Empirical Education
EMPOWERING EDUCATORS THROUGH EVIDENCE AND INSIGHT

Design Memo

The Rapid Cycle Evaluations of the *Dynamic Learning Project*

Denis Newman

Val Lazarev

Andrew P. Jaciw

Empirical Education Inc.

March 25, 2020

Empirical Education Inc. (Empirical) is conducting a research project to get evidence of the impact of the Dynamic Learning Project (DLP), a Google-funded project that provides coaching to K-12 teachers[1] on the use of education technology (edtech). Digital Promise (DP) has led a research study looking at the impact of DLP on teachers' reported use of edtech and other differences between teachers with and without DLP coaching (Bakhshaei et al., 2019, 2018). DP's data consisted of surveys of teachers, coaches and principals as well as interviews and classroom observations. Empirical's work was done under a contract to DP and goes beyond DP's work in looking at student academic outcomes in addition to edtech usage patterns. Empirical's work goes beyond DP's educator-reported data in using statistical controls called for by the Every Student Succeeds Act (ESSA) of 2015 especially when considering impact of a program on student outcomes.

[1] In the initial year that this project investigated, DLP was implemented for middle-school teachers only. While it was later expanded to all grades, we may maintain a focus of the evaluation on middle-school.

This memo provides the rationale for the research design that Empirical Education is using. We are going beyond descriptive analysis in two ways.

First, is the research design and how it uses ESSA evidence tiers. In particular we address "mediators" or actions by teachers and students that result from coaching and in turn result in impacts such as on student achievement.

Second, we address generalizability or the conditions under which a decision-maker in one district can find useful information from a study conducted in different school districts. Here it is important to understand how the populations and resources of the study districts moderate the impact and the extent to which those "moderators" are found in other districts. For example, as a measure of relative poverty levels in different districts researchers often use student enrollment in the free or reduced-price lunch (FRPL) program. Where the impact of a program varies according to the percent of FRPL students, we say that FRPL moderates the impact of the program. We introduce meta-analysis or the combination of multiple studies as a method for reliably measuring the moderator effect.

Note that *moderators* are characteristics of schools, teachers and students that are measured at baseline, before the study begins while *mediators* are actions taken by teachers or students during the implementation of the program being studied.

Research Design to Meet ED Evidence Standards

We begin by addressing the tiered evidence standards developed by the US Department of Education (ED) and written into ESSA. This section will introduce three interrelated issues:

1. Selection bias. With a focus on the school district decisions, we address biases related to the fact that edtech products and services are adopted and implemented based on personal preferences and market forces and not for the convenience of researchers[2].
2. Mediators. We show how intermediate results, i.e., the different patterns of edtech usage in the classroom, can help both explain the student achievement outcomes, and help us statistically control for the choice that teachers made in deciding to be coached. that is, control for selection bias.

[2] In a randomized control trial (RCT), users and non-users are assigned to experimental groups randomly in advance in order to avoid any bias that may result from people being able to select their own experimental condition. In our studies here, the teachers have already chosen to be coached or not, and as researchers we had no control over this process. We take seriously the problem of avoiding bias resulting from this personal choice through using appropriate variables in matching coached teacher to un-coached teachers.

3. Generalizability. We address how generalizable results can be obtained. And, we illustrate how a research design that uses several ESSA tiers of evidence can be combined within a study of mediating factors.

While the usual goal of academic research is to estimate an average impact, we share one of the important goals of DP's research on DLP, that is, to help school district decision-makers decide whether or not to adopt DLP. We believe that these stakeholders need information about whether the program is likely to work in their situation, with their teacher and student populations. Focusing exclusively on the overall average impact misses important variation in results between student groups that can lead to the increase or decrease of important achievement gaps.

We call these studies rapid cycle evaluations (RCEs) following the usage in the ED contract to Mathematica for a powerful tool that helps schools analyze their own data to evaluate educational technology used in the school. In other words, RCEs help schools use their routinely collected administrative and outcome data to conduct fast turn-around evaluations of instructional programs. While the schools are not conducting the research in this case, our idea is to illustrate the use of RCE methods that place very low burden on the school system beyond providing already collected administrative data.

The ESSA Standards and Definitions for Evidence

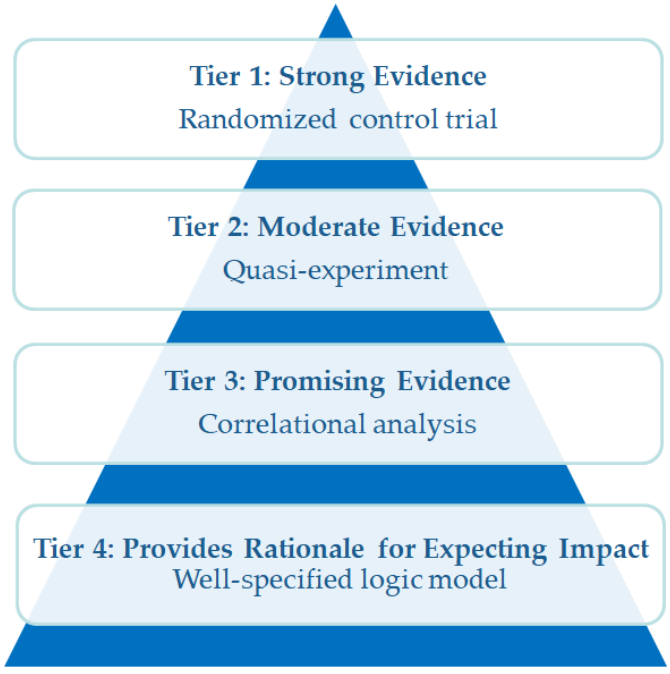


FIGURE 1. ESSA EVIDENCE TIERS

We think of the four tiers of evidence defined in ESSA as a pyramid. The base level or tier 4 is the expectation that any product should have a rationale for why it is likely to work. This is called for before any systematic research is undertaken. When we turn to our specific design, we start with a logic model displaying the causal connections that our studies are designed to measure. This study will be an illustration of how the tiers of evidence defined in ESSA as tiers 2, 3, and 4, fit together.

Each subsequent tier of evidence improves the "internal validity" rigor of the design. At tier 1, we find the Randomized Control Trial (RCT). At tier 2 we have the quasi-experiment (QE) or matched comparison study. Tier 3 provides evidence of promise through a non-causal correlational study. It is important to

understand that the hierarchy has nothing to do with whether the results can be generalized from the settings where the study took place and the district where the decision-maker resides (the so-called problem of "external validity"). As we explain in a later section, our study will attempt to demonstrate, through meta-analysis, how we can generalize useful information to other districts.

ESSA Tier 4: The Logic Model

We consider the logic model to be the basic rationale for the research—why we should consider the innovation to have any impact. The Empirical team, as an independent third-party, was not involved in the implementation of DLP. We refer here to the logic model (also called a "theory of change") developed by DP. The written description provided on the Digital Promise website explains in detail the way DLP coaching was implemented. Their logic model is shown in Figure 2.

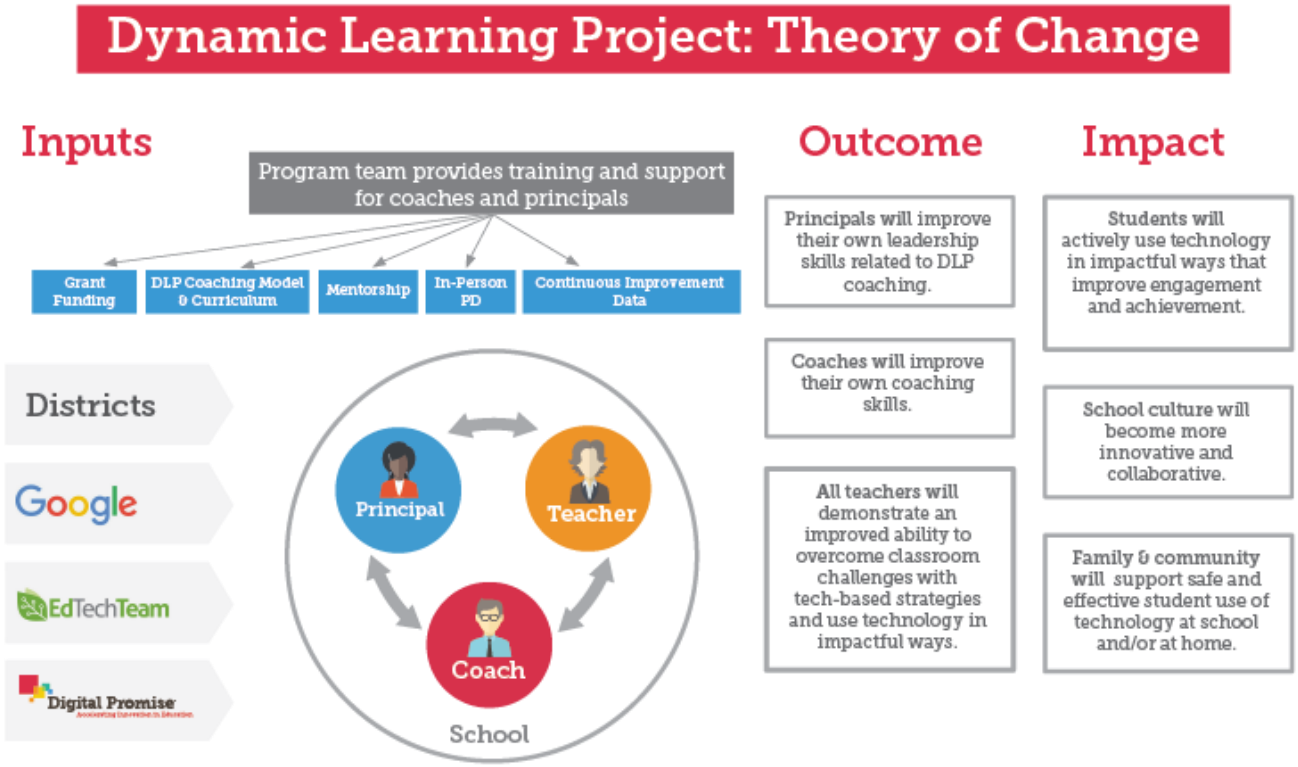


FIGURE 2. DIGITAL PROMISE'S DLP LOGIC MODEL

Inputs essentially describe the treatment. This is a complex process described in detail in the document and represented by the infographic showing the partnership among the coach, principal, and teacher. In considering potential for selection bias, we see the interactions resulting in the partnership are where the

enthusiasm for, and each teacher's selection into, coaching originates. The DP research collected surveys and interviews of the participants, which is the basis for their theory of action.

Our experiment requires identifying treatment teachers and matching them to teachers who were not coached, but, unfortunately, we were unable to use the detailed DP surveys because identifiers could only be provided for those selected into coaching and not for the larger set of teachers who did not choose to be coached. Although we had school rosters matching students to teachers and the demographics and test scores for students, we were unable to use the survey data for establishing the experimental groups. Fortunately, the DP team was able to identify teachers who received and did not receive coaching (although the non-coached teachers' surveys were not linked to the roster data.)

Outcome shows the immediate result of the Inputs. According to the DLP logic model the goal is to engender in teachers more Impactful Technology Use (ITU), which is about the use of technology to develop student 21st century skills. DP defines impactful technology use in terms of a framework where technology use is considered impactful when the teacher harnesses the tool to develop a student's collaboration, communication, creativity, critical thinking skills, and their agency.

Impact. Importantly, for the goals of this research project, the impact is shown under the right-most section: "Students will use technology in impactful ways that improve their 21st century skills and ultimately their engagement and achievement." This now points to a quantifiable research question. From this perspective, we can look at the use of edtech to be an intermediate outcome. Fortunately, we have been able to work with Hapara, whose tools were used in the classroom, which has been able to provide metrics associated with editing and collaboration on documents between students and teachers and other classroom activities that theoretically drive better writing achievement.

ESSA's Definitions for Tier 3

ESSA and the definitions provided in EDGAR give little specific guidance for tier 3. For tier 3, we are looking for a regression analysis that controls for "selection bias". (We assume that the selection bias requirement also applies to tier 2 although not stated there in ESSA.) Selection bias is important because, as has been pointed out by others, in a matched comparison (or quasi-experiment), such as the current study where teachers chose to participate in coaching, there may be unmeasured variables, technically "confounders" that affect outcomes. These variables are associated with the personal qualities that help explain their decision to pursue coaching and their motivation to excel in teaching. While recognizing in quasi-experiments (QEs)[3] there will always be,

[3] RCTs are often believed to be without bias (thus, "gold standard") but selection into the RCT can create a bias toward teachers or schools that are more willing to take risks and try new things. The results don't apply to less adventurous teachers. Also, most QEs are conducted where the agency has bought into the solution, while in many RCTs teachers receive stipends as part of the district agreeing to participate in the study. The bias to the average overall effect size resulting from the lack of buy-in can result in a lower effect estimate. These questions involve the generalizability of RCTs, an issue addressed by Tipton & Olsen (2018.)

by definition, an infinite number of unmeasured variables, we can, nevertheless control statistically for some important characteristics associated with selection into coaching. Motivation to use the edtech products can be controlled by amount of usage of a large set of products in a prior year of the products themselves. This control meets the ESSA requirement for Tier 3 (and 2).

We further note that the extent to which the remaining uncontrolled selection bias affects subgroup differences is not known. That is, we don't have data on the extent to which selection bias might change the effect size of the moderating effect of the subgroup characteristic. We also don't have a theory that would predict a difference in the effect for students in FRPL and non-FRPL based on the kind of coaching their teachers received.

The mediators and their relation to other outcomes is what schools and developers need to know about in order to improve the products and their implementation. This project's findings at this level, that is, those designed to provide "promising evidence" (tier 3) are offered primarily for helping DLP to improve the coaching and for districts to use in their own decisions about whether to support the implementation of coaching on the DLP model.

WWC's Definitions for Tier 2

Our design incorporates a set of QEs where the overall findings are intended to meet ED's What Works Clearinghouse (WWC) standards for QEs. The QE findings will fit the definition of moderate (tier 2) evidence in the Every Student Succeeds Act (ESSA). The ESSA law points to the WWC for criteria to use in determining the top two levels of evidence. Tier 1, strong evidence, calls for randomized control trials. Tier 2, moderate evidence, calls for "well-designed and well-implemented quasi-experimental [QE] studies." The only explicit requirement stated by the WWC is that the treatment and comparison groups in the study have to be equivalent at baseline (i.e., before implementation of the program or intervention begins) particularly, on the "pretest" or prior level of the outcome measure used in the study. A quasi-experiment by WWC rules (and common practice) allows for the identification of the treatment group according to pre-specified criteria providing the matched comparison group is equivalent at baseline.

None of the ESSA tiers address generalizability, to which we now turn.

Addressing Generalizability

ESSA and the definitions provided by the WCC do not address whether the information from a study can apply to districts other than the one in which the study was conducted. The approach we are taking is to provide moderate evidence about the DLP that will help districts in the process of deciding whether to implement this type of coaching. Our goal is to provide enough information for a district considering adoption to evaluate if the results generalize to their population. While addressing WWC and ESSA standards, we

caution, that a single RCE study in one school district, or even three RCEs in three school districts, may not provide enough useful information to generalize to other school districts.

The definitions provided in ESSA do not address how much information is needed from research studies to generalize from a particular study to what it will mean for implementing the program in other school districts. While we accept that a well-designed QE is necessary to establish an appropriate level of rigor, we do not believe a QE is sufficient to declare the program to have enough evidence to support implementation in another district. We note that the standards for Excellence in Education Research (SEER) recently adopted by the Institute of Education Sciences (IES), the research agency within ED, call for facilitating generalizability.

Our approach considers the generalizability of impacts for subgroups and the effects found through moderator analysis. Meta-analysis is a method for combining multiple studies to increase generalizability (Shadish, Cook, & Campbell, 2002). While our three sites may not support a full meta-analysis of this type, we can take it as far as it will go in producing generalized inferences about important quantities especially moderator effects across districts. We will demonstrate the approach by testing for stability of effects across sites and synthesizing those results, where warranted, based on specific statistical criteria. While moderator analysis is often considered exploratory and lacking in "statistical power," with meta-analysis, underpowered moderator analyses, can in combination provide confirmation of a differential impact.

This is especially the case, as in this study, where the QE is providing results at the student level, as well as for student subgroups. Each of these multiple subgroup findings will be essential for adoption decisions, where the schools want to know if the program is likely to work for their population of students and teachers. We can begin to address the question of generalizability by using how the project works within three districts with different characteristics.

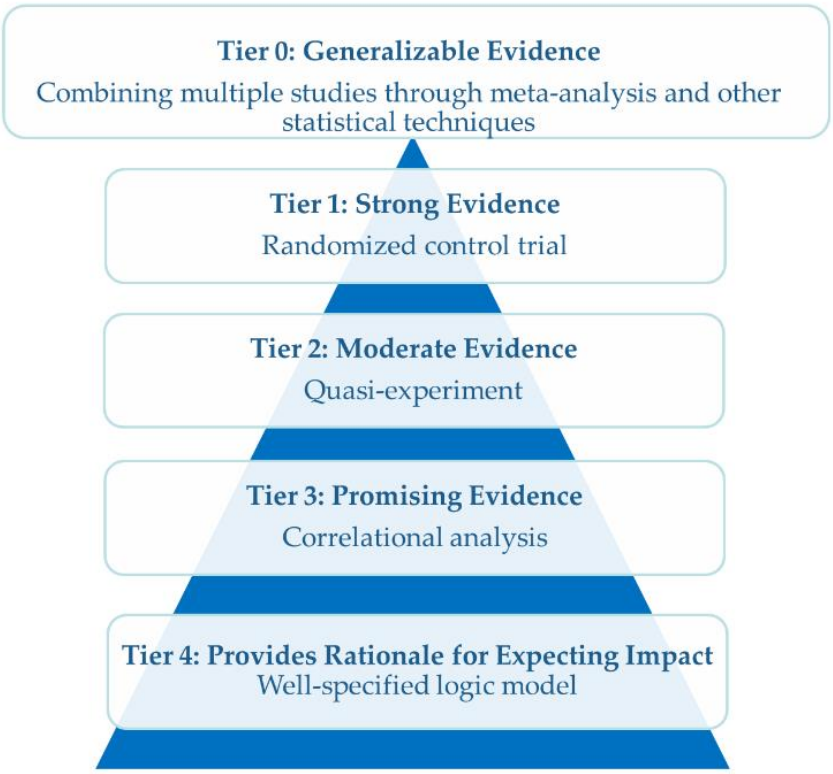


FIGURE 3. ESSA EVIDENCE TIERS TO INCLUDE GENERALIZABLE EVIDENCE

The general point we want to make about tier 2 moderate evidence is that findings from multiple studies can be combined through meta-analysis. While WWC does not consider subgroups to be reviewable results, such findings are nevertheless common in quasi-experimental results and of value in meta-analysis. While

cautioning about the limits to generalizability of single RCEs, we believe that our results can play a valid role in contextually-guided meta-analyses to provide generalizable evidence.

Our team would suggest that meta-analysis be a tier above the RCT in the ESSA pyramid (assuming that the studies going into it have an adequate level of rigor). Figure 3 suggests the new ESSA pyramid. This brings generalizability into the pyramid, which so far is absent from ESSA and WWC yet absolutely central in school decisions. We urge product developers and the schools considering implementing a new program to consider *generalizability* of the findings. This should focus on the generalizability of subgroup impacts, i.e., of moderating effects of subgroup characteristics, not just the average effect, which is less relevant to school decisions.

In addition to combining studies of the same program through meta-analysis of moderator effects, we will need a tool (similar to that developed by Tipton and colleagues that uses the demographics of the target district to generate a applicability metric.

Design for Our RCEs on DLP

We are conducting three QEs, each in a different school district: Lexington, SC, Talledega, AL, and Carrollton-Farmers Branch, TX[4].

Research Approach with Edtech

The Empirical team will make use of district-managed and commercially-generated data on the overall amount of edtech usage. These data will be available for all students and teachers, not just those taught by coached teachers. With this we will develop a measure of the intermediate effect (what we will designate as a potential mediator.) This measure will also be available as a pre-test at the teacher level that can be used in matching the treatment and the comparison groups (i.e., coached and not coached teachers) and addressing the potential selection bias issue—the teachers who are motivated to be coached are already more active in using edtech. With matching, the two groups will be equivalent with respect to motivation to use edtech.

Research Questions

The studies will address the following questions:

1. **Does receiving coaching result in improvements in student achievement?** This is the average overall impact question and is stated broadly as called for by WWC. Note that each of the three experiments will have results that can be combined to provide more broadly generalizable findings.

[4] Since intermediate outcomes are not available for this district, we are able only to test to direct link between Inputs and Impact on students.

2. **Does receiving coaching result in changes to the intermediate outcomes?** These outcomes will be defined at the teacher and student levels. We are fortunate to be working with Hapara on collecting the intermediate outcomes.
3. **Is the measure of usage** (the intermediate outcome or mediator**) correlated with the student achievement measure?** This will use standard regression and econometric techniques.
4. **Do the intermediate outcomes serve as mediators of the relationship between coaching and the student achievement outcome?** With questions 1, 2, and 3 combined, we have a method for providing evidence that coaching resulted in changes in edtech usage that is responsible for the effects on achievement found in question 1. The correlation between the usage metrics and the academic outcomes is tested (across all DLP and non-DLP classes.) If DLP predicts the usage metric(s) and if the metric is positively correlated with achievement, then we can say there is (promising) evidence that DLP results in academic benefits.
5. **For all results found in 1, 2, or 3, does the effect vary by subgroups?** This is the information most valuable to school district decision-makers. The moderation of the effect will be tested for the following subgroups[5]:
    a. Poverty: Students eligible for free or reduced-price lunch (FRPL)
    b. Ethnic/race category (for simplicity: white vs. Hispanic vs. Black)
    c. English learner status
    d. Gender: is the impact more pronounced for girls or for boys?
    e. Teacher characteristics: Data available will vary by district.

IDENTIFYING THE TREATMENT AND COMPARISON GROUPS

The district decision to join DLP put into motion a selection process by which some teachers adopted coaching and others did not. Some of this rationale was collected in surveys that DP conducted with principals, coaches, and teachers. Since the teachers receiving coaching were not identified prior to implementation, it is necessary to retrospectively identify teachers who received coaching, based on the surveys and coaching logs[6]. The coaching model was identified from the surveys using only questions about the coaching itself. This was the responsibility of the DLP team since the expertise in the model needs to inform the formation of the treatment group. Empirical will use standard methods for matching DLP-coached teachers to uncoached teachers once the treatment teachers are identified.

[5] These are all characteristics of students (and teachers) measured at baseline, i.e., before implementation of the program begins. These variables are commonly measured by school systems to be in compliance with US federal regulations. We note that the EdTech Genome project, seeks to define and provide measures for characteristics of districts, schools, teachers, or students that they believe will help explain success or failure of edtech implementations. If they succeed in validating them and making them widely available, such variables could be used in QE matching and in moderator and mediator analyses increasing the accuracy and generalizability of efficacy findings.

[6] Unfortunately, it was not possible to identify non-DLP teachers so the surveys could not be used in matching for the QE.

Identifying Mediators

A key characteristic of this study is the investigation of the intermediate outcomes, not just what conventional evidence standards count as the ultimate student achievement result of edtech implementation. Since the DLP coaching was not designed to improve test scores in specific areas, the mediator results are particularly important. They define a two-step process where the coaching results in changes in practices and capabilities and these lead to changes in student success. We will use Hapara's analysis of the usage of Google's G-suite within Google Classroom for the intermediate outcomes. Hapara's data are associated with individual students and teachers providing both teacher data and the student data on edtech usage that can be correlated with outcomes, especially writing outcomes, where available, the particular target of the Hapara analysis. When these data are available for the year prior to coaching, they provide a variable that can be used in matching.

How Our Design Combines ESSA Tiers 4, 3, and 2

Here we use a simplified logic model to illustrate the research design that links to the questions being asked.

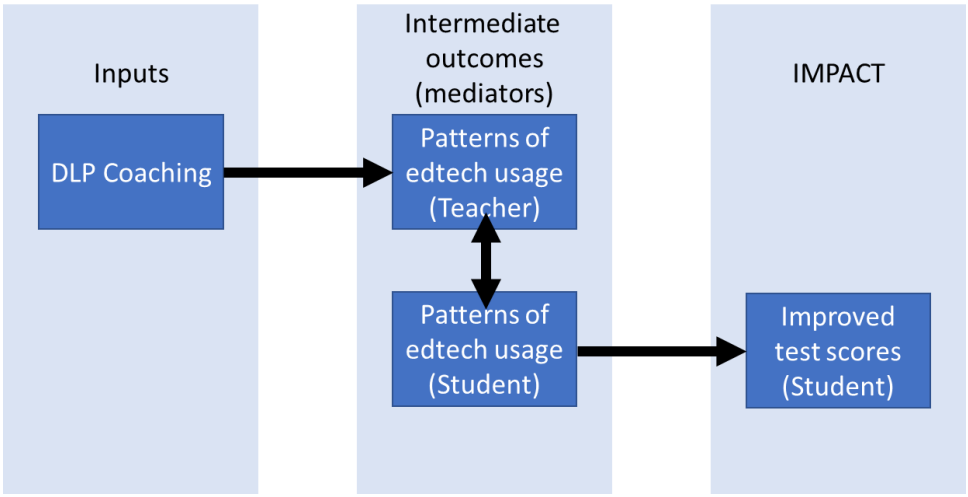


FIGURE 4. SIMPLIFIED LOGIC MODEL

Here DLP Coaching impacts the teacher, who may provide different experiences in interaction with students who benefit in ways that result in higher test scores. So we have an overall question, shown as a yellow arrow in Figure 5 and expressed as question 4: "Do the intermediate outcomes serve as mediators of the relationship between coaching and student achievement?". This is the question behind the DLP logic model's expectation that students will "actively use technology in meaningful ways that improve engagement and achievement."

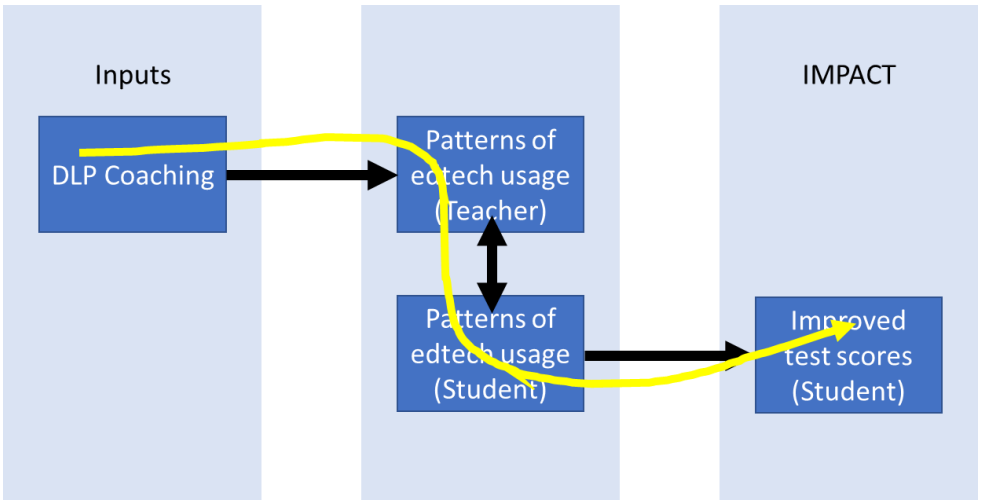


FIGURE 5. THE BROADEST QUESTION ABOUT DLP

Questions Answered by the QE

Tier 2 of ESSA is about showing that the elements on the left of the logic model cause something to the right through a matched comparison study. In Figure 6, the red lines represent the causal links that will be tested in the study. The first step for DLP is that some "teachers receive ongoing coaching". That's the initial cause of improvements. Tier 2 applies to all the quasi-experimental findings including questions 1 and 2. Understanding that the findings from a single RCE cannot be used to generalize, it is nevertheless valid as a contribution to looking at results across studies and potentially combining them to support generalization.

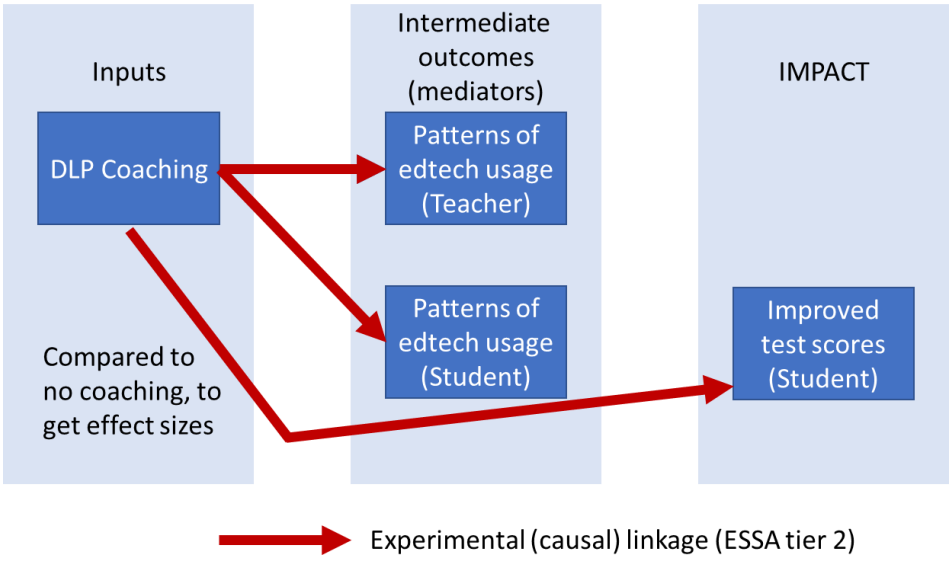


FIGURE 6. THE TIER 2 QUESTIONS

Answering the Correlational Questions

Tier 3 of ESSA evidence levels applies when a well-designed comparison group is not possible but where the association between a program or product use provides an important piece of promising evidence. Where tier 3 applies in this project is the intermediate outcomes (edtech usage) and their relationship to the final impact. This is shown in Figure 7. The intermediate outcomes are defined in terms of more-or-less usage by a teacher's classroom (as measured by the Hapara analytics.) Establishing the link between the intermediate outcomes and the long-term student outcomes is necessary if we want to show the relevance of an intervention and justifies the status of usage metrics as intermediate outcomes (mediators) but establishing causality is not required at this stage. This is addressed in research question 3.

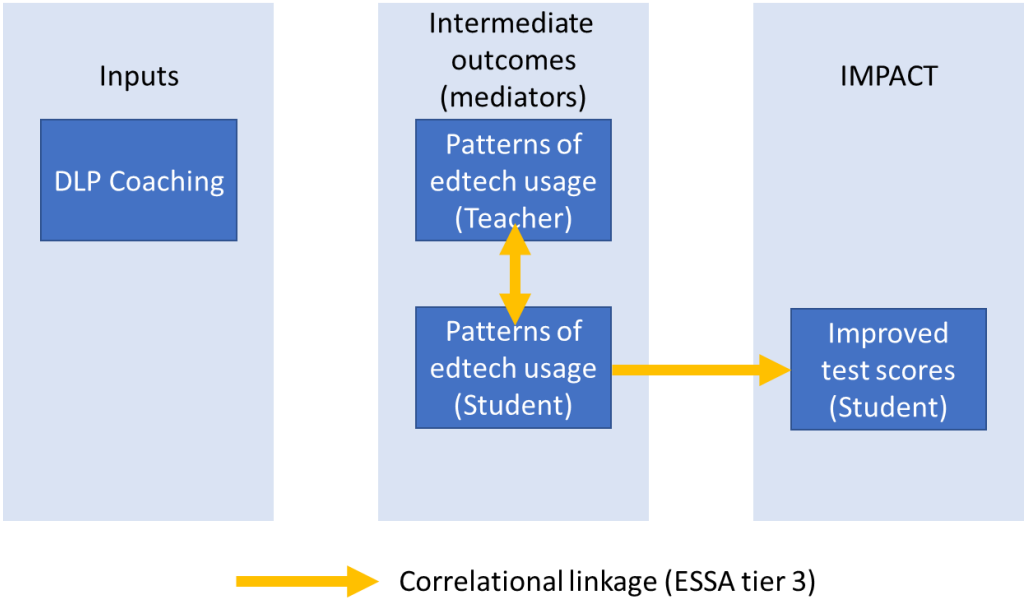


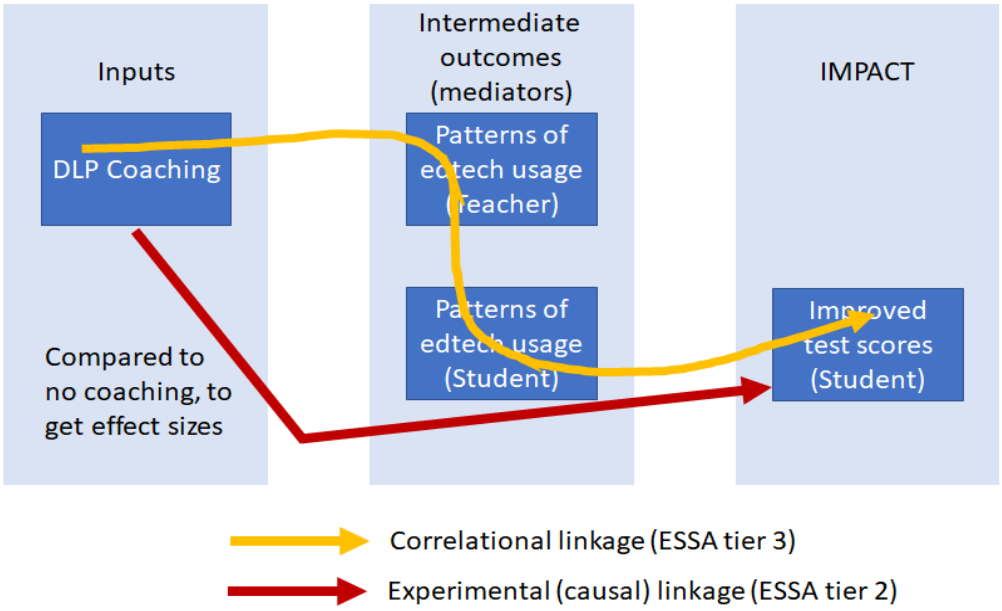
FIGURE 7. TIER 3 LINKAGES

Combining Figures 6 and 7 we have all the linkages among the levels of the logic model from Inputs to Impact. Since question 4 (Figure 5) depends on the results of question 3 (Figure 7), it is also a tier 3 inference[7]. Tier 3 is appropriate for a validation of a specific linkage and is a step above descriptive.

Answering the Impact Question

We now have two routes to answer question 1: "Does receiving coaching result in improvements in student achievement?" We see in Figure 8 that we can answer this question at tier 2 or at tier 3. If the Red arrow in Figure 8 does not show a positive effect then we can still ask whether the mediator is correlated with the outcomes.

[7] While there are experimental solutions to the question of whether the mediator is part of the causal chain leading to the student outcome, we are not using these approaches here because of the sample size requirement.



Tier 4 of ESSA regulations specify that the elements counted as intermediate outcomes are named in the logic model. The base on which any study stands is the rationale for why there should be an impact that is shown in the logic model.

FIGURE 8. TWO ROUTES FROM COACHING TO IMPROVED TEST SCORES

Conclusion: Efficacy Research in the Age of ESSA

The Empirical studies of DLP operate within the federal legislation and illustrate solutions to a number of problems.

- We show that it makes sense to combine different tiers of evidence in the same study.
- We show that it is essential to introduce generalizability and what that means for decision-makers in a school where their local conditions, not an abstract national average, is what is relevant.
- We believe the studies will illustrate how data from edtech usage can provide clear definitions of mediating implementation variables. These patterns of usage can (when used as moderator variables from the prior year) provide both an essential statistical control and data available for matching.

The Empirical team has great respect for the definitions of evidence tiers incorporated into ESSA. The tiers provide developmental stages that give any developer and any school a place to start. At the base, tier 4, you should at least know why the product, program, or policy should be expected to work. Tier 3 gives a simple correlational, low-risk approach that can be a starting point for data analysis. Tiers 2 and 1 are the experimental methods that can provide causal evidence. What's missing from the ESSA tiers is clear information on whether you can generalize from a study to the conditions found in a particular district, where the decision-maker resides. We are pleased with the growing attention being paid to how to make the leap from evaluation studies to a decision.

Reference this report: Newman, D., Lazarev, V., & Jaciw, A.P. (2020, March). *Design Memo: The Rapid Cycle Evaluations of the Dynamic Learning Project.* San Mateo, CA: Empirical Education Inc. Retrievable from https://www.empiricaleducation.com/blog/empirical-describes-innovative-approach-to-research-design/