation across 4 course iterations (N=151 total). Core finding: early Week 1 behavioral engagement predicted final course success with a 2.1x effect size.
Case Study 2 *Building and Evaluating AI in the Classroom*	A two-act study. Act 1 documents the design, failure modes, and iterative calibration of a ChatGPT-based feedback pipeline processing 640+ student submissions per term. Act 2 documents my evaluation of Canvas AI features as an Early Access Evaluator for Instructure, and how the practitioner experience of Act 1 directly shaped the product insights I delivered to the Canvas AI infrastructure team in Act 2.
Research Brief *Does the Project Actually Measure Learning?*	A construct validity analysis of project-based assessment in a General Psychology course (N=37). Tests whether a staged open-ended project measures the same underlying competency as a closed-note final exam, and finds that it does (r=.677, p<.001), with validity increasing as cognitive demand increases across project components. Directly relevant to the question of whether alternative assessments produce valid learning signals in AI-accessible environments.

Background

Academic research: PhD in Psychology, Stanford University (2019). Specialization in developmental psychology, pragmatic language, and computational modeling. Published in *Open Mind* (MIT Press), *Journal of Experimental Child Psychology,* and *Developmental Science.*

Applied work: Inaugural Teaching & Program Coordinator, Stanford Symbolic Systems Program (2019–2022). Associate Professor of Psychology, College of San Mateo (2022–2026). Canvas AI Early Access Evaluator, Instructure (Summer 2025). Google Certified Educator Level 1 (2026).

Recognition: K. Jon Barwise Award for Distinguished Contributions, Stanford Symbolic Systems Program (2023). Hastorf Teaching Award, Stanford Department of Psychology.

Same Objectives, Different Worlds

Designing Cross-Cultural Psychology for Two Learner Populations Simultaneously

Erica J. Yoon, PhD · PSYC 230, College of San Mateo · 2022–2026

MY ROLE

Context: Sole instructor and course designer across both modalities. No prior curriculum, no precedent for dual-modality delivery in this program.

What I did: Identified learner behavior differences through direct observation and Canvas analytics; designed two structurally distinct course architectures for the same learning objectives; built assessment systems; collected and analyzed outcome data; iterated across 4 course implementations (2 per modality).

Methods: Behavioral tracking (Canvas analytics), quasi-experimental comparison across cohorts, correlational analysis (project scores vs. exam scores), direct observation during class, informal user interviews embedded in 1:1 student meetings.

Sample: N=88 in-person (HS dual enrollment, 2 iterations); N=63 online async (mixed-age adult learners, 2 iterations).

Tools: Canvas LMS (data collection, delivery), Canvas analytics (Week 1 engagement tracking), R-adjacent correlational analysis, audio/video feedback tools, paper-based formative assessments.

RESEARCH QUESTION

When two structurally different learner populations are given identical learning objectives, how much does the course design need to differ — and how do you validate that both paths are actually working?

WHAT I OBSERVED (USER RESEARCH)

- **HS students experienced the instructor as unapproachable.** Dual-enrollment juniors and seniors from Hillsdale High School were accustomed to their own teachers; an external college instructor felt distant. I observed low initiation of contact, reluctance to ask questions, and passive attendance behavior. This was a trust gap, not a comprehension gap — and it required a deliberate structural solution.
- **Online students disengaged silently and early.** Without synchronous contact, early non-engagement was invisible. I began tracking Week 1 Canvas submission behavior across cohorts and found a consistent pattern: students who did not complete Week 1 activities passed at dramatically lower rates. This was a leading indicator problem: the course needed an early detection and intervention system, not end-of-semester grade checks.
- **AI availability changed what assessments could validly measure.** In the fully asynchronous online environment, traditional exams no longer reliably distinguished genuine learning from AI-assisted output. I observed students completing open-resource assessments without internalizing content. This required a fundamental redesign: shifting from output-testing to process-documentation, where the assignment architecture itself made genuine engagement the path of least resistance.

DESIGN DECISIONS ACROSS MODALITIES

Dimension	In-Person · Hillsdale HS Dual Enrollment (n=88)	Online Async · CSM Adult Learners (n=63)
Format	75-min structured lectures with activity breaks every 5–10 min (think-pair-share, board sharing, prediction tasks). Paper knowledge check at end of each class — formative, graded on engagement not correctness. Group unit test segment: individual attempt + open-note group Scantron (10% of test grade).	Fixed weekly sequence: video lectures + readings → knowledge log → knowledge check quiz → discussion forum → staged project. Consistent weekly routine designed to build engagement habits for learners managing work and other obligations.
Scaffolding	3-stage project. Stage 1: handwritten outline (two cultural sites, theory, predictions) → mandatory 1:1 instructor meeting per student to review outline, build trust, flag missing work. Stage 2: site visit + observations vs. predictions → group presentation with instructor feedback. Stage 3: APA-style paper.	5-stage project expanding Stage 1 into discrete checkpoints: (1) select sites, (2) identify theory + predictions, (3) site visit + observations, (4) limitations/caveats/implications, (5) final APA paper with revision history in appendix. Topic outline provided for each knowledge log. Instructor feedback within 1 week after every stage.
Assessment	Formative: per-lecture knowledge check (application + analysis, Bloom's levels 3–	No exams. Summative: 3–4 integrative reflection assignments (apply course theory to

	4). Summative: 3–4 unit tests → cumulative final (each unit test becomes practice for the next). Knowledge check questions recycled into unit tests; unit tests recycled into final. Major project: group cultural site visit (presentation + APA paper).	real video/documentary content — answerable only by engaging with assigned material) + 5-stage cultural site visit project (individual). 2 quiz questions per week answerable only by watching lecture — design disclosed to students at course outset.
Feedback	Immediate: verbal + written on knowledge checks; student questions earn citations from literature if well-reasoned. Post-test: group discussion of individual attempt, then instructor review of most-missed items in next lecture. Project: written feedback after each stage before student advances.	Audio/video instructor feedback on each of 5 project stages within 1 week of submission. Knowledge log: completion grade with spot-check for lecture-specific engagement. Instructor presence maintained through frequent, personalized feedback rather than synchronous contact.
Pacing	Activity break every 5–10 min; mandatory mid-lecture break (~40 min). Concept → example → student application before moving on. 3–4 unit tests spaced across semester; cumulative final.	Weekly rhythm fixed; daily pace student-controlled. Embedded reflection prompts in video lectures ('pause and write'). Stages released sequentially; students cannot advance without instructor feedback approval on prior stage.

FINDINGS

- **Early engagement predicts outcomes with high reliability.** Across two online async cohorts (N=65 with complete data), students who completed all Week 1 activities passed at 100%, compared to 47% for non-completers — a ~2.1× difference. This pattern held across both iterations, identifying Week 1 completion as the single strongest leading indicator of course success.
- **Process-based assessments measure the same constructs as traditional exams.** Correlational analysis of open-ended project scores vs. closed-note exam scores (in-person cohort) yielded r = .77; confirming that the staged project measured equivalent underlying competencies. This validated the online async assessment redesign as both AI-resistant and construct-valid.
- **Structural contact points close the trust gap.** Following introduction of mandatory 1:1 Stage 1 meetings with HS students, student-initiated contact increased substantially across the semester. Students who completed the 1:1 meeting asked more questions, flagged confusion earlier, and completed missing work at higher rates than in prior iterations without this structure.
- **Online async success rate exceeded department baseline by 19 points.** 85.2% success rate vs. 65.8% department online async average — sustained across both iterations, suggesting the design improvements were replicable, not incidental.

RESEARCH → DESIGN → IMPACT

- **Leading indicators beat lagging metrics.** Waiting for grades to flag disengagement is too late. Week 1 behavioral data enabled proactive outreach within Days 5–7: the equivalent of a real-time user signal triggering a product intervention.
- **Valid measurement requires context-aware design.** The same assessment format can measure completely different things depending on the surrounding system. In AI-accessible environments, output-testing measures tool access, not learning. Redesigning around process-documentation restored signal validity.
- **Instructor presence must be engineered.** For both populations (unapproachability in HS and silent disengagement online), the solution was identical: deliberately designed contact points built into the course structure, not added on top of it.

ARTIFACT: CULTURAL SITE VISIT PROJECT RUBRIC (PRESENTATION STAGE)

Used across both modalities (adapted for group vs. individual). Designed to assess application and critical thinking (Bloom's levels 3-4), not recall.

Criterion	Full Marks Description
Overview & Justification	Locations clearly described and thoughtfully selected. Comparison grounded in relevant cultural markers (ethnicity, language, norms), with a clear rationale for why the two are worth comparing. Includes a working definition of culture to frame the analysis.
Theoretical Application	Theory accurately summarized and appropriately applied to both locations. Predictions logically follow from the theory.
Culture Cycle & Predictions	At least two correct layers of the culture cycle identified. Predictions stem clearly from those layers and match the chosen theory.

Observations & Critical Thinking	Observations are vivid, specific, and clearly linked to predictions. Insightful evaluation of how observations match or contradict the theory, with reasoning and consideration of non-cultural explanations.
Reflection	Demonstrates thoughtful reflection on challenges, surprises, personal or group growth, and acknowledges possible caveats or limits in interpreting cultural differences.
Presentation Quality	Slides visually clear and organized; photos labeled and enhance key points; group presents cohesively and attentively.

Full course materials, knowledge log examples, assessment data, and additional rubrics available on request.

Building and Evaluating AI in the Classroom

A Two-Act Study: Deploying an AI Feedback Pipeline in Production, Then Evaluating What a Major LMS Got Wrong

Erica J. Yoon, PhD · College of San Mateo + Instructure (Canvas) · 2022–2026

MY ROLE

Act 1: Sole designer and operator of a ChatGPT-based AI feedback pipeline deployed across my online asynchronous courses at College of San Mateo, covering 640+ student submissions per term across 5 subjects.

Act 2: Invited by Instructure as a Canvas AI Early Access Evaluator (Summer 2025); assessed 5 AI features across product surfaces in active development; produced written efficacy evaluations delivered directly to the AI infrastructure team.

Methods: Iterative prompt engineering, systematic output evaluation, rubric-anchored calibration, batch processing design, mid-semester student survey, direct observation of student engagement patterns.

Tools: ChatGPT (GPT-4), Canvas LMS, Canvas AI Beta features (Rubric, AI Grading Assistant, Discussion Insights, Agent, Smart Search), mid-semester survey instrument.

ACT 1 — BUILDING THE PIPELINE

The Problem

Grading 640+ written submissions per term across online async courses was consuming ~53 hours per term, time that could not be reinvested in direct student interaction. The core challenge was not speed alone: feedback needed to be calibrated to expert rubric standards, specific to each student's work, and consistent across all submissions regardless of grading order. Generic AI feedback (the default output without careful prompt design) was not acceptable.

What I Built

- **Rubric-anchored system prompt:** Developed a locked system prompt embedding the full assignment rubric, grading criteria, and anchor examples (strong, moderate, weak responses with reasoning) before any student submissions were introduced. The prompt was locked prior to each grading session to prevent drift across submissions.
- **Batch processing architecture:** Rather than submitting one response at a time (which produced inconsistent outputs) I fed up to 10 submissions simultaneously, the maximum the system could handle. This forced the model to calibrate relative quality across a set of responses rather than scoring each in isolation, substantially reducing variance.
- **Iterative calibration:** Early outputs revealed systematic bias: scoring and feedback quality shifted unpredictably — sometimes skewing high, sometimes low. The root cause was that the model was treating each submission as new instructional context, effectively allowing student work to redefine what a good response looked like rather than holding to a fixed standard. The fix was architectural: locking the full instruction set (rubric, criteria, anchor examples, and scoring rationale) before any student submissions were introduced, and keeping that instruction identical across all submissions in a batch. This prevented the model from drifting based on what it had just read.

Failure mode → fix

Early outputs produced feedback that was either generically positive ("Great job addressing the prompt!") or inconsistently scored depending on what came before it in the batch. Locking the prompt before any submissions were introduced and requiring criterion-by-criterion justification with student-specific language citations corrected both problems across subsequent iterations.

Outcomes

- **~53 hours/term redirected:** Time previously spent on written feedback was reinvested in direct student interaction:office hours, 1:1 check-ins, and responding to student questions.
- **Student evaluation of feedback quality:** Mid-semester survey (formal, administered each semester for all online async courses) included questions on feedback helpfulness. Students consistently reported that feedback was specific and useful for their learning — with comments noting that the feedback helped them understand exactly what to improve before the next stage.
- **Evaluations held:** Course evaluations sustained at 4.7-4.9/5.0 across terms where the AI pipeline was in use, with 5.0/5.0 for course clarity and organization in Fall 2025.

ACT 2 — EVALUATING CANVAS AI

The Context

In Summer 2025, Instructure invited me as a subject-matter expert to evaluate Canvas AI features in active development across 5 product surfaces: the Rubric feature, AI Grading Assistant, Discussion Insights, Agent, and Smart Search. My evaluations were produced as written reports and delivered directly to the Canvas AI infrastructure team. The experience of having built and iterated my own AI feedback system in production made several product-level problems immediately visible.

Key Findings

Finding	Detail
Generic AI, not LMS-native AI	Most features replicated what ChatGPT already does, rather than leveraging Canvas's proprietary data advantage — engagement trajectories, submission histories, cohort comparisons, dropout signals. The competitive moat for an LMS-integrated AI is not the model; it's the data. Canvas AI was not yet using it.
Discoverability failures	Agent was hidden in the top-right corner with no onboarding or example workflows. Smart Search scope was ambiguous; unclear whether it searched files, pages, or assignment content. Instructors cannot adopt tools they cannot find or understand.
Narrow practical applicability	AI Grading Assistant only supported text-entry assignments; not file uploads, PDFs, or multimedia. Discussion Insights produced surface-level summaries that missed main ideas and did not work with video replies. Tools only functional under narrow conditions.
Lack of instructor control	Tools made automatic decisions without allowing instructors to guide output, directly limiting alignment with course outcomes. The Rubric feature generated criteria based on generic content patterns rather than instructor-specified learning goals (e.g., 'understand concept' vs. 'apply cultural theory to a real-world site'). No option to input keywords, anchor phrases, or weight criteria.
Missing transparency	No visibility into how AI decisions were made for rubric generation or discussion summaries. Instructors need to understand and override AI outputs, not accept them as black boxes.

Core implication from Act 1 → Act 2

Having built my own AI grading pipeline, I knew exactly what the failure modes looked like: inconsistent calibration, rubric misalignment, and loss of instructor voice. Canvas AI was reproducing those same failure modes at the product level — not because the AI was weak, but because the product design had not solved the instructor-control and calibration problems that make AI feedback actually useful. The fix is the same at both levels: lock the instructional parameters before the model touches student work.

RESEARCH IMPLICATIONS FOR AI-FIRST LEARNING PRODUCTS

- **Calibration is a design problem, not a prompt problem.** Consistent, rubric-aligned AI feedback requires architectural decisions (locking parameters before input, batching submissions for relative calibration, requiring criterion-level justification) not just better prompts.
- **LMS AI should leverage proprietary data, not replicate generic AI.** The value of an LMS-integrated AI is access to longitudinal behavioral data: engagement trajectories, dropout signals, cohort comparisons. Products that ignore this data and replicate ChatGPT are competing on the wrong dimension.
- **Instructor control is a measurement validity requirement.** AI feedback that instructors cannot guide or override will misalign with course-specific learning goals, producing feedback that doesn't measure what it claims to measure. Getting this right is both a UX and a research design problem.

Written Canvas AI evaluation reports, mid-semester survey instrument, and prompt architecture documentation available on request.

Does the Project Actually Measure Learning?

A Construct Validity Study of Project-Based Assessment in the AI Era

Erica J. Yoon, PhD · PSYC 100, College of San Mateo · Spring 2025

MY ROLE & METHODS

Context: Sole instructor and course designer for PSYC 100 (General Psychology), hybrid 16-week format. Designed both the Mythbuster Project and the Final Exam independently.

Research question: Does the Mythbuster Project (a staged, open-ended application assignment) validly measure genuine student learning, or does it measure effort and compliance separately from understanding?

Participants: N=37 students who completed at least 30% of total coursework (Spring 2025 hybrid cohort).

Criterion measure: Final Exam score; closed-note, comprehensive, covering recall through application (Bloom's levels 1–4).

Predictor measures: Aggregate Mythbuster Project score (Parts 1–4 combined); Lecture Quiz scores (used as comparison).

Analysis: Pearson correlation (r) between each predictor and the Final Exam; Williams' Test (Steiger's Z) to compare whether the difference between correlations was statistically significant.

BACKGROUND & DESIGN PROBLEM

In the AI era, the standard case for project-based assessment rests on a practical argument: projects are harder to fake than multiple-choice exams. But that argument alone doesn't establish that projects actually measure learning — it only establishes that they're harder to game. The more important question is whether a project score predicts the same underlying competency as a traditional exam. If it does, the project is a valid substitute. If it doesn't, high project scores may reflect effort, formatting, or AI-assisted polish rather than genuine understanding.

I designed this analysis to answer that question directly for the Mythbuster Project (a 4-part staged assignment requiring students to identify a psychological myth, find real research evidence, apply course concepts, and reflect critically) because I wanted to know whether the grade I was giving actually corresponded to what students understood.

FINDINGS

Finding 1: The project predicts exam performance significantly better than quizzes.

Measure	Correlation with Final Exam (r)	Statistical Significance (p)
Mythbuster Project (aggregate)	0.677	$p < .001$
Lecture Quizzes	0.428	$p = .008$

The project is a significantly stronger predictor of final exam performance than quizzes ($t(34) = 2.42$, $p = .021$). This is not a trivial finding: quizzes in this course were designed as retrieval practice: low-stakes, frequent, completion-graded. The fact that the project outperforms quizzes as a predictor of the exam suggests the project is capturing something closer to genuine mastery, not just study behavior.

Finding 2: Validity increases with cognitive demand (the staircase effect).

Breaking the project into its four parts reveals a consistent pattern: the correlation with the final exam increases as the cognitive demand of the task increases.

Project Component	r with Final Exam	Cognitive Task (Bloom's)
Part 4: Final Essay & Revision	0.690	Synthesis & Evaluation
Part 3: Critical Analysis	0.556	Analysis
Part 2: Article Annotation	0.367	Understanding & Summarizing
Part 1: Introduction	0.361	Recall & Foundation

The staircase pattern is not incidental as it directly mirrors Bloom's Taxonomy. Parts requiring recall and summarizing show weaker alignment with the exam; parts requiring analysis, synthesis, and evaluation show the strongest alignment. This is what a valid, well-designed assessment should look like: the higher-order components of the project are measuring the same higher-order competencies captured by the final exam.

INTERPRETATION & IMPLICATIONS

- **Quizzes measure learning behavior; the project measures learning.** The lower quiz correlation is consistent with their design role as retrieval practice; students use them to learn, not to demonstrate mastery. This is expected since quizzes are designed to help students learn, not to demonstrate mastery. The project captures the endpoint: what did the student actually retain and understand?
- **The staircase validates the staged design.** If the project were measuring effort or compliance rather than understanding, we would expect relatively uniform correlations across all four parts. The increasing pattern confirms that the project is working as designed: each stage building toward the higher-order competencies the course is trying to develop.
- **This matters most for AI-era assessment design.** The practical case for project-based assessment in AI-accessible environments is often framed as 'harder to fake.' This analysis adds a more important claim: the project is measuring the same underlying competency as a closed-note exam, which means it is not just AI-resistant: it is construct-valid. The grade reflects genuine understanding, not AI-assisted output quality.
- **Implications for learning platforms.** Engagement metrics and assignment completion rates are commonly used as proxies for learning on digital platforms. This analysis is a reminder that proxy validity requires empirical verification; a student can complete every assignment without the completion data predicting actual learning. Designing for valid measurement requires asking which behaviors actually correlate with the outcomes you care about.

Full dataset, assignment rubrics, and quiz instruments available on request.