Evidence for ESSA: Standards and Procedures
Version 2.2
February, 2024

Evidence for ESSA is intended to provide educators with reliable, easy-to-use information on programs and practices that meet the standards of evidence in the Every Student Succeeds Act (ESSA). EvidenceforESSA.org is using an updated set of standards so that we can include a wider range of studies that are rigorous enough to meet the expectations of the Act. Previously posted information will be reviewed using these new standards. All new submissions will be reviewed using these criteria as of November, 2023.

Defining ESSA Evidence Categories

ESSA defines strong, moderate, and promising evidence of effectiveness. It also lists a fourth category indicating programs lacking evidence of effectiveness, though they may be under evaluation currently.

The ESSA evidence standards are a giant step forward in defining what it means to have evidence of effectiveness for educational programs. However, the legislation does not provide sufficient detail or resources to permit educators to easily evaluate the evidence supporting specific programs. The purpose of Evidence for ESSA is to provide further definition, to evaluate the evidence bases for PK-12 programs, and to communicate this information fairly and clearly.

Consistent with the law and guidance, we place programs in categories according to the following procedures:

Strong (Tier 1)- A program is placed in "strong" if it has a well-conducted, randomized study showing a statistically significant positive effect on at least one outcome measure (e.g., state test or national standardized test) analyzed at the proper level of clustering (class/school or student) with a multi-site sample of at least 350 students.

Moderate (Tier 2)- A program is placed in "moderate" if it meets all standards for "strong" stated above, except that instead of using a randomized design, qualifying studies are prospective quasi-experiments (i.e., matched studies).

Promising (Tier 3) –A program is placed in "promising" if it has a study that would have qualified for "strong" or "moderate" but did not because it failed to account for clustering (but did obtain significantly positive outcomes at the student level) or did not meet the sample size requirements for Moderate. Post-hoc or retrospective studies may also qualify as "promising."

Procedures

Finding Eligible Studies

Our intention is to evaluate every program currently in general use, so that users of our website can look up any program they have ever heard of and find a rating. To find the evidence, we carry out a comprehensive search of the literature by topic through a multi-step process that

includes an electronic database search, a web-based search of educational research sites and educational publishers' websites, search of recent meta-analyses, a hand search of relevant peer-reviewed journals, and a final review of citations found in relevant documents retrieved from the first search wave. In addition, we review studies sent to us by program developers, program evaluators, and educators. We also obtain state lists of approved programs and other lists of programs currently in use and evaluate the evidence supporting all programs. This is an ongoing process, and we accept studies at any point.

Inclusion Criteria for Studies

1. Studies must be of programs available today to schools in the U.S.
2. Studies must have been carried out in the U.S. from 2000 to the present.
3. To be accepted for review, studies must compare intervention groups to comparison groups. Either random assignment to conditions or quasi-experimental assignment based on matched schools, classes, or students; regression discontinuity designs, must be used. Comparisons to norming groups, pre-post comparisons, or other non-experimental comparisons are not accepted. Single-case design studies are not currently being accepted. Comparisons without a control group representing ordinary practice already in place or "business as usual", are not accepted. Treatment on the treated analyses (those that compare the impacts for students who received a specified "dose" of the treatment with students who received no treatment) are not accepted.
4. Studies must not have any confounding factors that are perfectly aligned with group assignment. For example, if all the treatment schools were charter schools and all the comparison schools were not charter schools, school type would be a confounding factor and the study would not be accepted. Time can also be a confounding factor, such as in successive cohort designs where the entire treatment sample comes from one school year, and the entire comparison sample comes from a different school year. This would make the study ineligible.
5. Quasi-experimental studies that conduct the matching/weighting prior to posttest collection and define the experimental group as all students who received any treatment, are eligible for a Moderate (ESSA Tier 2) rating. The quality of the match must be demonstrated through comparable baseline characteristics including pretest equivalence and similar demographics.
6. Post-hoc studies (retrospective studies where the sample is identified after the outcome assessment has been given) are considered under the following conditions:
    a. The treatment group is all students who received any treatment and were not restricted by dosage or usage.
    b. The matching is done using only baseline characteristics.
    c. The quality of the match must be demonstrated through comparable baseline characteristics including pretest equivalence and similar demographics.

    Post-hoc studies that meet these requirements will be considered correlational studies, so the highest rating they can achieve is Promising (Tier 3). The key distinguishing characteristic between post-hoc and prospective quasi-experimental studies is the timing of the matching/weighting.

7. Regression discontinuity design studies are reviewed using the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0, and may be considered for Strong (Tier 1).
8. Quasi-experiments or small randomized studies (fewer than 50 clusters for cluster-randomized studies or 350 students for student-randomized studies) must provide pretest data to establish equivalence of the analytic sample. On achievement measures, the average pretest difference for the analytic sample must not exceed 0.25 standard deviations. Large, randomized studies (at least 50 clusters for cluster-randomized studies or at least 350 students for student-randomized studies) do not need to demonstrate baseline equivalence.
9. Studies' dependent variable(s) have to include a quantitative measure of student academic achievement, social-emotional learning, or attendance. These outcome measures could be a standardized test, or a test created by test developers not involved with the research, but tests made by the researchers or developers themselves are not acceptable. Also, tests that are aligned with content taught in the experimental but not the control group are not acceptable. Tests administered individually by students' own teachers or others with a potential stake in the outcome only for the purposes of the study or the intervention are not accepted.
10. Studies in the preschool years present unique challenges in terms of outcome measurement (Chambers et al., 2016; Slavin et al., 2009). For example, it is possible for a study to find positive effects of programs that introduce skills not ordinarily taught in preschool on measures of those skills. To address these concerns, for preschool studies with outcomes measured at the end of preschool only, the highest rating the study could receive is Promising (Tier 3). For preschool studies with outcomes measured at later time points (e.g., end of kindergarten), the study may be eligible to receive any rating (Tiers 1, 2, and 3). To reflect developmentally appropriate skills at the preschool stage, acceptable outcomes of studies of reading interventions may include quantitative measures of student language in addition to reading achievement and literacy skills.
11. Study durations have to be at least 12 weeks from program inception to posttest.
12. Studies have to have at least 2 teachers and 30 students per condition (which qualifies for Promising [Tier 3]). In order to avoid possible confounding factors, at least two schools per condition are required when randomization/matching of students/teachers takes place outside of a single school.
13. From pretest to posttest, attrition (dropout) must be similar between experimental and control groups. Studies with differential attrition of more than 15 percentage points are rejected. Attrition should be calculated for both randomized and prospective quasi-experimental studies. Attrition is not expected to be assessed for post-hoc studies because the groups are formed after implementation and collection of the outcome, so the matching/weighting is often done on only the complete data, thereby resulting in no attrition.
14. Studies must use a form of a program that could in principle be replicated. Studies that provided exceptional, non-replicable resources, such as having the researcher or his or her students provide tutoring or placing a graduate student in each class to help teachers every day, are not included.

Evaluating Study Outcomes

Statistical Significance

The ESSA evidence standards place a strong reliance on determination of statistical significance, as it requires at least one study with significant positive effects for each of its three top levels.

1. If random assignment and treatment are at the individual student level, statistical significance is usually determined using analysis of covariance, controlling for pretests and possibly other factors, or using equivalent procedures, such as multiple regression.
2. If subjects were assigned or treated in clusters (classes or schools), statistical significance for clustered designs should use HLM, with pretests and other variables as covariates, or other multi-level methods accounting for clustering.
    a. If a clustered experiment failed to take clustering into account, we use a formula in the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0, that recalculates statistical significance accounting for clustering.
    b. If a clustered study failed to account for clustering, or if a study used HLM or other methods that account for clustering, but did not find a statistically significant result, we re-analyze the data ignoring clustering for possible inclusion of the study in the ESSA "Promising" category. This procedure generally produces unbiased effect sizes, but it overstates statistical significance, so studies rated "Promising" should be interpreted as preliminary indications of what a program's true effects might be.
3. Each outcome measure is considered separately. If outcomes are analyzed by pooling across grade levels, the reported statistical significance is used. If outcomes are analyzed separately by grade levels, then the findings must be statistically significant across a grade band (preK-K, 1-2, 3-6, middle school, high school), rather than a single grade demonstrating statistically significant results. For example:
    a. A study examines results for kindergarten, first, and second grade separately. Significant results are found in first grade, but not kindergarten or second grade. The result is also not significant when pooled across those grades. This would not qualify as a significant positive effect.
    b. A study examines results for third, fourth, and fifth grade. Significant results are found in third and fourth grade, but not fifth. When results are pooled across third, fourth, and fifth grade, the result is statistically significant. This would qualify as a significant positive effect.
    c. A study examines results for eighth, ninth, and tenth grade. Significant results are found in eighth grade only. Eighth grade is in a different grade band from ninth and tenth grade, so the results for eighth grade are looked at separately from the results for ninth and tenth grade (which are pooled together). The significant result from eighth grade would qualify this as a significant positive effect.
4. In studies that have multiple randomized sites (states/districts), significant results from one state/district may qualify a program for a rating (unless there are significantly negative results elsewhere) but effect sizes will be averaged across sites. In this case, the different sites will each be treated as their own "mini-study" that must meet the criteria

outlined in this document (including sample size requirements for Strong [Tier 1] and Moderate [Tier 2]). The same approach will be applied in studies with multiple cohorts. For example, a researcher conducts a multi-site study of a literacy intervention in Kansas and Iowa, with each site serving as a "mini-study". The results from Kansas meets criteria on its own for Moderate [Tier 2]. The results from Iowa do not yield significant positive or negative effects. Results will be aggregated across the two sites to determine overall impacts, but the Kansas study will still qualify on its own for Moderate [Tier 2] status.

5. In studies of universal or whole class/school models, outcomes for the full sample as well as outcomes for the subset of students most at risk (e.g. performing below grade level or in the lowest quartile at baseline) will be considered for ratings. Each of these samples will each be treated as its own "mini-study" that must meet the criteria outlined in this document (including sample size requirements for Strong [Tier 1] and Moderate [Tier 2]).

Effect Sizes

Ordinarily, effect sizes should be computed as the experimental-control difference in means (adjusted for covariates) divided by the unadjusted posttest standard deviation for the control group (Glass' delta) (or a pooled SD [Cohen's d or Hedges g] if the control group SD is not available).

Standard deviations adjusted for pretests or other covariates may not be used as the denominator of the effect size formula. SDs of gain scores may not be used. Only unadjusted SDs are acceptable.

Difference-in-differences ($ES_{post} - ES_{pre}$) can be used when adjusted scores are not reported. Lipsey & Wilson (2002) provide other formulas for estimating effect sizes when adjusted SDs are not reported. For example, ES can be estimated from exact t and f values, B in regressions, and odds ratios.

Pooling Effect Sizes Within Studies

Effect sizes are pooled (averaged) across outcomes within a single study to find the average effect for that study.

In reading and mathematics studies, the overall study level effect size is generally the reading total or math total. For example, if the study reports state test scores or standardized tests such as GRADE/GMADE, we would calculate total reading or math. Otherwise, if only separate subscales are reported, we combine appropriate measures.

Measures given at grade levels far from usual (e.g., phonics measures in secondary schools) are not accepted. Individually administered measures given by the student's teacher or tutor solely for the purposes of the study or the intervention are never accepted.

Acceptable tests include, for example: GMRT, GRADE, GORT, Woodcock-Johnson, CST, CAT, MAP, ERDA, STAR, ITBS, Terra Nova, CTBS, SAT, i-Ready, iSAT, ISTEP, SDRT, DRP, ETS, NWEA, TOSREC, TERA, Durrell, WIAT, DIBELS, AIMSweb, and state standardized tests, assuming that the measure is not over-aligned with the intervention.

Procedures are somewhat different for SEL studies because there are many quite diverse measures used in such studies. We have organized SEL measures under four major categories, and within these we have 17 individual variables. The categories and variables are as follows:

1) Academic
   a) Academic performance
   b) Academic engagement
2) Problem Behaviors
   a) Aggression/misconduct
   b) Bullying
   c) Disruptive behavior
   d) Drug/alcohol abuse
   e) Sexual/Racial Harassment or Aggression
   f) Early/Risky Sexual Behavior
3) Social Relationships
   a) Empathy
   b) Interpersonal relationships
   c) Pro-social behavior
   d) Social skills
   e) School climate
4) Emotional Well-Being
   a) Reduction of anxiety/depression
   b) Coping skills/stress management
   c) Emotional regulation
   d) Self-esteem/self-efficacy

Attendance has just two categories, "average daily attendance" and "chronic absenteeism," defined as the number of students absent more than 10% of school days each year, which are common school accountability metrics many states use under ESSA.

SEL and attendance variables include behaviors we want to see more of (such as emotional regulation and average daily attendance) and those we want to see less of (such as bullying, anxiety/depression, and chronic absenteeism). We have reversed the +/- signs on negative behaviors, so that all outcomes (in effect sizes) with a positive sign are desirable. Some articles list, for example, "absenteeism" as a negative number if it is going down more in the experimental group than in the control group. We have recoded "less absenteeism" as a positive number.

Aggregating Effect Sizes Across Multiple Studies

While the outcomes from a single study can earn a rating, we report the average effect size across all eligible studies. After a single effect size is computed for each study within a

domain/subject, then effect sizes are averaged across studies, weighted by sample size using an inverse variance procedure.

Because we have multiple, very different outcomes across SEL studies and two for attendance studies, our reports on these topics are more complicated than those for reading and mathematics. The summary pages for SEL show outcomes on the four categories across all eligible studies of that program. Following definitions in the ESSA law, the categories are marked in green if at least one variable in at least one study in that category met the ESSA "Strong" criterion, in orange if at least one variable in at least one study met the ESSA "Moderate" criterion, and in fuchsia if at least one variable in at least one study met the ESSA "Promising" criterion. If a category contains variables that met different ESSA levels, the summary page shows the highest level reached across all variables measured. For example, if one variable under Social Relationships had a "Strong" rating and another had a "Promising" rating, the summary page would indicate social relationships in green (i.e., "Strong").

The summary also shows the average effect size for the category, including all variables in that category. This means that it is possible that a given category is marked in green ("Strong"), but the mean effect size is very low. This could happen if a program had a statistically significant positive outcome on one variable but a zero or negative outcome on one or more other variables in that category within a single study or across studies. If the category average is zero or negative, however, we do not list the category as meeting ESSA standards, even if it does have one or more outcomes meeting one of the three ESSA criteria.

Pooling Studies To Increase Sample Size

One of the requirements to receive a rating of either Strong (Tier 1) or Moderate (Tier 2) is that the total sample size is at least 350 students. In order to reach this threshold, studies may be combined (pooled), so that the total sample across all studies is at least 350 students. For this combination of studies to receive a rating of either Strong (Tier 1) or Moderate (Tier 2), the following requirements must be met:

1. To be eligible for a rating of Strong (Tier 1), all combined studies must meet the research design requirements for Strong (Tier 1). To be eligible for a rating of Moderate (Tier 2), all combined studies must meet the research design requirements for either Moderate (Tier 2) or Strong (Tier 1).
2. To meet the requirement of statistical significance, either all the combined studies must be statistically significant on their own OR the pooled effect, using a fixed effects meta-analysis, must be statistically significant.

Placement in Tables

On the Evidence for ESSA website, programs are categorized as strong, moderate, or promising, as defined in the law. However, it is also useful to represent distinctions *within* categories, to help educators select the programs most likely to have a positive effect on their students.

We sequence programs within ESSA evidence categories according to the following:

1. Number and quality of studies.
2. Recency of studies.
3. Collective sample size across all qualifying studies.
4. Weighted mean effect size across all qualifying studies.

Badged Studies

To indicate programs with particularly strong evidence, we put a "badge" on programs with at least two studies meeting the "strong" category.

For additional information

Amanda J. Neitzel
Deputy Director for Evidence Research
Center for Research and Reform in Education
Johns Hopkins University
aneitzel@jhu.edu

References

Chambers, B., Cheung, A. C. K., & Slavin, R. E. (2016). Literacy and language outcomes of comprehensive and developmental-constructivist approaches to early childhood education: A systematic review. *Educational Research Review*, *18*, 88–111. https://doi.org/10.1016/j.edurev.2016.03.003

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, *79*(4), 1391–1465.