Proceedings of Machine Learning Research 1:202–211, 2025 iRAISE Workshop at AAAI 2025

The LEVI Trialing Hub Evidence Matrix: Providing Progressive Measures of AI-Driven EdTech Research & Development

J.M.Alexandra L. Andres ALEXIS@LD-INSIGHTS.COM
John Whitmer JOHN@LD-INSIGHTS.COM
United States

Abstract

The rapid growth of education technology (ed tech) tools, including AI-powered applications, has highlighted the need for robust evaluation frameworks, particularly at early development stages. Current evaluation models, such as the Every Student Succeeds Act (ESSA) evidence tiers created by the U.S. Department of Education, may be appropriate for many education research activities, but miss critical stages in emerging AI-driven interventions. To support the Learning Engineering Virtual Institute (LEVI), a research collaboratory with the goal of doubling math learning rates in middle school students, we have developed a new evidence matrix to bridge this gap. This matrix incorporates a two-dimensional approach that evaluates research methods alongside outcome variables, enabling nuanced assessments of interventions along an ordered process. By categorizing research methods into five levels — ranging from randomized controlled trials to qualitative studies and modeling efforts, this matrix ensures comprehensive evaluation. Complementary outcome measures, emphasizing math learning gains, engagement, and model performance, contextualize these findings. This framework fosters alignment between research rigor and practical application, offering valuable insights into scaling educational innovations responsibly.

Keywords: education technology, evaluation, artificial intelligence, efficacy, math education

1. Introduction

Education technology applications are widely deployed in classrooms across the United States; in the 2022-2023 school year alone, an average of 2,591 unique ed tech tools were used per school district (Instructure, 2024). New applications are continually being invented, and recent Artificial Intelligence (AI) innovations have only increased the speed of innovation. However, it is much less clear how well these applications work to improve learning outcomes, especially in the case of AI. (Grant, 2024).

While there are widely-accepted approaches to evaluate stable, late-stage products (e.g., Randomized Controlled Trials), there is much less clarity about how to conduct these evaluations at earlier stages of product development. Given the potential risks in AI-powered solutions due to potential hallucinations and concerns about bias results for students from historically marginalized communities, education administrators, teachers, parents and other stakeholders have indicated that evidence-based evaluations are more important than ever (DOE, 2023).

2. The Learning Engineering Virtual Institute (LEVI)

The Learning Engineering Virtual Institute (LEVI) is a five year research and development program with seven research and development teams creating diverse education technology interventions with the goal of doubling the rate of math learning among middle school students. These teams were selected in 2022 through a competitive RFP process that included criteria such as advanced computational method innovation, education technology platforms that were widely deployed (or had the capability of being widely deployed), and incorporated a learning engineering approach. While these criteria and the program goals of doubling the rate of instruction were clearly defined, the type of intervention was left open to the creativity of proposers. The program is fashioned after a DARPA-type model of research and development, with active and ongoing management of the teams (Bonvillian and Van Atta, 2011). Teams are reviewed annually and grant renewals are contingent upon performance toward goals set during that year. The current list of teams and interventions follows below.

Carnegie Learning	**Carnegie Melon University**	**University of Colorado Boulder**	**EEDI**	**Khan Academy**	**Rising Academy**	**University of Florida**
MATHstream, Upgrade, Mathia	PLUS Tutoring	Saga, HAT	EEDI	Khan Academy, Khan-migo	RORI	MathNation

Table 1: LEVI Teams and Interventions

3. Background

Several taxonomies have been created to interpret and evaluate research methods and findings in interventional research. Each taxonomy is driven by a different set of assumptions to guide evaluation of research across various disciplines, including education (Midwest, 2023), medicine (Brighton et al., 2003), and psychology (Spring, 2007). The most widely used hierarchy in education is the Every Student Succeeds Act (ESSA), which was developed in 2015 to assist with state and district level decision making within schools (Act, 2015; Midwest, 2023). Guidelines within the ESSA provide a set of clear criteria of increasing robustness in evidence to identify the efficacy of education technology interventions. These guidelines are summarized into four tiers of evidence: 1) strong evidence, 2) moderate evidence, 3) promising evidence, or 4) demonstrates a rationale. Each tier is evaluated using five factors, including: 1) study design, 2) results of the study, 3) findings from related studies, 4) sample size and setting, and 5) match (similarity of the study sample to that of the school adopting an intervention). School improvements, if conducted within one of the first three tiers, are eligible for financial support by federal grant programs, however, evidence from any of the four tiers is sufficient for evidence-based adoption.

This guidance encourages the use of the strongest available evidence (more discussion is available in Herman et al. (2017) and Midwest (2023). Other notable frameworks include the hierarchy of evidence and evidence based practice. The "hierarchy of evidence" (Jamshidi and Pati, 2023), developed initially for healthcare research, proposed a method to improve upon the explicit definition of research area relevant methods, evaluation of accuracy and reliability, and comparison of evidence. The primary goal of this approach was to determine the ranking of studies based on the strength of internal validity (Jamshidi and Pati, 2023). The evidence-based practice hierarchy, also created for medical research (Force et al., 1979), has continually been referenced and adapted since its creation (Guyatt et al., 2001). Levels within the evidence-based practice hierarchy include 1) evidence obtained from at least one RCT, 2a) evidence obtained from well designed cohort or case-control analytic studies, 2b) evidence obtained from comparisons between times or places with or without the intervention, and 3) opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees (Elamin and Montori, 2012).

4. Fitting LEVI Evaluation with Interventions

The hierarchies and models listed above identify important criteria for research into the efficacy of completed interventions, but have limitations for education technology interventions in earlier stages. To address this gap, through a series of literature reviews, discussions with stakeholders, and iterative processes we developed an "evidence matrix" that provides a more complete set of evaluation criteria for the full set of analyses conducted in LEVI research studies. Our first step was the conceptualization and drafting of the categories. Drawing on the approaches referenced above, our team initially came up with a tiered approach to categorizing research.

This resulted in three categories: 1) Robust with Randomized/Quasi-Randomized Selections, 2) Moderately Robust With/Without Comparison Groups, and 3) Initial Evidence (descriptive analytics, anecdotal evidence, or simulations). However, upon conferring with teams and reviewing research study results, we found that this approach did not represent the depth of research nor did it represent the differences between studies with substantively different levels of effort and impact. For example, Team A had conducted a randomized experimental study (RES) on the learning outcomes of tutors using their platform. Their study had investigated the influence of feedback mechanisms and worked examples on e-learning outcomes while using their tutoring platform. Team B, on the other hand, conducted an RES as well but, instead studied the rate of learning for students across a school year compared based on student access to tutor chatbots. With the three tiered approach, the target audience and impact of the study results would be obscured and important differences of effort and significance in findings overlooked.

We worked with teams and reviewed studies to create a better approach that would be accessible across diverse audiences and stakeholders. After further reflecting on this process, we revised the approach into a two-dimensional matrix that included a) research methods and b) outcome variables, with a focus on investigating learning gains. Each category below serves as a guide to identify the trajectory of research activities as they build off one another towards the north star goal of doubling math learning gains. Specifically, studies and outcomes involving model testing establish the safety, validity, and accuracy

of algorithms while initiatives around platform-related and usage-related changes provide insights into engagement, usage, and other more generalizable "public goods" that can be adopted by future research and development efforts.

5. Detailed Descriptions of Research Methods & Outcome Variables

This section describes the different categories of research design methods, ordered according to the strength of the analytical method and its capacity to produce the most accurate estimate of the effect of an intervention. The categories below have been adapted from frameworks on the hierarchies of evidence (Brighton et al., 2003; Evans, 2003; of Minnesota, 2023)

5.1. Level 1 Randomized Experimental Studies

Studies in this category are intended to demonstrate clear distinctions or to isolate differences that may emerge from the use of a LEVI-funded platform, allowing for the identification of intervention outcomes isolated from spurious differences or confounding variables. Studies in this category are required to have comparison or control groups to which participants are randomly assigned and to have the demographic characteristics of their populations. This assignment should be made at the most fine-grained level possible, although in many cases classroom-level assignment is the best approach within the constraints of school requirements and what is feasible to conduct within an actual classroom. By designing studies whose control and treatment groups only differ by exposure to the intervention/platform usage, experimental designs isolate intervention effects from other factors both known and unknown that might impact student learning gains. Studies in this category feature experimental designs, usually conducted using a Randomized Controlled Trial methodology (Bhide et al., 2018). RCTs are regarded as one of the most reliable methods for evidence collection derived from their capacity to minimize the influences of confounding factors (Sackett, 1997).

5.2. Level 2 Quasi-Experimental Studies (QES)

Studies in this tier examine whether there is a likely causal relationship between independent and dependent variables. Similar to studies in the first tier, QESs require the use of a control or comparison group; however, these studies no longer require random assignment between experimental and control groups as this is often not feasible in naturalistic contexts or in large-scale deployments. Nonetheless, there are many robust methods that include a comparison group and provide significant insights (Cook and Wong, 2008). The combination of scale and naturalistic environments make these studies extremely valuable as a source of insights. Some examples of research methods within this tier include Quasi-Experimental Studies, Propensity Score Matching, A/B testing results, Time-Series Design, and Regression Discontinuity Designs.

5.3. Level 3 Analytic Studies

Studies in this category identify relationships between use of LEVI interventions and improvements in math achievement. Unlike Level 1 or Level 2 studies, these studies might

not have a control group or a complete understanding of the participant demographics and other characteristics to ensure that intervention effects are not caused by external factors or might have assignment to groups without randomization (Brighton et al., 2003). Some examples of research methods within this tier include Observational and Correlational Designs, Interrupted Time Series comparisons, and Statistical Comparative Analysis.

5.4. Level 4 Qualitative or Small-Scale Studies

Studies in this category include early-stage research initiatives and development. Qualitative and other descriptive methods applied in these studies are essential to offer important insights into how stakeholders interact with and perceive the LEVI interventions, and can uncover deeper factors that may affect the scaling and implementation (Bhide et al., 2018). Descriptive studies support exploration in relation to interventions or conditions and provide evidence to contextualize their implementation (Grimes and Schulz, 2002). These studies may be used as a "springboard" into more rigorous research that investigates the impact of these desired interventions on students. Some examples of research methods within this tier include Qualitative Analysis, Focus Groups, Descriptive Statistics, Case Studies, Usability Testing, and Pilot Testing.

5.5. Level 5 Modeling & Feature Usage Studies

Studies in this category use data from previous studies or simulated data sets that are used to develop statistical or computational models for LEVI platforms. They may also be used to investigate prior platform usage to provide insights about student learning behavior. These studies can also be used to demonstrate the accuracy of predictive models or provide evidence for the development and refinement of features, analytics, and interventions that will be deployed. Some examples of research methods within this tier include Model Testing and Validation and Comparative Analysis.

6. Outcome Variable Type

This section outlines the categories of outcome measures for use with the research plans and completed studies. These categories are arranged to offer a framework that guides the application of each outcome variable, enabling researchers to structure their study results in a way that substantiates evidence of learning gains. Additionally, the categorization maps to a progression from product improvement to increasing learning gains. This progression includes product improvement, increasing learning gains at the student level, and increasing learning gains at the classroom level. This progression is presented below beginning with outcome variables that are most directly connected to learning gains and moving towards those more relevant to product improvement.

- **Type A Learning Gains - Robust, Externally Validated Measures of Math Learning Gains** - These include assessments that have been developed and validated by an external organization with robust testing and validation. These assessments may involve testing of content knowledge beyond mathematics but should provide specific math scores to demonstrate concrete changes for the sample being studied. These

include results from a full test administration or subscale that has been validated for stand-alone use.

- **Type B Learning Gains - Modified, Externally Validated Measures of Math Learning Gains** - These assessments include externally-created and validated measures, but may include less than a complete test administration or validated subscale.
- **Type C Learning Gains - Internally Developed Measures of Math Learning Gains** - These include assessments developed by teams to provide evidence of math learning gains. The assessments should be rigorous and include pilot testing and other psychometric validation.
- **Type D Platform Engagement and Activity Changes** - These include metrics collected or qualitative reports from platform activity instrumentation or less robust platform-based assessments (e.g. productive discussion in tutorial sessions). These may include usage logs, interactions, time on platform, and performance on less rigorous measures of learning. Platform analytics include constructs beyond math learning gains, this may include measuring improvements in cognitive skills like problem-solving and critical thinking, changes in engagement and motivation levels, or developments in soft skills such as collaboration and communication.
- **Type E Model Performance** - These include performance metrics for statistical, predictive, or computational models that have been integrated into LEVI products. These measures allow teams to evaluate the effectiveness and accuracy of these models and demonstrate improvement over time. These may include the precision, recall, or accuracy of predictive models for learning outcomes, the robustness of statistical models in analyzing educational data, or the efficiency and scalability of computational models in processing large datasets.

7. The Matrix in Practice

The Trialing Hub Evidence Matrix has been applied to the evaluation of LEVI team research activities in the second and third year of the program. This section demonstrates some of the analyses conducted on activities started and completed by the first quarter of 2024. The results below include findings and studies from the 2023-2024 year that were not completed in time for a prior reporting period, leading to a longer performance period than a standard quarter.

Application of the categories in the matrix allows for flexibility in visualization and comparison of studies conducted and their trajectory over a period of time. When examining the research and studies conducted by teams, the projections in Figure 1 indicate that teams are on track to exceed the number of studies they conducted in the second year of the LEVI program. Based on the research plans submitted by teams, considerable increases can be expected in the number of studies conducted in tier 1 and 2 for year 3; in contrast to the increase in the number of randomized experimental studies to be conducted, a decrease in the number participants is expected within this research tier.

The matrix itself provides a foundation to examine a snapshot of activities across the different categories of research design and outcome variables. As can be seen in Figure 2,

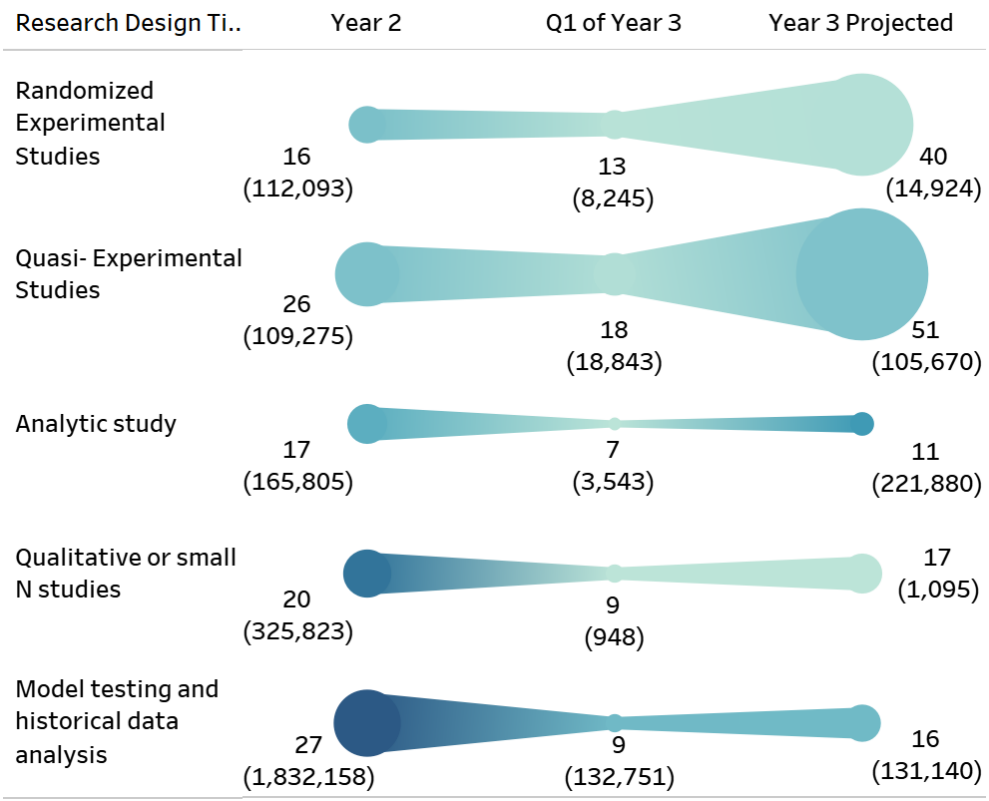

Figure 1: Synthesis of Studies for Year 2 and 3 of the LEVI Program, Sample Size (color) and Number of Studies (size)

model testing has the largest number of participants (81%, 132,829 participants), due to using large datasets collected using automated tools. Platform engagement and activity are the outcome variables in the largest number of studies (45%, 28 studies). Additionally, across all the currently reported studies completed and started, the majority of studies (54%, 34 studies) use quasi-experimental designs and Randomized Experiments. This is a positive finding as these studies compare the LEVI intervention to 'business as usual' to identify the true effect of the intervention, and are either conducted in the platform with standard users or are tested in well-controlled environments with detailed information about the users. All of these studies test for statistical significance. With these designs, studies demonstrate the generalizability of LEVI interventions.

8. Summary & Next Steps

Using these categories over the past year has enabled us to better represent the experimental activities conducted by LEVI teams and distinguish which teams are closer to readiness for an RCT-like research study from those that have additional research that needs to be conducted. This approach has also enabled us to provide more consistent (and robust) de-

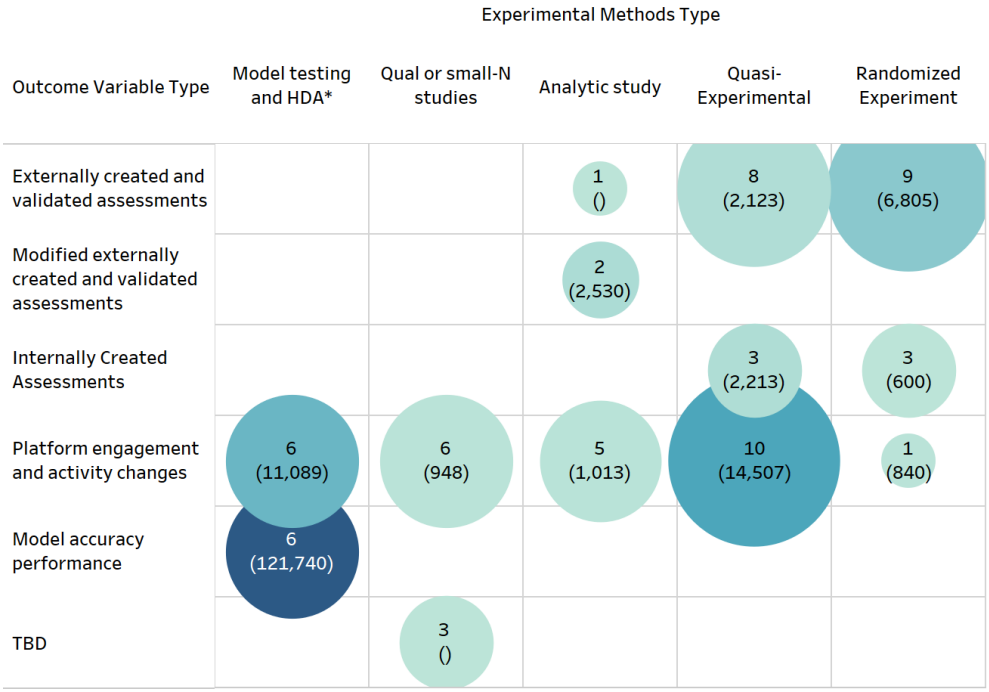

Figure 2: Studies Started and Completed for Year 3 of the LEVI Program, Number of Participants (color) and Number of Studies (size) for Each Method and Outcome Variable)

scriptions of research across the entire portfolio – while recognizing the variability between teams and projects.

Although this framework has improved consistency in describing and evaluating LEVI's research activities, there is still room for refinement. Key areas for future work include addressing the relative value of quasi-experimental designs versus randomized controlled trials and expanding the matrix to accommodate broader variables and contexts as well as expanding the use of the matrix to other research outside of math education. Ongoing discussions with researchers, educators, and stakeholders will help ensure that the framework evolves alongside technological advancements and educational priorities. LEVI's approach not only underscores the importance of iterative development and evidence-based practices but also emphasizes the need for continuous learning and adaptation to maximize the potential of AI-powered educational innovations.

References

Every Student Succeeds Act. Every student succeeds act. *Public law*, pages 114–95, 2015.

Amar Bhide, Prakesh S Shah, and Ganesh Acharya. A simplified guide to randomized controlled trials. *Acta obstetricia et gynecologica Scandinavica*, 97(4):380–387, 2018.

William B Bonvillian and Richard Van Atta. Arpa-e and darpa: Applying the darpa model to energy innovation. *The Journal of Technology Transfer*, 36(5):469–513, 2011.

Brian Brighton, Mohit Bhandari, Paul Tornetta, David T Felson, et al. Hierarchy of evidence: from case reports to randomized controlled trials. *Clinical Orthopaedics and Related Research®*, 413:19–24, 2003.

Thomas D Cook and Vivian C Wong. Better quasi-experimental practice. *The Sage handbook of social research methods*, pages 134–164, 2008.

U.S. DOE. Artificial intelligence and the future of teaching and learning, 2023. URL https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf.

Mohamed B Elamin and Victor M Montori. The hierarchy of evidence: from unsystematic clinical observations to systematic reviews. *Neurology: An evidence-based approach*, pages 11–24, 2012.

David Evans. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of clinical nursing*, 12(1):77–84, 2003.

Canadian Task Force et al. The periodic health examination. *Can Med Assoc J*, 121: 1193–1254, 1979.

Morrison J. Cook M. Reid A. Reilly J. Grant, A. A systematic review of evaluation methods in ed tech math, 2024. URL https://sree.confex.com/sree/2024/meetingapp.cgi/Paper/5213.

David A Grimes and Kenneth F Schulz. Descriptive studies: what they can and cannot do. *The Lancet*, 359(9301):145–149, 2002.

Gordon Guyatt, Holger Schunemann, Deborah Cook, Roman Jaeschke, Stephen Pauker, and Heiner Bucher. Grades of recommendation for antithrombotic agents. *Chest*, 119(1): 3S–7S, 2001.

Rebecca Herman, Susan M Gates, Aziza Arifkhanova, Andriy Bega, Emilio R Chavez-Herrerias, Eugeniu Han, Mark Harris, Jennifer Tamargo, and Stephani L Wrabel. School leadership interventions under the every student succeeds act: Evidence review. 2017.

Instructure. The edtech top 40: K-12 edtech engagement during the 2022-23 school year, 2024. URL https://www.instructure.com/resources/research-reports/edtech-top-40-look-k-12-edtech-engagement-during-2022-23-school-year?filled.

Saman Jamshidi and Debajyoti Pati. Hierarchy of evidence: An appraisal tool for weighting the evidence in healthcare design research based on internal validity. *HERD: Health Environments Research & Design Journal*, 16(3):19–38, 2023.

Regional Education Laboratory Midwest. Essa tiers of evidence, 2023. URL https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/blogs/RELMW-ESSA-Tiers-Video-Handout-508.pdf.

University of Minnesota. Evidence-based practice, 2023. URL https://pressbooks.umn.edu/evidencebasedpractice/.

David L Sackett. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier, 1997.

Bonnie Spring. Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of clinical psychology*, 63(7):611–631, 2007.