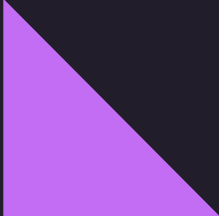




Ministry of Education

720



Data Point

Evaluating Complex AI Educational Programs



THE CHALLENGE

Evaluating Complex Educational Programs: When Everything Matters

Evaluating complex educational programs is inherently challenging

Multiple Delivery Methods

Abstract concepts
meet concrete delivery

Multiple stakeholders
(students → Ministry)

Limited timeframe
for scale-up evidence



CASE STUDY:

- Multi-platform personalized learning at scale
- 28 schools, ~300 staff, ~3000 7th graders
- 4 platforms, pilot timeframe

What We're Trying to Measure (& Why It's Hard)

The Measurement Tension: Comprehensiveness vs. Feasibility

What we would like to measure

Six agency components
(Self-Awareness, Planning, Independence,
Responsibility, Perseverance, Growth Mindset)

Deep teacher mentoring quality

Platform effectiveness & comparisons

School implementation quality

What is feasible to measure (at this stage)

Initial agency manifestations
(including SEL aspects)

Preliminary mentory capabilities

Program-level usage patterns

Implementation signals + sample quality



Our Strategic Framework – What We Chose

Trade-offs: Our 4 Evaluation Principles

1

Measure what
we actually
impact

2

Unified tools
(black boxes)

3

Multiple
measurement
strategies

4

Clear construct
boundaries

Our strategic framework – What we chose

Trade-offs: Our 4 Evaluation Principles

1

Measure what we
actually impact



Focus on initial,
observable indicators



Let go of long-term
manifestations

*Example:
Deep internalization
requires more time
than pilot allows*

2

Unified tools
(black boxes)



Treat different platforms
as “**black boxes**”



Measure **usage patterns**,
not features

*Students:
logins, time, completion*

*Teachers/Mentors:
login frequency*



No platform-to-platform
comparison



AI role: MoE-level
monitoring dashboards

Our strategic framework – What we chose

Trade-offs: Our 4 Evaluation Principles

3

Measure what we
actually impact

Quantitative:

- Self-assessment surveys
- Simulations
- Platform usage data

Qualitative:

- Simulations (open-ended)
- Interviews
- Classroom observations



AI role: Large-scale
analysis of open-ended data

4

Unified tools
(black boxes)



Define upfront what
counts as **agency**



Don't capture everything
under “personalised learning”



AI role: Create scales
and categories

What we're still wrestling with

Open Dilemmas: Where We Need Your Input



THE COMPARISON DILEMMA – When comparison is necessary but problematic

"**Semi-Black box**" approach risks overlooking differences

Mitigation: Examine Specific Pedagogical parameters and School visits, observations and case studies (10-15)



THE TIMELINE DILEMMA – When outcomes develop slowly but decisions need speed

Agency develops over time - **what's valid in a limited timeframe?**

Mitigation: Focus on initial indicators only



THE AI DILEMMA – When scale requires automation that can lack precision

AI expands capacity but **risks revealing biases**, not insights

Mitigation: Human review and iterative crosschecking

Key takeaway 1: Influence vs. interest

Lessons for Evaluating Complex Educational Programs

Measure what you can actually influence, NOT what's interesting

- Map what you **intend to influence**
- Design measurements around **causal pathway**
- Resist measuring everything that's "interesting"

In 720:

Mapped agency pathway → **measured initial manifestations** only using diverse **validation** tools



Key takeaway 1: Influence vs. interest

Lessons for Evaluating Complex Educational Programs

Fall in love with effectiveness, NOT with measurement

- The goal is **actionable insights**
- **Strategic trade-offs** are preferable to comprehensive impossibility
- Good evaluation supports **learning and decision making**, not just reporting

In 720:

- Avoided survey fatigue via:
- Brief questionnaires + simulations
- Auto-extracted platform analytics
- Observations
- AI-powered analysis

Result: Rich insights, no exhaustion

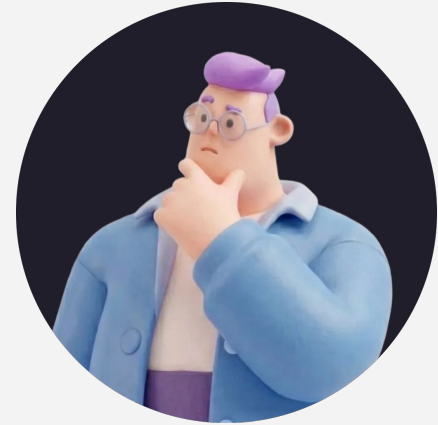


DISCUSSION

Lessons for Evaluating Complex Educational Programs

1

When rigorous evaluation demands multiple data sources but schools complain of "survey fatigue," how do you balance methodological quality with field burden - and what criteria guide your trade-offs?



2

What should we measure, what can we let go of, and how do we justify these choices to stakeholders?

Appendix 1: Scenario-based simulation example

Measuring Agency Through Authentic Learning Scenarios

How it works

Students respond to realistic learning dilemmas from their daily school life. Each scenario includes 3-4 open-ended questions that reveal their planning, decision-making, and problem-solving approaches.

Example Scenario - Summarised

Study complex biology topic for test in 2 weeks. Multiple resources available. Question: "How will you study this material?".

High Agency Response			Low Agency Response		
First, I'll watch a video to get the big picture. Then I'll read the textbook and take notes. After that, I'll practice with online exercises to check myself. If something's unclear, I'll use ChatGPT to ask specific questions."			"I don't know, maybe I'll read or ask a friend."		
AI Coding	Self-Awareness:	5 / 5	AI Coding	Self-Awareness:	1 / 5
	Self-Regulation:	5 / 5		Self-Regulation:	1 / 5
	Agency:	5 / 5		Agency:	1 / 5

KEY INSIGHT: Simulations capture actual decision-making behaviors, not just students' beliefs about themselves.

Appendix 2: Platform usage data

Measuring Effectiveness Through Usage Patterns

What we measure from the platforms?

Student Signals (Examples)

- Session length
- Task start, success and completion
- Task length
- Drop off rate
- Active usage

Teacher Signals

- Unique logins
- Usage of features
- Active usage

Dashboards with Actionable data for ministry level using AI

- Program wide trends
- Implementation signals
- Early warning indicators



KEY INSIGHT: Platforms as "black boxes" - we measure inputs (usage) and outcomes (agency), not platform features.

Appendix 3: AI-powered qualitative analysis

AI Coding: Enabling Depth at Scale

EXAMPLE: Student response to "What will you do if you get stuck?"

"If I don't understand, I'll find a different explanation - maybe a YouTube video or ask ChatGPT for an example. If still stuck, I'll mark which part I don't understand and ask the teacher or a friend. I'll also check if I missed something earlier that I need to go back to."

AI Analysis

- ✓ Self-Awareness: 4/5 - Recognizes gaps, identifies specific difficulties
- ✓ Self-Regulation: 5/5 - Multiple strategies, monitors understanding
- ✓ Agency: 5/5 - Proactive problem-solving, strategic help-seeking
- ✓ Growth Mindset: 4/5 - Views difficulty as solvable, persistence
- ✓ Support-Seeking: 5/5 - Strategic timing, specific needs

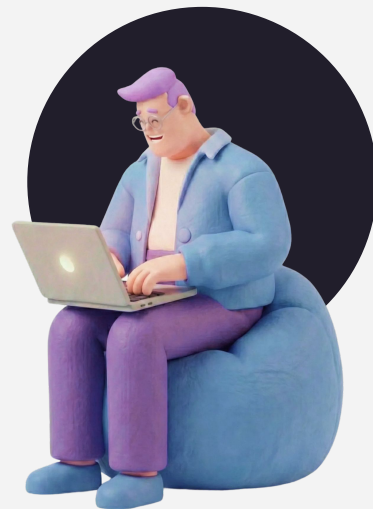
Scale and Efficiency

Traditional Coding: 4,000 responses = 400 hours, moderate consistency

AI Coding: 4,000 responses = 48 hours, high consistency

Validation: Human experts code 200 → AI learns → codes 3,800 → 10% validated → iterative refinement

Applied to: Simulations, interviews, observations, learning journals



RESULT: Comprehensive qualitative insights at quantitative scale.

Appendix 4: Ministry-level dashboard

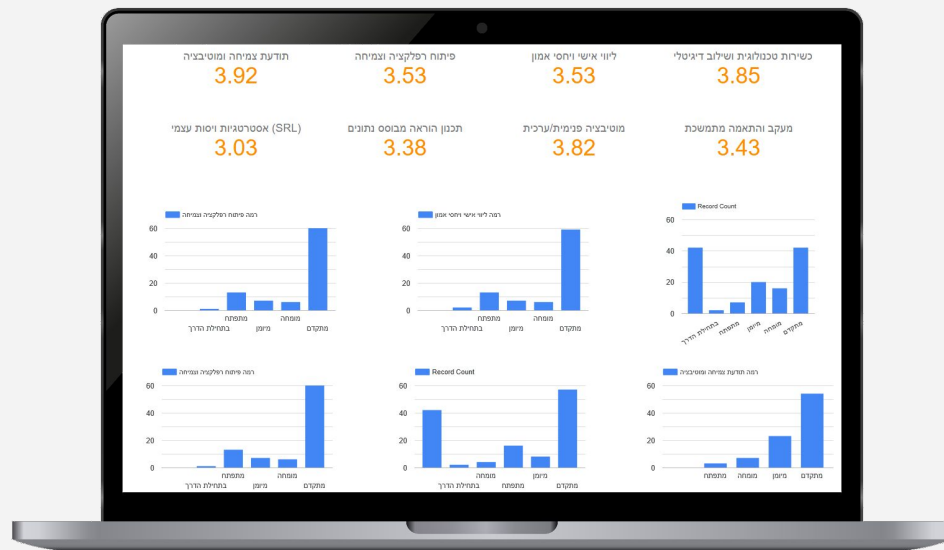
From Data to Insights in Real-Time

Teacher Mentoring Competencies

- Growth mindset development 3.92
- Reflective practice: 3.53
- Personalized instruction: 3.53
- Digital pedagogy: 3.85
- Self-regulated learning support: 3.03
- Data-driven instruction: 3.38
- Competency-based learning: 3.82
- Ongoing assessment: 3.43

Distribution Patterns

- Teacher competency levels (Beginner → Advanced)
- Implementation signals



KEY INSIGHT: School/Ministry-level view identifies professional development needs and effective mentoring practices.



Ministry of Education

720

Thank you

