

Purpose

- To improve student outcomes for schools that are identified as needing comprehensive or targeted support and improvement, the Every Student Succeeds Act (ESSA) (SEC. 1003) requires that SEAs, LEAs, educators and other education stakeholders, use evidence-based interventions (EBIs) in their improvement plans.
- Under ESSA section 8101(21)(B), any evidence-based intervention, activity, or strategy that the LEA pays for with the ESSA 1003 funds, must comply with one of the first three tiers of evidence under ESSA section 8101(21)(A) (strong, moderate, or promising evidence). Four levels of evidence were developed by the U.S. Department of Education.
- The purpose of this rubric is to clarify what is meant by an evidence-based intervention for ESSA 1003 so that the reader can determine if the intervention proposed by the applicant fits within the criteria issued by the U.S. Department of Education into one of the three tiers of evidence.

Terms Referenced in Rubric:¹

Attrition

- The loss of sample during the course of a study. It occurs when individuals initially randomly assigned in a study are not included when researchers examine the outcome of interest. Attrition is a common issue in education research, and it occurs for many reasons (e.g. a student in the study relocates to another school). Attrition occurs when an outcome variable is not available for all subjects initially assigned to the intervention and comparison groups. If a randomized controlled trial (RCT) or regression discontinuity design (RDD) study has high levels of attrition, the validity of the study results can be called into question.

Baseline Equivalence

- A demonstration of the similarity of the analysis groups at baseline. Randomized controlled trials with high attrition and quasi-experimental designs must establish that the intervention and comparison groups used in the analysis were equivalent on observable characteristics at baseline. Characteristics for which equivalence must be established are outlined in the protocol for a review.

Confounding Factor

- A component of a study that is completely aligned with one of the study conditions. For example, a study may have one intervention school and a different comparison school. In this case, it is impossible to separate how much of the observed effect was due to the intervention and how much was due to the particular school in which the intervention was used. Therefore, a study with a

¹ Adapted from: What Works Clearinghouse. (n.d.). *WWC / Glossary of terms*. Retrieved from <https://ies.ed.gov/ncee/wwc/Glossary/improvement%20index>

confounding factor cannot meet standards, as the impact cannot be attributed solely to the intervention.

Effect Size

- A standardized measure of the magnitude of an effect. The effect size represents the change (measured in standard deviations) in an average student's outcome that can be expected if that student is given the intervention. Because effect sizes are standardized, they can be compared across outcomes and studies.

Face Validity

- When an outcome measures what it claims to be measuring.

Moderate Evidence

- Limited evidence for a practice guide recommendation. This rating is assigned when the panel finds high quality causal research that links a practice with positive results, but the research may not adequately rule out other causes of the positive results, or the school and classrooms are not similar to those targeted by the guide.

Promising Evidence

- No causal evidence for a practice guide recommendation. The panel cannot point to a body of research that demonstrates the practices' positive effect on student achievement. In some cases, this simply means that the recommended practices would be difficult to study in a rigorous, experimental fashion. In other cases, it indicates that researchers have not yet studied this practice, or that there is weak or conflicting evidence of effectiveness. A minimal evidence rating does not indicate that the recommendation is any less important than other recommendations with a strong evidence or moderate evidence rating.

Quasi-Experimental Design (QED)

- A design in which one or more groups receive the intervention, and one or more groups do not receive the intervention. For example, one group of 5th and 6th grade teachers receive funding to implement project-based learning at their school, while another group of 5th and 6th grade teachers did not receive funding and continue business as usual. These groups were not randomly assigned, so a study measuring the impact of the project-based learning program on outcomes would be using a quasi-experimental design. For a quasi-experimental design to be rigorous, the intervention and comparison groups must be similar, demonstrating baseline equivalence on observed characteristics, before the intervention is started. Strong quasi-experimental designs will, at best, be rated as Meets WWC Group Design Standards with Reservations (the middle possible rating).

Randomized Control Trial (RCT)

- A design in which groups are created through a process that is random. Carried out correctly, random assignment results in groups that are similar on average in both observable and unobservable characteristics, and any differences in outcomes between the groups are due to the intervention alone. Strong randomized controlled trials can receive the highest rating of Meets WWC Group Design Standards without Reservations.

Regression Discontinuity Design

- A design in which groups are created using a continuous scoring rule. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and an effect occurs if there is a discontinuity in the two regression lines at the cutoff.

Reliability

- An outcome measure is reliable if it produces the same scores when it is administered at different times to different people in different settings. In other words, the outcome measure can be replicated.

Statistical Significance

- The likelihood that a finding is due to chance rather than a real difference. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than five percent ($p = 0.05$).

Strong Evidence

- Consistent evidence supporting a practice guide recommendation. This rating is assigned when the panel finds high quality causal research that links a practice with positive results, ruling out other causes of the positive results.

Understanding Citations to Evidence-Based Interventions

*At least one **citation (included as an attachment by the applicant)** that illustrates how this EBI has had a **statistically significant positive effect on student outcomes** and thus met the **evidentiary threshold required by ESSA***

The above statement includes the following (each column should have the relevant checkboxes):

1. Is the citation to a reputable source?	Yes	No	Comments
<p>Citations to well-designed and well-implemented studies referenced from the following (Applicant should identify what source where they found the study):</p> <ul style="list-style-type: none"> <input type="checkbox"/> An entity from the National Center for Educational Evaluation and Regional Assistance (NCEE), including the What Works Clearinghouse (WWC) the Regional Educational Laboratories (REL), or Educational Resources Information Center (ERIC) <i>OR</i> <input type="checkbox"/> A peer-reviewed academic journal (e.g. <i>American Educational Research Journal</i>) <i>OR</i> <input type="checkbox"/> A report published by a reputable organization focused on education research and evaluation (e.g., AIR, WestEd, Abt Associates, Evidence for ESSA, Best Evidence Encyclopedia, Results First Clearinghouse, RAND Report on School Leadership Interventions) 			
2. What is the level of evidence?	Yes	No	Comments
<p>For the evidentiary threshold required by ESSA, the applicant <i>must</i> include a well-designed and well-implemented study in one of the following categories (this can be found in the abstract, executive summary or methodology section of the cited study).</p>			
<input type="checkbox"/> Strong Evidence. Studies that demonstrate strong evidence are typically randomized control trials (RCT) , where participants of the study were <i>randomly</i> assigned into a treatment and control groups. There was also some sort of intervention (e.g., new type of instructional technique) used to change outcomes. Participants (e.g., students, teachers, schools) in the treatment group are those that receive the intervention. Participants that do not receive the intervention are those in the control group.			
<input type="checkbox"/> Moderate Evidence. Studies that demonstrate moderate evidence are typically quasi-experimental designs (QED) , where participants were <i>not</i> randomly assigned into treatment and comparison groups. Some natural change happened, such as a newly-funded program implementation that created a group that received the treatment/intervention, and one that did not. However these groups were <i>not</i> randomly assigned. Other methods were used to attempt to create baseline equivalence between groups, so that each participant in the treatment group has a comparable “twin” in the group that did not receive the treatment (i.e., the comparison group). Many times, statistical processes are used to match those in the treatment group to similar participants that did not receive the intervention. Characteristics that each member of the treatment and comparison group would be matched on include gender, race/ethnicity, free/reduced price meal eligibility, prior academic performance, teacher, and school characteristics. <i>Note.</i> Regression discontinuity designs can also be considered strong evidence, though they are a type of QED. To do so, they will need to meet certain standards from IES.			
<input type="checkbox"/> Promising Evidence. Promising evidence includes correlational studies, with statistical controls for selection bias. These studies will <i>not</i> include treatment and control/comparison groups, and participants are part of one large group. In these studies, researchers examine relationships among specific variables and the outcomes. Background characteristics such as gender, race/ethnicity, free/reduced price meal eligibility, prior academic performance, teacher and school characteristics, and other variables are used in these correlational studies to control for participant characteristics that might impact the outcomes, so that the potential effect of the intervention on the outcome can be measured. For example, a study in this category might examine the relationship between the hours spent by students on a new Algebra I e-learning module and their performance on the Algebra I ECAs, when controlling for student background characteristics, such as prior math performance, demographic factors, teacher experience and PD hours with this module, etc. <i>Note.</i> If there are no controls, then the study <i>cannot</i> be considered promising evidence.			
X Demonstrates a rationale. Studies categorized as “Demonstrates a Rationale” (e.g., any other type of study not referenced above) are <i>not</i> an acceptable level of evidence, though they might have a strong logic behind the intervention, or some promising preliminary data.			
<p><i>Note.</i> Determining if a study is well-designed and implemented can be challenging. Two clues to look for are low attrition (loss of participants) and the absence of confounding factors (additional variables that might explain differences in the outcomes between the groups that were not addressed in the study). Please see the glossary above or reference the WWC.</p>			

3. Was a statistically significant positive effect found?	Yes	No	Comments
<p>Statistically significant positive effects include <i>both</i> of the following:</p> <p><input type="checkbox"/> The study authors report that the intervention was statistically significant (i.e., the change in outcomes were not due to chance for the students who received the intervention compared to students who did not receive it) AND</p>			
<p><input type="checkbox"/> Positive (favorable) effect size:</p> <ul style="list-style-type: none"> ○ For studies that are categorized as strong/moderate evidence, the effect size listed will typically be a standardized difference in the outcome between the treatment and control/comparison groups, or in other words, a proportion of a standard deviation whereby the treatment group surpasses the control/comparison group. ○ The effect size should be positive, so that the treatment group experienced a better overall outcome (e.g., better test scores, fewer suspensions, etc.) than the control/comparison group. A rule of thumb is an effect size greater than or equal to 0.25, which means that the average student in the treatment group would be predicted to score 0.25 standard deviations better than if they had been part of the control or comparison group. ○ For studies that are considered promising evidence, the type of effect size used can be a bit more complex. A rule of thumb is to look at a statistic called the R², or coefficient of determination, which shows the amount of variance explained by the combination of the variables to predict the outcome in the correlation study with statistical controls. This can be found at the bottom of most tables with the statistical results of the correlation analyses. The same rule of thumb above can be used, so that an R² greater than or equal to 0.25 would be acceptable. <p><i>Note:</i> Larger sample sizes are favored (at least 10 classes or 250 students), though they will typically result in smaller effect sizes and potentially negligible results for statistical significance</p>			
4. Are relevant student outcomes used?	Yes	No	Comments
<p>Outcomes measured in the study should be relevant to the setting and student population that the LEA serves. The applicant should state the extent to which the participants in the study are similar to those who would participate in the intervention. It helps if this study was conducted at multiple sites.</p> <p><input type="checkbox"/> Is the setting similar to the LEA (grade level, urban/rural/suburban)?</p> <p><input type="checkbox"/> Is the student population (FRL EL, SpED, race/ethnicity) similar to the LEA?</p> <p><i>Note:</i> The outcome(s) measured should also be reliable and have face validity.</p>			

If you have questions or need clarification, please contact Gina Romano, Senior Data Coach Specialist at gromano@doe.in.gov or 317-234-4746