ELEVATING EVIDENCE

EMPOWERED BY DATA:

Evaluating Intervention Impact

Kentucky Department of Education

Our Children, Our Commonwealth

Office of Continuous Improvement and Support

Introduction

The Every Student Succeeds Act (ESSA) requires schools to ensure that improvement initiatives are rooted in "evidence-based activities, strategies, or interventions." ESSA divides evidence-based activities into four differing levels. The guidance in this handbook corresponds to *Level IV: Demonstrates a Rationale,* which is defined in Section 8101(21)(A) of the ESSA as evidence "based on high-quality research findings or positive evaluation that such activity, strategy, or intervention is likely to improve student outcomes or other relevant outcomes and includes ongoing efforts to examine the effects of such activity, strategy, or intervention."

Schools wishing to utilize interventions that align with Level IV evidence are tasked with collecting data to support their continued use of the intervention. Ideally, this data should be collected in an academically rigorous way to produce *Level III: Promising Evidence* or higher. This document provides general guidance for schools seeking to collect data to support their use of an intervention.

This guidance divides the evaluation of an intervention into four distinct phases:

- Phase I: Creating a Logic Model
- Phase II: Gathering Evidence
- Phase III: Evaluating Effectiveness
- Phase IV: Making Educational Decisions

During each phase, educational leaders will take the necessary steps to select, monitor and evaluate the long-term effectiveness of an intervention on their students. The protocols outlined in this guidance are unlikely to produce research findings that are as rigorous as those published in peer-reviewed publications; however, the steps outlined within this guidance should provide educational leaders with a better understanding of the impact of an intervention in their school.

While this handbook provides a brief overview of some of the more common ways to evaluate intervention effectiveness, it does not provide an exhaustive discussion on the topic. There are many evaluation methods aligned to the federal regulations that are not mentioned in this guide. The Kentucky Department of Education (KDE) encourages schools to collaborate with community partners who can assist with the rigorous monitoring and evaluation of new interventions.

As you work through the evaluation process, you may find it beneficial to consult other pieces of guidance listed on the KDE Evidence-based Practices webpage.

Phase I: Creating a Logic Model

The first step in evaluating the impact of an intervention is to develop a logic model. A logic model is a "well-specified conceptual framework that identifies key components of the proposed process, product, strategy, or practice (i.e. the active "ingredients" that are hypothesized to be critical to achieving the relevant outcomes) and describes the relationships among the key components and outcomes, theoretically and operationally (34 CFR 77.1)." In short, the logic model provides a mechanism for the careful analysis of the various parts of an intervention.

Ideally, the logic model will be developed well before the implementation of an intervention begins, but it can also be a useful tool for reflection, refinement and future planning.

A logic model is crucial when evaluating the impact of an intervention on a student outcome or related outcome. It ensures that education leaders are able to view the intervention in its entirety and systematically evaluate the relevant pieces of the intervention. For example, a logic model may help leaders to identify relevant changes that caused an intervention to stop working or identify why an intervention had greater success in one year over another.

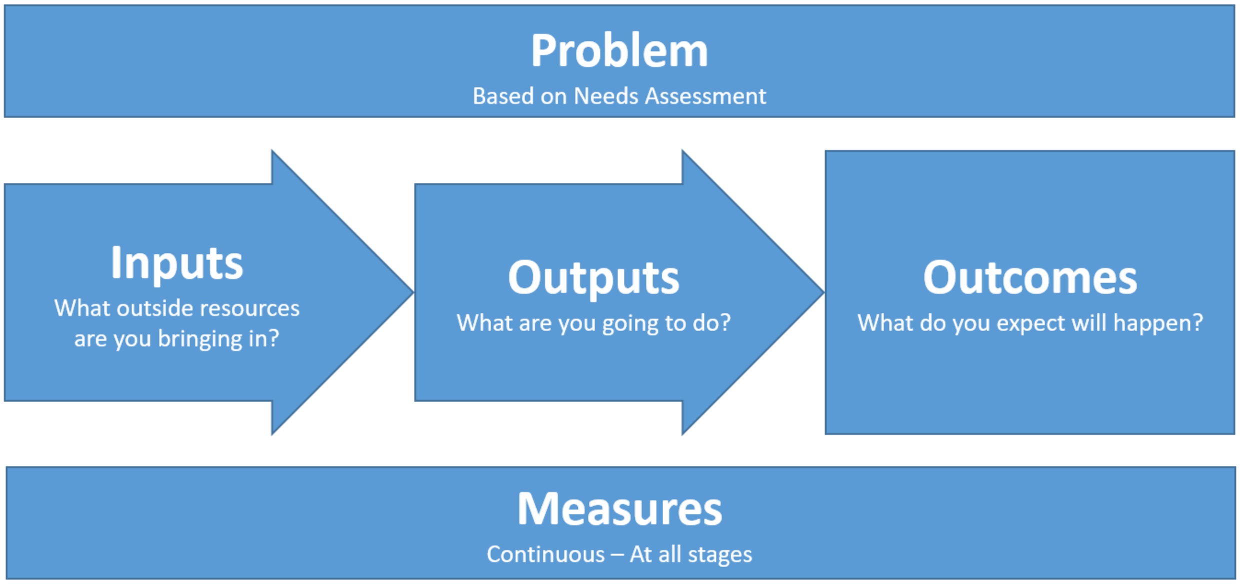
While there are many logic model frameworks available in the literature (Alter & Murty, 1997; Blase, Fixsen, & Jackson, 2015; Hernandez, 2000; Lawton et al., 2014; Stegemann & Jaciw, 2018), a basic logic model includes five key components. These components are outlined in *Figure 1: Creating a Logic Model.*

The first component is the problem. Before you can begin to identify or evaluate a solution, you must clearly articulate the problem. The problem should be based on a thorough local needs assessment. It may also be helpful to do a root-cause analysis. Your problem guides the rest of the work when creating a logic model, and all pieces of the logic model should be directly aligned to the problem.

The next phase in the creation of a logic model is to identify the inputs. Inputs are the outside sources that you bring in to address the problem. These outside resources may include items such as curriculum, instructional materials or professional development. Inputs expand your capacity and introduce new ideas and fresh perspective to your school.

After you have identified your inputs you should consider your outputs. The outputs are the individual steps that you are going to take to implement the intervention with fidelity. Be as detailed as possible in describing your outputs and be sure to include a responsible point person for each activity.

Figure 1: Creating a Logic Model


Problem
Based on Needs Assessment
Inputs
What outside resources are you bringing in?
Outputs
What are you going to do?
Outcomes
What do you expect will happen?
Measures
Continuous – At all stages

Finally, your outcomes describe what you expect will happen when you implement the intervention. Again, be as specific as possible. Your outcomes may include a combination of student and non-student outcomes. For example, you may anticipate that an intervention may improve proficiency in math while also improving the collaborative climate among teachers. The SMART format for goal setting provides an easy formula for writing outcome statements.

Throughout the development of the logic model, you should also be considering the various measures you will use to collect documentation along-the-way. Measures should be continuous and varied and be aligned to all stages of the logic model. Be sure to include as many relevant measures as possible; this will make it easier for you to evaluate the impact of the intervention later.

The logic model provides a record of the various steps taken to address the problem. It should be revisited and evaluated regularly to ensure that organic changes to the plan are documented throughout the year. It may also be helpful to document relevant research that you used to inform your logic model. You can embed your documentation within the framework described above, or you may find it easier to create a references document to house your research documentation.

EVIDENCE IN ESSA

ESSA's evidence provisions require the development of a logic model for evidence that falls under *Level IV: Demonstrates a Rationale*.

Phase II: Gathering Evidence

Before you begin to implement your intervention, you need to consider how you will study the impact of the intervention. ESSA's evidence provisions state that schools seeking to use intervention under *Level IV: Demonstrates a Rationale* should make an effort to study the effects of the intervention in a way that ideally produces promising evidence or higher (34 CFR 77.1).

In this section, we will discuss the primary considerations for conducting a study at your school.

Clarifying Your Questions

When you are ready to evaluate the impact of an intervention, you must first take time to clarify your questions. It is not enough to simply ask whether an intervention is working. Rather, you must create questions that are clear, focused, concise, complex and arguable. While it may feel like a sophomoric exercise, taking the time to write a quality question will make the process of answering the question easier and provide greater insight into what you really need to know. It will also provide a clear starting point for discussing the impact of an intervention with stakeholders later.

The Center for Innovation in Research and Teaching offers the following four questions for use in evaluating the quality of your research question:

1. Is the research question one that is of interest to the researcher and potentially to others? Is it a new issue or problem that needs to be solved or is attempting to shed light on previously researched topic?
2. Is the research question researchable? Consider the available time frame and the required resources. Is the methodology to conduct the research feasible?
3. Is the research question measurable and will the process produce data that can be supported or contradicted?
4. Is the research too broad or too narrow?

These four questions allow you to self-evaluate your own research questions and make appropriate adjustments. For example, most school-based research begins with a question like, "Did the intervention work on our students?" While this question is certainly of interest to you and others and therefore meets the criteria for question one, it can be improved.

A better question would read, "Did students who participated in Read 180 outperform their peers who did not participate in Read 180 on the Woodcock Johnson IV Tests of Cognitive Abilities?" This question is of interest to the researcher and others, is researchable with a feasible methodology (discussed below), is measurable and will produce data and is neither too broad nor too narrow. In short, this question is specific enough to give you the answers you need while remaining broad enough to allow you to see other possibilities from its findings.

Selecting a Study Design

Once you have identified and clarified your research question, you must design a study to rigorously collect data for analysis. There are many possible study designs available to education researchers. This guidance will explore two common study designs that are feasible in most schools: single-case design and quasi-experimental study design.

Single-Case Design

Many schools find the single-case study design to be a natural extension of their normal data collection process. For schools with fewer resources, using this design to study the impact of an intervention may be beneficial.

A single-case study design is one in which data from a single student, classroom or school is collected and analyzed to measure growth. In single-case design, the performance of a subject is compared against past performance to draw a loose correlation.

There are typically three phases to a single-case study; baseline, intervention and reversal. During the baseline phase, preliminary data is collected to establish a point of comparison. The intervention is implemented during the intervention phase and further data is collected. Finally, during the reversal phase, an intervention is removed to allow an additional post-intervention comparison point. The reversal phase can help researchers understand if the intervention has had long-term effects or if the intervention may work only during implementation. The phases may be cycled through multiple times or staggered over multiple groups of students to collect more data for analysis (Kratochwill et al., 2010).

This type of study is common in studying interventions that seek to change behaviors over time. For example, if a school is attempting to increase student attendance rates, then it would collect baseline data, implement an intervention designed to increase attendance and then remove the intervention to determine the impact of the intervention. Ideally, during the intervention period, the researchers would observe higher levels of attendance than during the baseline period. During the reversal period, when the intervention condition is removed, the researcher will be able to use the new data to determine if the intervention created a short-term or long-term improvement in student attendance.

Similarly, a single-case design could be used to monitor academic achievements. A school seeking to increase phonemic awareness in third-grade students may use this study design to monitor changes in the rate of growth over time. Student achievement could be regularly monitored to determine a rate of growth – this is the baseline phase. As the intervention is implemented, the researcher would expect to see an increase in the rate of growth over time. During the reversal phase, researchers would continue to monitor the rate of student growth. By comparing the rate of growth during baseline to the rate of growth during reversal, the researcher will be able to determine how an intervention influenced the long-term rate of growth.

Single-case designs require a significant commitment to the regular and consistent collection of data. Attention should be paid to ensure that collection instruments used are both valid and reliable (discussed below) and are monitored for potential bias. In some instances, it may be appropriate to use multiple years of standardized test data to draw a causal inference. A potential pitfall of single-case studies is that they may include irreversible conditions that can skew the outcome of the study and make the findings difficult to interpret.

Under ESSA's evidence provisions, a well-implemented single-case study design could create *Level 3: Promising Evidence,* because the statistical analysis of the study findings could be used to draw a correlational conclusion. In this instance, multiple single-case studies should be pooled to produce a stronger argument for the use of the intervention. Schools may choose to partner with other schools implementing the same intervention to collect multiple cases for comparison.

It is important to note that correlation does not equal causation, meaning this type of study is unlikely to produce definitive results. The small sample size and inability to control for other variables make findings from single-case studies unable to override contrary results from a larger, more rigorous study.

Quasi-Experimental Study Design

Schools with more resources may find that the quasi-experimental study design provides a more accurate determination of the impact of an intervention. In a quasi-experimental study design, data from previously assigned groups of students, such as classrooms, could be compared for evaluation. One group, the intervention group, receives the intervention while the other group, the control group, does not receive the intervention. By comparing the results of these two groups, researchers can determine the magnitude of the impact of the intervention (What Works Clearinghouse Standards Handbook v. 4, 2017).

Under ESSA's evidence provisions, a well-constructed quasi-experimental study is considered *Level 2: Moderate Evidence*, assuming that it also meets the size and site expectations set out in the Code of Federal Regulations (CFR). The desired size of an analytic sample is 350 or more students, or 50 or more groups of 10 or more students. The CFR also states that studies should include multi-site samples in which site can mean a school district, state or locality (34 CFR 77.1).

Schools that wish to monitor interventions using a quasi-experimental model may benefit from partnering with schools in other districts who are using the same intervention. For example, if 10th-grade teachers at Commonwealth County High School and Bluegrass County High School are both implementing an intervention designed to improve students' understanding of algebraic expressions, they could pool their data to create a large and multi-site sample as required by the CFR. In this example, the schools may decide to offer the intervention only to students in even-numbered periods and use students in odd-numbered periods as their control group.

A major consideration of a successful quasi-experimental study is the establishment of baseline equivalence. Baseline equivalence is a measure that demonstrates that the intervention group and control group are similar enough at baseline to provide a clear comparison. Baseline equivalence is determined by analyzing the baseline data using the formula for effect size (discussed below). The What Works Clearinghouse (WWC) sets the standards for baseline equivalence and may direct researchers to include a statistical adjustment to correct for selection bias.

Other Study Design Considerations

While single-case and quasi-experimental study designs may be more accessible to schools, other more rigorous study designs are useful for interpreting the effectiveness of an intervention.

Experimental study designs examine interventions by comparing the achievement of two randomly assigned groups of students. The nature of random assignment applies a higher level of rigor to the study design by controlling for more external variables. Studies conducted using experimental study designs are eligible to be *Level 1: Strong Evidence* under ESSA's evidence provisions (What Works Clearinghouse Standards Handbook version 4, 2017; 34 CFR 77.1).

Another common study design employed by education researchers is the regression discontinuity design (RDD). In an RDD, the researcher assigns a cutoff threshold and analyzes scores lying closely to the threshold to estimate its effect. RDD's are often used when random assignment is not possible for a given population (What Works Clearinghouse Standards Handbook version 4, 2017).

Selecting a Collection Protocol

When designing your study, it is important to select a data collection protocol that will provide you with an accurate picture of change over time. Typically, education researchers rely on quantitative assessments of student performance to measure the impact of an intervention. When selecting a student assessment, it is important that you choose an instrument that is both valid and reliable.

Reliability and validity are two statistical terms used to describe the dependability of a data collection instrument (such as a test or survey). Reliability refers to the dependability or consistency of a measure. Instruments that are said to have high levels of reliability should produce consistent results over multiple administrations. Validity refers to the quality or soundness of a measure. An instrument has high levels of validity if it can be proven that the instrument collects the data that it is said to collect (Spatz, 2011).

Reliability and validity play a significant role in monitoring for bias. Testing and observational bias are important factors to consider when selecting or designing a data collection protocol. Test bias occurs when subgroup performance differs for reasons that are unrelated to the testing protocol (Salkind, 2006). Another good definition was developed by T. Anne Clearly (1968) and states that test bias occurs when a test measures different things for different people.

By recognizing our own potential for bias and collaborating with a diverse group of professionals during the reliability and validity testing phase, we can reduce the rate of potential bias in homegrown assessment.

Standardized tests are thoroughly vetted and are generally accepted to have high levels of reliability and validity and low levels of bias. Their results can be trusted to provide an accurate picture of student achievement. It is the best practice to use standardized tests when measuring the impact of an intervention.

Teacher-, school- or district-created assessments can also be used assuming that the assessments are both valid and reliable. Validity can generally be determined if a committee of reviewers agrees that the test measures what it is said to measure. Reliability should be calculated using statistical measures. The most common statistical measure is the test-retest model, in which the same group of students are given the assessment twice and the correlation between the two tests are used to estimate reliability using the Pearson product-moment correlation coefficient.

Depending on the nature of the intervention, it may be appropriate to use other methods of data collection. Qualitative methods, such as surveys or direct observation, may be beneficial when monitoring changes in behavior, perception or participation. When designing these types of protocols, it is important to take steps to reduce potential bias by standardizing observational norms.

EVIDENCE IN ESSA

When evaluating evidence under ESSA's evidence provisions, be sure to take steps to regulate your data collection process and increase scholarly rigor. This is a good opportunity to partner with university researchers who can assist you.

Phase III: Evaluating Effectiveness

Once an intervention has been implemented and data has been collected, education leaders should apply appropriate methods of data analysis to evaluate the effectiveness of the intervention. By using descriptive statistics and calculating the statistical significance and effect size, education leaders can turn raw data into meaningful, standardized measures for discussing the impact of an intervention over time.

Descriptive Statistics

Descriptive statistics are easily calculated statistical tests that allow you to succinctly and accurately describe study findings. There are three broad categories of descriptive statistics: distribution, measures of central tendency and variability.

Distribution is a method of summarizing the frequency of values for a variable and is typically represented by a percentage (Spatz, 2011). In education settings, distribution is commonly used to compare student achievement before and after the implementation of an intervention and may be presented as a bar graph, as demonstrated in *Figure 2: Distribution Graph Example.*

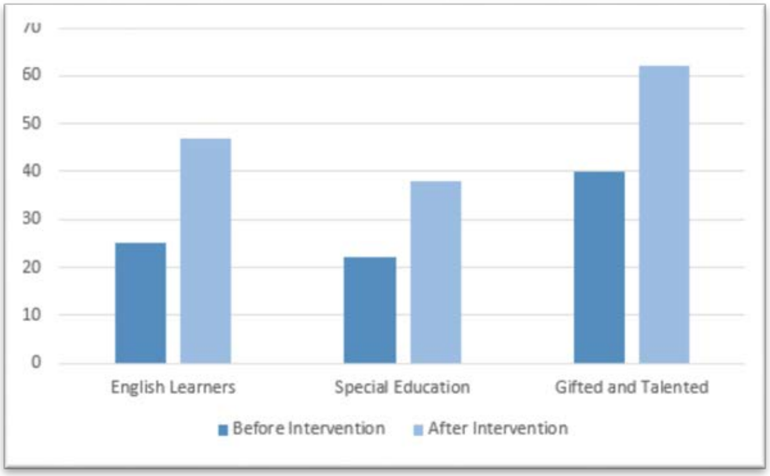


Figure 2: Distribution Graph Example

Measures of central tendency estimate the central point of a distribution of scores. This type of descriptive analysis is one of the most common methods used to describe data because it gives you a single number to describe a data set. Central tendency is measured by mean (the average of the scores), median (the middle number in a distribution) and mode (the value that shows up most often) (Spatz, 2011).

Finally, variability, sometimes called dispersion, seeks to describe the spread of values around the central tendency. The primary methods for calculating variability are range and standard deviation. Range is calculated by subtracting the lowest number in a data set from the highest number. Standard deviation is a figure that describes how close the scores in a data set are to the mean. A low standard deviation indicates that scores are clustered closer together, while a higher standard deviation indicates that scores are more spread apart. Standard deviations are often represented by a bell curve graph called a plot of normal distribution (Spatz, 2011).

Statistical Significance

The CFR states that evidence should "show a statistically significant and positive (i.e. favorable) effect of the intervention on a student outcome or other relevant outcome (34 CFR 7:11)." This means that education leaders must take the time to evaluate the statistical significance of study findings before making decisions about an intervention.

There are a wide variety of statistical tests that may be used to analyze a data set. It is unrealistic for this guidance to include a discussion of all potential statistical measures available to education leaders. Care should be taken to ensure that appropriate statistical measures are chosen for the data collected.

WWC sets the standards for statistical significance in education research. The statistical reporting measure that describes outcome significance is known as the *p* value. It is generally accepted that an outcome is significant when *p* is less than 0.05 ($p<.05$). Study findings can also be designated *substantively important* if the difference between the performance of the intervention group and the control group is greater than 0.25 standard deviations (What Works Clearinghouse Standards version 4, 2017).

Effect Size

Effect size is the standardized measure of the magnitude of a difference. Once you have determined that an outcome is significant, calculating the effect size can tell you how big of a difference the intervention created in the student outcome. One of the most straight forward methods for calculating effect size is called Cohen's *d.*

The formula for Cohen's *d* is

$$d = \frac{M_{intervention} - M_{control}}{SD_{pooled}}$$

where $M_{intervention}$ represents the mean performance of the intervention group, $M_{control}$ represents the mean performance of the control group, and SD_{pooled} represents the pooled standard deviation.

The *d* index is used to discuss the impact of an intervention. It is generally accepted that an effect size is small when $d = 0.20$, medium when $d = 0.50$, and large when $d = 0.80$ (Spatz, 2011). In his seminal work, *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement,* John Hattie (2009) suggests that schools should focus their efforts and resources on interventions with an effect size of $d = 0.40$ or higher, rationalizing that students demonstrate developmental growth at a rate of $d = 0.00$-0.15 and typical teaching (that is, teaching without additional interventions) generally effects achievement at a rate of $d = 0.15$-0.40.

Effect size is an important consideration for education leaders. It is plausible that a study might have statistically significant results but a relatively low effect size. While we shouldn't discount interventions with low effect size for that reason alone, it is a worthwhile consideration when resources are scarce. It may be more beneficial for schools to direct resources towards interventions with a higher effect size.

Considering Outside Factors

When analyzing the potential impact of an intervention, it is important to consider the other outside factors that may have influenced student performance. A rigorously developed study design should take steps to limit the impact of outside influences on study participants; however, there may be times when influences cannot be removed.

Izci (2016) divides the outside influences that impact intervention implementation into three broad categories: external factors, contextual factors and resource-related factors. Generally speaking, resource-related factors should not dramatically influence the results of a well-constructed study because the necessary resources should be provided to study participants. However, resources such as reliable access to high speed internet may influence and skew the results of a study.

Contextual and external factors are more difficult to filter out of a study and should be carefully considered by education leaders when monitoring intervention impact. Contextual factors include items like school context and internal policy, student attitudes and parent viewpoints, while external factors can include things like curriculum developers, state policy and pressure from high-stakes tests (Izci, 2016). Other external factors, such as legislative priorities, can also impact the implementation of an intervention or deployment of a curriculum (National Research Council, 2002).

In sum, it is important for education leaders and researchers to consider the wide range of impacts that external factors have on the successful implementation and study of an intervention. These external factors must be acknowledged, accounted for and addressed by study authors seeking to create meaningful evidence to support the continued use of an intervention.

EVIDENCE IN ESSA

ESSA's evidence provisions, as defined in the Code of Federal Regulations, require that evidence "shows a statistically significant favorable impact on a relevant outcome (34 CFR 77.1)."

Phase IV: Making Educational Decisions

After an intervention has been implemented with fidelity and data has been collected and analyzed, education leaders must incorporate this new knowledge into their educational decision making process. The National Center for Research in Policy and Practice identified the following three uses of research for educational decision making:

- Instrumental use: Research is applied to guide or inform a specific decision.
- Conceptual use: Research induces changes in the way a person views either a problem or the possible solution space for a problem.
- Symbolic/political use: Research is used to validate a decision or legitimate a decision already made (Penuel et al, 2016).

All three research uses are relevant in the daily decision-making of education leaders. The results of a study conducted in a school can be used to inform future decisions, update our perspective on a problem and argue for or against a position.

In addition to these common tasks, education leaders should use homegrown research in the planning of instruction. This is a natural extension of DuFour's Four Essential Questions for Professional Learning Communities (PLC) (DuFour et al, 2010). In DuFour's PLC model, teachers focus on the standard they want their students to know, determine how they will measure success and plan how they will either remediate or extend the learning experience. When reviewing the impact of an intervention on a student outcome, education leaders should reflect on how the students performed, whether or not the intervention was successful at providing remediation or expansion and how to shape instructional decisions moving forward.

In most cases, the completion of a study will prompt education decision-makers to determine if the future implementation of the intervention is beneficial to students long-term. If the study has shown a statistically significant and favorable effect, or is determined to substantively important, then the study may qualify for a higher level of evidence under ESSA's evidence provisions and warrant further use. If a study fails to meet this threshold, then it may be in the best interest of the students to select a new intervention and begin the study process again.

EVIDENCE IN ESSA

At the end of the process, newly created evidence should be submitted for vetting by peer-reviewed journals to create higher levels of evidence.

References List

Alter, C., & Murty, S. (1997). Logic modeling: A tool for teaching practice valuation. *Journal of Social Work Education, 33*(1), 103-118.

Blase, K., Fixsen, D., & Jackson, K.R. (2015). Creating meaningful change in education: A cascading logic model. Scaling-up brief. *State Implementation & Scaling-up of Evidence-based Practices.*

Center for Innovation in Research and Teaching. (n.d.). Writing a good research question. Retrieved August 1, 2018, from https://cirt.gcu.edu/research/developmentresources/tutorials/question

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement (5)*2, 115-124.

Definitions that Apply to Department Regulations, 34 C.F.R. 77.1 (2015).

DuFour, R., DuFour, R., Eaker, R., & Many, T. (2010). *Learning by doing: A handbook for professional learning communities at work.* Bloomington: Solution Tree Press.

Hattie, J. (2009). Visible learning: A synthesis of meta-analyses related to achievement. New York: Rutledge.

Hernandez, M. (2000). Using logic-models and program theory to build outcome accountability. *Education & Treatment of Children, 23*(1), 24.

Izci, K. (2016). Internal and external factors affecting teachers' adoption of formative assessment to support learning. *International Journal of Educational and Pedagogical Sciences (10)*8, 2800-2807.

Kratochwill, T.R., Hitchcock, J., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M. & Shadish, W.R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Lawton, B., Brandon, P.R., Cicchinelli, L., & Kekahio, W. (2014). Logic models: A tool for designing and monitoring program evaluations. *Regional Educational Laboratory Pacific.*

National Research Council. (2002). *Investigating the influence of standards: A framework for research in mathematics, science, and technology education.* I.R.Weiss, M.S.Knapp, K.S.Hollweg, and G.Burrill (Eds.), Committee on Understanding the Influence of Standards in K-12 Science, Mathematics, and Technology Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington: National Academy Press.

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., . . . Allen, A. R. (2016). *Findings from a national study on research use among school and district leaders* (Tech. No. 1). National Center for Research in Policy and Practice.

Salkind, N.J. (2006). *Test and measurements for people who (think they) hate tests and measurements.* Thousand Oaks, Calif.: Sage Publications Inc.

Spatz, C. (2011). Basic statistics: Tales of distributions. Australia: Wadsworth.

Stegemann, K.C., & Jaciw, A.P. (2018). Making it logical: Implementation of inclusive education using a logic model framework. *Learning Disabilities Worldwide, 16*(1), 3-18.

United States, Department of Education, Institute for Educational Sciences. (2017, October). *What Works Clearinghouse Standards Handbook v. 4*. Retrieved August 1, 2018, from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf