Reward Follows Legibility

Why the education market can't reward what it can't see, and the infrastructure that changes that.

Eli Jameson, Founder, PILLAR

Response · June 2026 · A reply to *The $30 Billion Question* (AEI, 2025)[1]

KEY POINTS

- The report's diagnosis is the most honest from inside the system, but it stops one level short: it frames the fix as a bridge to fund and an evaluator to reform, and both quietly recreate the capture problem they set out to escape.
- The real failure is legibility at the moment money moves. Reward follows legibility. A falsifiable, independently computed, decision-point evidence layer, built once as shared infrastructure, dissolves the capture problem no committee can.
- This is a proof of existence, not a proposal. The evidence spine is built: seven states of real district adoptions on a national outcome ruler, the methodology fixed as falsifiable invariants, and the funding side now legible in the same query.

A note on method

This response was written from the builder's chair, not the policy desk, and it was developed deliberately as a triangulation. It takes seriously five vantage points on a single market failure: Mark Schneider's, from inside the federal evidence apparatus he once led; Auditi Chakravarty's, from inside the philanthropy trying to fund the gap; Matt Pasternack's, from the founder's seat of a company selling into the very market in question;[2] my own, as the founder of a company building the data infrastructure underneath it; and a fifth, synthetic vantage point that none of the four reaches alone, which this analysis works to hold in view throughout. The argument, and any errors in it, are mine.

Executive summary

Chakravarty and Schneider have written the most honest diagnosis of the education market I have read from inside the system. They are right that the market rewards visibility over impact; right that a "messy middle" of translation funding is missing; right that the What Works Clearinghouse failed by measuring the rigor of evidence instead of its usefulness; and right that schools need probabilistic, contextual evidence, "what works for whom, under what conditions", rather than binary verdicts.

I want to extend their diagnosis to its conclusion, because I think they stop one level short of it. The report frames the solution as a **bridge to be funded and an evaluator to be reformed**. I will argue that both framings quietly recreate the problem they are trying to escape, and that the real fix is neither capital nor a committee but **infrastructure**: a falsifiable, independently-computed, decision-point evidence layer, built once as a shared public good and then crossed by everyone.

The report leaves three questions unanswered, *mechanism, sequencing,* and *capture*, and I will answer each. And because a response report that only argues is just another white paper, I will close by showing that this is not a proposal but a **proof of existence**: the first load-bearing span of that bridge has been built, and it works.

The thesis in three words: **reward follows legibility.** Markets reward what they can see. Today they can see marketing budgets and incumbent familiarity. Make efficacy as legible, at the moment money moves, as those things are, and the market mechanism does the rewarding, with no kingmaker required.

I. Five chairs around one table

The reason this debate has not resolved is that each participant is looking at a different face of the same elephant, and each is right about the face they can see.

Schneider, the evaluator. From the chair that ran the Institute of Education Sciences, the failure is structural: federal R&D rewards process over results. Contractors are rewarded for navigating procurement, academics for methodological precision, agencies for continuity, and the WWC, the one body meant to tell schools what works, retreated into summarizing evidence strength rather than rendering useful judgment. This is insider testimony of the highest value. Its blind spot is the assumption that a fixed evaluator can be *made* useful, that the WWC's drift was a design error rather than the predictable end-state of any single, capturable arbiter.

Chakravarty, the funder-operator. From the chair that runs an applied-R&D fund, the failure is financial: research gets funded, commercial growth gets funded, and the translational middle, turning evidence into a usable product, piloting it, iterating with educators, gets funded by neither, "where most promising ideas die." This is true, and it is the report's strongest contribution. Its blind spot is that the prescribed remedy, patient and blended capital, plus convening, names the *what* without the *how*, and that a report co-authored by the CEO of the fund whose portfolio supplies its three success stories carries a structural conflict that disclosure notes but does not neutralize.[3]

Pasternack, the founder in the arena. From the chair of a company actually selling into districts, the failure is one of demand: buyers are "overwhelmed with thousands of products with overinflated efficacy data that they can't distinguish, limited budgets, and quickly changing mandates."[2] His instinct, that the fix is *opinionated* signal delivered to the people who buy, is the right instinct, and it is the one the report under-develops. His blind spot is that he is partly arguing against a caricature (the report already agrees the market is a demand problem and already indicts the WWC for being insufficiently opinionated), and that, as a founder, he has a natural preference for a world where you win by building rather than by being anointed.

Jameson, the builder. From my chair, the failure is one of *legibility and delivery*: the evidence the market needs largely exists, scattered across public catalogs, but it is not computed, not contextual, not independent of the seller, and not present at the moment and place where a purchasing decision is made. The job is therefore engineering, not curation. My own blind spot, named honestly: infrastructure optimism. Building the spine is necessary but not sufficient; observational efficacy inference is genuinely hard, selection bias is real, and the work is early.

The fifth chair, the synthesis. Held across the analysis is a simpler claim that none of the four chairs, alone, quite reaches: every proposed fix that routes through a *funded institution* or a *staffed evaluator* inherits the capture problem, and the only escape is to move the judgment out of any institution and into falsifiable, independently-computed infrastructure. Reward follows legibility, and legibility is a property you can engineer.

II. Where the report is right

Let me be specific, because the critique that follows is an extension, not a rebuttal, and it has to earn that.

The report is right that **the strongest market signal today is familiarity, not effectiveness**, that districts make "high-stakes adoption decisions driven by a tangle of misaligned incentives: procurement rules that favor incumbents… and a marketplace where evidence is hard to compare and often unavailable when a decision must be made." That single sentence is the truest thing in the document.

It is right about the **messy middle**, the translational valley between a research grant and a venture term sheet that neither instrument is designed to cross.

It is right about the **What Works Clearinghouse**: a body that "does not actually declare whether a program works," that "lost sight of its main purpose," is a precise autopsy of why neutral evidence-strength summaries do not move classrooms.

And it is right about the **shape of the evidence we need**: "a probabilistic understanding of what is likely to work for which students, in which settings, and under which conditions." That is exactly correct, and exactly the standard against which I will hold the report's own proposed remedy.

On all of this, there is no daylight between us. The disagreement is about altitude.

III. The reframe: a legibility problem, and a bridge that must be built, not funded

The report's organizing metaphor is a **bridge** from research to scale, and its prescriptions follow from that metaphor: fund the messy middle (philanthropy, SBIR, impact capital), and reform the evaluator (a more opinionated WWC, From-Seedlings-to-Scale, ART). These are good and I support them. But notice what kind of thing each remedy is. Each is **an institution to be funded or a body to be staffed**, and each therefore inherits the exact failure mode the report so carefully diagnosed in the WWC.

If the cure for an under-opinionated, capturable evaluator is a *more* opinionated evaluator that picks winners, you have not escaped capture; you have relocated it from the evidence-strength layer to the verdict layer, where the stakes, and the incentive to lobby, to teach to the rubric, to fund the convening, are higher. The report is rightly wary of any single funder positioning itself as the arbiter of what works, then proposes a system in which someone still declares the best evidence. Pasternack raises the same worry from the founder's chair, asking whether this is "foundations declaring themselves to be the best innovators." The someone is the problem, wherever you put them.

So I want to reframe the diagnosis one level down. The market does not reward what works because **what works is not legible at the moment and place where money moves.** A superintendent signing a purchase order in a budget window does not have, in front of her, an independent, contextual estimate of how this product performed for districts like hers. She has a sales deck and a familiar logo. The market is not irrational; it is *under-informed at the point of decision*, and it rationally rewards the only signals it can see.

That reframing changes the nature of the fix. A legibility problem is not solved by funding a bridge or staffing an evaluator. It is solved by building **infrastructure**: a shared, falsifiable, continuously-updated evidence layer that any buyer, founder, funder, or policymaker can query, computed from data the seller does not control, delivered into the decision itself. You build it once. Then the market does the rewarding.

The report, to its great credit, actually reaches for this analogy and then sets it down too soon. It invokes SpaceX: education R&D needs "new entrants that pressure the system to reform, not just compete alongside it." Yes, but the SpaceX lesson is not "a better-funded contractor" or "a new review panel." It is **reusable infrastructure** that collapsed the marginal cost of every future mission. The evidence spine is the reusable rocket. Fund a hundred more incubations and you have a hundred more bespoke exceptions; build the spine once and every future product is legible by default.

IV. The three questions the report does not answer

Read closely, the report's back half is a manifesto of worthy reforms, fix the WWC, rewrite the Federal Acquisition Regulation, broaden peer review, settle IP early, fund the middle, build local evaluation capacity, with no answer to three questions that determine whether any of it works.

1. Mechanism. The report says to "make evidence work for schools" and to render it probabilistic and contextual. It never says *how* evidence becomes that. By what process does a fact about a product's effect become a contextual, uncertainty-bounded estimate that arrives in a buyer's hands at the moment of decision? "Build the capacity" is a placeholder where a mechanism should be.

2. Sequencing. The reforms are presented as a flat list, but they are not independent, they have a dependency order, and the report supplies none. You cannot make evidence actionable before you can compute it; you cannot compute it before you have linked what districts bought to how their students did. A plan is an ordered thing. This is an inventory.

3. Capture. This is the deep one, and it is the report's own unanswered question turned back on its own remedy. The WWC was captured by neutrality; an "opinionated" successor invites capture by influence. Who runs the new, actionable evaluations? What stops a vendor from lobbying the rubric, gaming the trial design, or funding the convening? The report diagnoses capture brilliantly in the institution it is replacing and is silent on capture in the institution it proposes.

A serious response has to answer all three. Here is the answer.

V. The delivery gap: why even perfect evidence loses at the point of sale

First, an extension that both the report and Pasternack under-weight, because it determines whether any evidence layer actually changes behavior.

Suppose the perfect efficacy estimate existed and sat in a beautifully written report. It would still lose, because the K-12 purchasing decision is not made on the merits. It is **relationship-driven** (the rep who shows up), **RFP-gamed** (specs written around an incumbent), **contract-vehicle-gated** (you buy what's on the state or cooperative contract), **budget-cycle-bound** (the money is use-it-or-lose-it in a window), and **split-incentive** (the buyer is not the user is not the payer). Evidence that lives in a PDF does not survive contact with that machine.

This is why "convene the evidence and publish it," whether by a reformed WWC or by Pasternack's gathering of the 500 largest districts, is necessary but not sufficient. The signal has to be **injected into the buying workflow itself**, at the budget moment, on the contract vehicle, contextualized to *this* district's students and mandates, or it does not move the purchase. Legibility is not a publication; it is a delivery system. The evidence layer and the decision layer have to be the same layer.

This is also, I think, the most generous and useful reading of Pasternack's contribution. His real point is not that the report is wrong about supply; it is that evidence which never reaches the buyer at the decision is inert. He is right. The disagreement between him and the report is smaller than his post suggests, and the synthesis of the two is precisely: *opinionated, independent evidence, delivered into the decision*.

VI. The answer: a falsifiable, independent, decision-point evidence layer

Here is how an infrastructure approach answers the three questions, mechanism, sequencing, capture, concretely rather than aspirationally.

Mechanism. Efficacy is computed, not adjudicated. You assemble three layers on one identifier (the federal district ID that every U.S. school district already carries): a **treatment** layer, which district adopted which specific product, in which year[4]; an **outcome** layer, student learning results over time, drawn exclusively from independent government assessment data the vendor cannot generate or edit; and a **covariate** layer, the poverty, enrollment, funding, and demographic measures needed to build a fair comparison. Then you estimate a product's effect by comparing adopter districts to *matched* non-adopters, and you express the result the way the report itself demands: a probabilistic, context-stratified estimate, "associated with a gain of this size for high-poverty elementary districts in these states; insufficient evidence elsewhere", with its uncertainty stated and its observational limits attached. That estimate then rides into the buyer's workflow at the procurement window. That is the mechanism the report gestures at; it is buildable today because the inputs already sit in public catalogs.

Sequencing. The dependency order is not a matter of taste; the data dictates it. (1) Stand up the independent outcome-and-covariate spine on the district identifier. (2) Prove you can turn that spine into a calibrated, decision-point signal that backtests against reality. (3) Build the treatment layer, what districts actually adopted. (4) Run the matched-cohort inference. (5) Close the loop with post-adoption monitoring, so the estimate is a living thing that updates every assessment cycle, the part the WWC never had. You climb this ladder in order because each rung is the previous rung's prerequisite. That is what a plan looks like.

Capture, answered structurally, not procedurally. This is the heart of the matter, and it is where infrastructure decisively beats any institution:

- *Independent inputs*. The outcome variable is computed from government assessment results the vendor never touches. The single most corrosive failure mode the market has, "overinflated efficacy data", becomes structurally impossible, because the seller cannot fabricate a state's own test scores. A vendor can lobby a committee; it cannot lobby a state testing system. - *Falsifiable, versioned methodology*. The method is not a panel's judgment; it is written down as enforceable invariants, version-controlled, and gated by automated tests. You cannot quietly re-weight the rubric, because re-weighting it is a visible change that fails a test. Capture becomes a red build. - *No discretionary intake*. There is no "submit your product for review" form, no allowlist, no inclusion or exclusion path a vendor could influence. The engine scores every product-and-context pair that can be derived from public catalogs, the whole universe, or nothing. You cannot bribe a pipeline that has no door. - *Backtest as immune system*. Estimates are checked against held-out years. An estimate that has been gamed to flatter a product fails its out-of-sample test against next year's real outcomes. Drift away from the truth is not just detectable; it is detected on a schedule. - *Probabilistic, not binary*. The WWC's "works: yes/no" created one high-value gate worth capturing. A contextual estimate with an uncertainty band has no single lever; to game it you would have to game actual student outcomes across many independent jurisdictions, which is the same thing as actually improving learning.

That is the answer to the question the report could not answer about its own remedy. You do not prevent capture by choosing better people to run the evaluator. You prevent it by removing the evaluator, by making the judgment a falsifiable computation over independent data, with no door to lobby and a backtest for an immune system.

VII. This is a proof of existence, not a proposal

I am wary of white papers that answer a hard question with an architecture diagram. So this response does not end at theory. While writing it, we built the first load-bearing span, and in the work since, it has grown into a multi-state structure with a live inference engine on top.

We wrote the methodology down first, as twelve falsifiable invariants, before any estimate was computed, so that the system is honest by construction rather than by retrofit. (Outcome data must be independent of the seller; an estimate must be an interval, never a verdict; suppressed student-privacy cells must never be imputed; every estimate must publish a reproducible manifest of its inputs; the engine must compute the whole catalog universe with no discretionary intake; and so on.) Specifying the constraints before the computation is the opposite of how a capturable evaluator is built.

Then we built the spine. PILLAR now ingests the *actual* district-level curriculum each district uses, real selections rather than a state-approved menu, across **seven states, Nebraska, Massachusetts, Rhode Island, New Mexico, Oregon, Ohio, and Tennessee, more than 1,500 districts and several thousand program adoptions,** publisher and edition parsed automatically, every district matched to its federal identifier.[5] The outcome side is national: a codified, license-cleared Stanford Education Data Archive panel (15,591 districts across fifteen years) plus PILLAR's own NAEP-anchored achievement layer that reproduces it at r = 0.94, so the seller-independent outcome variable exists for essentially every district in the country. And the engine that turns treatment times outcome into an estimate is live, not notional: matched-cohort comparison and difference-in-differences, now with a parallel-trends pre-test and an event-study that let the method check its own identifying assumption and refuse when it fails, every result an interval with its poverty-confound flag attached, every invariant enforced by automated tests so the rubric cannot be quietly re-weighted.

We also built the surface that answers the report's question most directly. A cross now places each program's *state quality endorsement*, Louisiana's Tier 1/2/3, Texas's and California's approval gates, next to its *measured outcome* across the districts that adopted it. Where a state rates a program highly but its adopters' results do not follow, the divergence is surfaced, not buried. That is the $30 Billion question made answerable one program at a time, by computation over independent data rather than by a committee's verdict.

And the spine has begun to render the *funding* side legible in the same breath. From the same public catalogs, federal Title I and state categorical expenditure alike, the engine can now show which specific dollars at a given school are eligible to pay for a given intervention. When the evidence and the money become legible in one query, the signal finally arrives in the form the purchasing decision actually takes, which is the gap Section V argued no report on a shelf can cross.

This is real and it bears weight, and I will still not overstate it. The estimates are observational, not causal; at honest grade-band granularity most program-and-context cells *still* cannot statistically separate one product from another, and the engine is built to say exactly that rather than manufacture a verdict, which is the discipline the WWC never had. The treatment layer is seven states deep, not fifty, because district-level adoption data exists only where a state's own law or initiative created it; widening it is a question of state policy, not of engineering. But the central claim of *The $30 Billion Question* is that America "lacks a reliable bridge between research and scale," and that claim now has a standing counter-example: the bridge is built of public data and test-gated code rather than of a new institution, it spans seven states on a national outcome ruler with an enforced-by-tests methodology, and it carries load today.

VIII. What this asks of the field

If the diagnosis is legibility and the fix is infrastructure, then each actor in the report's drama has a clearer job than the report assigns.

To funders, including AERDF. The highest-leverage philanthropic dollar in education is no longer a fourth incubated product. It is the **public good of an independent evidence spine**, the infrastructure that makes every product, including the ones you did not fund, legible by default. Fund the bridge, not another crossing. This is also the cleanest answer to your own conflict problem: infrastructure that measures everyone, including your own portfolio, by a falsifiable method you do not control is the one form of evidence a fund can produce without a thumb on the scale.

To district leaders. Stop rewarding the best deck. Demand, and reward, decision-point evidence computed from your peers' real outcomes, and favor vendors willing to submit to independent, government-data-based measurement they cannot edit. Your purchasing power is the demand signal that makes efficacy worth investing in.

To founders, Pasternack's tribe. A legible market is the one you have been asking for. When efficacy is independently visible, "we have evidence" stops being a multi-year, self-funded gamble and becomes a queryable, fundable asset, and patient capital, which has been waiting for a path to distribution that is not acquisition, finally has one. The clearer path you want is downstream of legibility.

To policymakers, Schneider's world. The FAR rewrite, peer-review reform, and a fixed WWC all matter. But pair them with the one thing the report leaves out: an **infrastructure mandate**. The federal government already collects the assessment and directory data the spine runs on; the missing public good is the layer that links and computes it. That is a more durable legacy than another contract recompete.

And the moral frame, because it is the only one that matters in the end. Every year we wait for the next bespoke exception, the next Magpie, the next CueThink, hand-built and individually shepherded to scale, is a year in which millions of students learn from materials no one independently checked. Exceptions do not scale. Infrastructure does. The report says the students "deserve a system designed to reach them, not one that leaves their chances to exceptional circumstances." Exactly. So let us stop manufacturing exceptions and build the system.

IX. Coda: where we agree

The $30 Billion Question closes on a line I would sign without changing a word: "The bridge can be built. We know this because it has been built by people willing to fund and do the work of translation."

I agree, and I would add one sentence. The bridge does not need to be rebuilt for every product, by every foundation, in every state, forever. It needs to be built **once, as infrastructure**, falsifiable, independent of the sellers it measures, and present at the moment money moves, and then everyone crosses it. The report asked why the market doesn't reward what works. The answer is that it cannot reward what it cannot see. Make it visible, at the decision, by a method no one can capture, and the $30 Billion will start finding the things that work, not because we picked them, but because, at last, the market can.

We have started building. The invitation is open.

ABOUT THE AUTHOR

Eli Jameson is the founder of PILLAR, which builds revenue and evidence infrastructure for the EdTech and GovTech markets. This response engages the AEI report in good faith and with respect for its authors, whose diagnosis it extends rather than disputes.

NOTES

1. Auditi Chakravarty and Mark Schneider, *The $30 Billion Question: Why Doesn't the Education Market Reward What Works?* (American Enterprise Institute, October 2025). https://www.aei.org/research-products/report/the-30-billion-question-why-doesnt-the-education-market-reward-what-works/. Chakravarty is CEO of the Advanced Education Research & Development Fund (AERDF); Schneider is a former director of the Institute of Education Sciences and a nonresident senior fellow at AEI. All quotations of the report are from this text. ↩

2. Matt Pasternack, CEO of Once, public commentary on the report (LinkedIn, 2026). Quotations are from that public post; characterizations of his position are my own reading and any misreading is mine. ↩

3. The report discloses that one author is AERDF's CEO, and two of its three flagship examples (Magpie Literacy, CueThink) are AERDF-funded. I raise this as a structural matter, not a personal one; it is precisely the kind of conflict that an independent, falsifiable evidence layer is designed to make irrelevant, because such a layer measures a funder's own portfolio by the same method it applies to everything else. ↩

4. On the "adopted ≠ used" caveat and on independent, nationally-comparable district achievement measures, the relevant literature (e.g., RAND's American Instructional Resources Surveys; Stanford's Education Data Archive) is clear that formal adoption is a proxy for, not a guarantee of, classroom use, and that any efficacy estimate must carry that limitation explicitly. PILLAR's methodology encodes this as a required, non-removable caveat on every estimate. ↩

5. A note on the proof of existence: the treatment data is published by each state's department of education (Nebraska, Massachusetts, Rhode Island, New Mexico, Oregon, Ohio, Tennessee); the achievement outcomes are public government assessment data (the Stanford Education Data Archive and a NAEP-anchored owned layer); the linkage is on the NCES district identifier. The figures cited (seven states, more than 1,500 districts; a 15,591-district, fifteen-year SEDA outcome panel; an owned layer reproducing it at $r = 0.94$) are from PILLAR's own ingestion as of June 2026. The efficacy estimates the engine produces are observational and interval-valued, not validated causal results, and the response says so wherever it matters. ↩