Evidence for ESSA: Standards and Procedures

Version 3
June, 2026

What's New in Version 3

Version 3 reflects methodological updates intended to improve transparency, consistency, and alignment between Evidence for ESSA ratings and current research standards. Major changes include**:**

- **Updated approach to quasi-experimental studies (QEDs).** Previous versions distinguished between prospective and retrospective QEDs, with retrospective studies generally limited to the Promising category. Version 3 instead shifts evaluation focuses to how analytic samples and comparison groups are constructed. This change reflects the principle that risk of bias depends more on sample-selection procedures than on the timing of the evaluation. Therefore, we introduce baseline-defined and outcome-defined analytic sample standards. To improve transparency and consistency, we explicitly distinguish between QED studies that define analytic samples from the full baseline sample (maximum rating of Moderate) and those that restrict samples based on outcome availability (maximum rating of Promising).
- **Expanded standards for public-data evaluations**. Version 3 introduces specific criteria governing evaluations that rely solely on publicly available aggregated data. These studies may be considered when they use transparent, replicable procedures and are conducted by independent evaluators, but they remain limited to the Promising category.
- **Clarified baseline equivalence requirements for QEDs.** In addition to pretest equivalence, QEDs must demonstrate balance on key baseline demographic and student-characteristic variables. These requirements are intended to improve confidence that treatment and comparison groups are comparable before implementation.
- **Updated preschool outcome standards**. Previous versions limited preschool studies measured only at the end of preschool to the Promising category. This restriction has been removed. Preschool studies may now achieve any rating if there is a valid preschool outcome based on a developmentally appropriate and fair assessment of program impacts, regardless of when outcomes are measured.
- **Simplified grade-level rules.** Previous versions required statistically significant results across a grade band (e.g., K, 1-2, 3-6) rather than a single grade. Version 3 simplifies this requirement: a statistically significant positive result in at least one grade level may now qualify as a positive effect, provided the grade-level analysis independently meets all other study design requirements (e.g., baseline equivalence, attrition, sample size) for the relevant evidence tier, and positive effects are not overridden by negative effects.
- **Clarified rules for handling negative effects.** Version 3 adds explicit guidance on how statistically significant negative effects are handled. For a program to receive a rating, positive effects must not be overridden by negative effects. This rule applies across all levels of analysis: across outcomes within a study, across grade levels, across sites in multi-site studies, across cohorts, and across studies of a program. Null (non-significant) results do not count as negative effects.

- **Expanded outcome standards.** Version 3 provides additional guidance regarding acceptable outcome measures, including requirements related to measure validity, independence from intervention content, developmental appropriateness, and use of researcher- or developer-created measures.
- **Greater transparency and review procedures.** Version 3 adds formal conflict-of-interest procedures, expanded descriptions of study identification and review processes, and a commitment to ongoing re-review of programs as standards evolve.

I. Purpose

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 uses an up-to-date set of standards based on research findings and expert recommendations regarding the internal and external validity of studies. These standards are reviewed and updated on an ongoing basis to reflect new research findings, methodological advances, and developments in the field. When updates are made to the Evidence for ESSA standards, previously posted information will be re-reviewed using these new standards. In the present set of standards, all new submissions will be reviewed using these criteria as of June, 2026.

ESSA defines *Strong*, *Moderate*, and *Promising* evidence of effectiveness. It also lists a fourth category of studies (*Demonstrates a Rationale*) for programs presently lacking evidence at one of these three tiers but planning or currently conducting efficacy research. This category, however, is not currently included in Evidence for ESSA.

The ESSA evidence standards are a giant step forward in defining varying levels of 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. Evidence for ESSA intends to address this important need by interpreting the Evidence for ESSA guidelines in relation to contemporary program usage by practitioners and advances in research methodology and communicate this information fairly and clearly both to educators and to the public.

A. Scope of Review

Programs for K-12 schools are rated based on a comprehensive, up-to-date review of all available outcome research on programs in reading, math, and science as well as in the related non-academic areas of attendance, social-emotional learning, and family engagement. We find our studies through purposeful searches by subject, and also receive submissions from program developers, research scientists, and educators who email their studies to evidenceforessa@jhu.edu. See the *Study Identification and Review Process* section for more details.

B. Version History and Implementation Dates

Evidence for ESSA began in February, 2017. The updated standards have been effective since June, 2026.

C. Commitment to Transparency and Accuracy

The Evidence for ESSA team encourages an open forum for ideas and debate. If users have a question about how we rated a program, or if they find errors in something we have posted, we ask them to contact us at evidenceforessa@jhu.edu.

II. ESSA Evidence Categories

A. Overview of Evidence Tiers

Based on our interpretation of ESSA criteria and guidance from our Technical Work Group (experts from diverse areas in education who serve as a sounding board to offer advice and guidance) as well as the Non-Regulatory Guidance provided by the Department of Education, we place programs in categories according to the following procedures:

Strong (Tier 1) - A program is classified as 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. A program may also be placed in Strong if it has multiple smaller randomized studies that, when pooled, have 350 or more students and either has statistically significant effects in each of those smaller randomized studies or pooled across all of those smaller randomized studies (see *Pooling Studies To Increase Sample Size*).

Moderate (Tier 2) - A program is classified as Moderate if it has a well-conducted, quasi-experimental design (QED) 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. A program may also be placed in Moderate if it has multiple studies that meet all other inclusion requirements for Strong or Moderate except for the sample size requirement, provided that the studies together include at least 350 students and either has statistically significant effects in each study or pooled across studies (see section on pooling). For QEDs, groups must be formed using all participants with baseline (pretest) data (i.e., prior to restricting the analysis sample to only those cases having complete outcome data). This requirement ensures that overall and differential attrition can be reported and evaluated. There are additional limitations and acceptable design variations concerning eligible Moderate studies; see *Acceptable Research Designs*.

Promising (Tier 3) - A program is classified as Promising if it obtained significant positive outcomes but did not meet the criteria for a Strong or Moderate rating. Common cases are obtaining significance at the (a) student level but failing to account for clustering, or (b) cluster

level but failing to meet the sample size requirements for Strong or Moderate. The Promising category also includes QEDs where the analytic sample only includes students having complete outcome data (e.g., pre- and post-test data), such that attrition and risk for attrition bias cannot be reported/assessed. Evaluations that use only public aggregated data may qualify for and be capped at Promising provided they meet specified research design standards for minimizing potential bias (see *Aggregated Public-Data Evaluations in* Section IV.B**).**

For all tiers, positive effects must not be overridden by statistically significant negative effects (see section on *No Overriding Negative Effects*).

B. Evidence for ESSA Additional Categories

No Studies Met Inclusion Requirements – If a program has no studies or a study of a particular program does not meet the inclusion criteria listed below, the program will be entered into the website with the "No Studies Met Inclusion Requirements" designation.

Qualifying Studies Found No Significant Positive Outcomes – If a study does meet inclusion criteria, but the acceptable study outcomes are not significant, the program will be entered into the website with the "Qualifying Studies Found No Significant Positive Outcomes" designation.

Program No Longer Disseminated – Programs that were previously reviewed but are believed to be no longer disseminated receive the "Program No Longer Disseminated" designation. If the program previously received a rating, that is noted.

III. Study Identification and Review Process

A. Procedures for 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 an ongoing comprehensive search of the literature by topic through a process that includes:

1. **An electronic search,** a web-based search including educational databases (e.g., ERIC, JSTOR), educational research sites, educational publishers' websites, and searches of recent meta-analyses.
2. **A hand search** of relevant peer-reviewed journals, and a final review of citations found in relevant documents retrieved from the first search wave.
3. **Publisher and developer submissions** sent to us by program developers, program evaluators, and educators.
4. **Review of state-approved lists** of programs and other lists of programs currently in use.

B. Conflict of Interest and Reviewer Independence

The following study-review situations may present potential conflicts of interest.

1. A study with an author employed by Johns Hopkins University (JHU), because the Center for Research and Reform in Education at Johns Hopkins University produces the Evidence for ESSA website.
2. A study that evaluates a program having an affiliation with JHU, such as when a JHU staff member is part of the development team.

When a potential conflict of interest exists, the following protocol is followed:

1. The study is sent for review to an external reviewer who has been trained in Evidence for ESSA review methods.
2. The external reviewer performs the review and returns the results to the Evidence for ESSA team.

IV. Inclusion Criteria for Studies

A. Program and Study Eligibility

To be eligible for review, studies must:

- Be conducted in U.S. schools and be available in English.
- Have been carried out from 2000 to present.
- Have a program duration of at least 10 weeks from the first day of student exposure to the program to the posttest, with the exception of Science (8 weeks) and Family Engagement (8 weeks).
- Compare an intervention group to a comparison group. Construction of groups can include studies using random assignment to conditions, QED assignment based on, for example, matched schools, classes, or students; or regression discontinuity designs.
- Be currently available to schools in a replicable form. This requirement disqualifies studies including extraordinary supports such as researchers providing tutoring, graduate students helping in class every day in ways that are not available to typical schools, or the program provider offering resources or support not available in other contexts.
- Have one or more dependent variables that are valid quantitative measures of student academic achievement, social-emotional learning, family engagement, or attendance. (See section on *Outcome Standards*.)

B. Acceptable Research Designs

- **Randomized controlled trials (RCT)**
- **Regression discontinuity designs (RDD).** RDD studies are reviewed using the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0, and may be considered for Strong.
- **QEDs (General Requirements):** Quasi-experimental design (QED) studies are eligible for review when treatment and comparison groups are constructed using a process other than random assignment. To be considered for inclusion, all QEDs must meet a common set of design requirements intended to reduce bias and support valid causal inference.

    - Studies must clearly document how treatment units were selected or assigned, specifically., the exact processes by which teachers, classrooms, or schools were chosen to receive the intervention. Assigning programs to units based on prior performance, prior implementation success, or perceived likelihood of strong implementation (for example, selecting "high-performing" or "high-capacity" teachers) can introduce systematic differences between treatment and comparison groups and therefore disqualify the study for acceptance.
    - The quality of the match between treatment and comparison groups must be demonstrated through baseline equivalence on relevant pre-intervention measures, including outcome-relevant pretests and key demographic characteristics, as described in the *Baseline Equivalence Requirements* section.
- **Baseline-Defined Analytic Sample QEDs (Eligible for Moderate)**: Baseline-defined analytic sample QED studies are eligible for a Moderate rating when the comparison group is created based on pre-intervention (baseline) characteristics (i.e., information measured before implementation begins). The comparison group must not be created or adjusted using post-baseline information including but not limited to program participation, level of implementation, completion of professional development, or availability of outcome data. Analytic groups must be defined from the full baseline sample (all participants with pretest/baseline data), prior to restricting the actual analytic sample to only those participants having observed outcomes, so that overall and differential attrition can be computed and evaluated. The treatment group must include all students who received any treatment exposure and must not be restricted by dosage or usage. The quality of the match must be demonstrated through pretest equivalence and comparable student demographic characteristics. (See section on *Baseline Equivalence*).
- **Outcome-Defined Analytic Sample QEDs (Limited to Promising)**: Outcome-defined analytic sample QED studies are eligible for a Promising rating when the analytic sample and/or groups are created *after* restricting the dataset to students with complete outcome data (e.g., only students with both pre- and post-tests), such that attrition cannot be calculated for the baseline-defined sample. The treatment group must be defined as all students who received any treatment and is not restricted by dosage or usage. The quality of the match must be demonstrated through pretest equivalence and comparable baseline characteristics. (See section on *Baseline Equivalence*).
- **Aggregated public-data evaluations (Limited to Promising):** Studies that rely solely on public, aggregated (school/district/state-level) outcome data to construct treatment and comparison groups may be considered only if the evaluation is conducted by an independent third-party evaluator who (a) defines treatment/comparison groups and any matching/weighting procedures using pre-specified, replicable rules, (b) documents and justifies all inclusion/exclusion criteria and any "treatment" definitions, and (c) provides explanation of how they reduced bias from outcome-informed selection, data-mining, or p-hacking[1] to the extent possible. Because outcomes are often observable prior to analysis (i.e., in a retrospective design) and the risk of bias is higher than in QED studies obtaining student-level data from districts or states, the rating is capped at Promising.

[1] Manipulating sampling to attain a significant probability (p) level.

Not accepted:

- Norm-referenced comparisons (e.g., comparing the treatment group gain to the average state-level gain)
- Pre-post only (studies of an intervention group without a comparison group).
- Treatment-on-the-treated (those that compare the impacts for students who received a specified "dose" of the treatment with students who received no treatment).
- Single-case or single-subject designs such as multiple-baseline designs across students/classrooms/settings or alternating-treatments designs, even when replicated across a number of cases.
- No business-as-usual comparison, such as comparing one intervention to another new practice.

C. Baseline Equivalence Requirements

- 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.
- Small randomized studies (fewer than 50 clusters for cluster-randomized studies or 350 students for student-randomized studies) must provide baseline data on student achievement. The average pretest difference for the analytic sample must not exceed 0.25 standard deviations.
- QEDs must demonstrate the quality of the match through comparable baseline characteristics of the analytic sample (e.g., after inverse weighting) and must include both of the following:
    - Pretest equivalence: For each outcome domain used to estimate program impacts, the matched treatment and comparison groups must be equivalent on a pre-intervention pretest in the same outcome domain (or, if unavailable, a broader pre-intervention measure in the same content area encompassing the outcome), with an absolute standardized mean difference < 0.25 SD.
    - Demographic and student-characteristic balance: Groups must be similar on key baseline student characteristics. At least three of the following variables must be included when assessing demographic balance: grade, race/ethnicity, disability status, English learner/dual language learner status, and economic disadvantage. For binary or categorial variables, balance must be assessed using an effect size calculated with the arcsine transformation (Lipsey & Wilson, 2001). The effect size must be < 0.5 SDs on each available demographic variable.
    -

D. Confounding and Internal Validity Safeguards

- No perfect confounds (e.g., all charter vs. all non-charter).
- No time-aligned confounds, 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[2].

[2] Evidence for ESSA is currently exploring design properties that would make successive cohort designs potentially eligible for review capped at Tier 3 status.

- No sample-aligned confounds. Studies that compare students or teachers from different schools must include at least two schools in the treatment condition and two schools in the comparison condition. For example, a one-school treatment versus one-school comparison design is not accepted because school membership is completely confounded with treatment status. The same restriction applies to any type of cluster assignment, such as comparing one classroom in a school to another classroom.
- Studies in which either the treatment or comparison group comprises less than 20% of the total sample will not be accepted unless the authors provide a plausible and detailed account of the selection or assignment process demonstrating the absence of systematic bias.

E. Attrition Standards

- RCTs and Baseline-Defined Analytic Sample QEDs: Differential attrition must be ≤ 15 percentage points from the baseline-defined sample to the final analytic sample, to qualify for Strong (RCT) or Moderate (QED) ratings.
- Outcome-Defined Analytic Sample QEDs: For QEDs in which analysis samples are restricted to only cases having complete outcome data, attrition cannot be determined for the baseline-defined sample; therefore, these studies are limited to a Promising rating.

F. Sample Size and Structural Minimums

- Strong and Moderate: 350-student minimum.
- Promising: 2 teachers and 30 students per condition minimum; 2 schools per condition when applicable (See *Confounding and Internal Validity Safeguards*).

G. Considerations for Preschool

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, preschool studies are reviewed based on whether the outcome measures provide a developmentally appropriate and fair test of program impacts. Developmentally appropriate outcomes (eligible for any tier) are those assessed using validated, age-appropriate measures of skills and competencies that are reasonably expected to be taught and learned during preschool (e.g., developmental language and literacy measures, early numeracy, self-regulation/executive function, social-emotional development) and the study may be eligible for any rating (Promising, Moderate, or Strong), whether measured at the end of preschool or at later follow-up points.

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.

V. Outcome Standards

A. Core Requirement

Studies must include a quantitative measure of student outcomes in:

- Academic achievement.
- Social-emotional learning (e.g., engagement, suspensions, emotional well-being, social skills; see *Social-Emotional Learning Outcomes*).
- Attendance (e.g., average daily attendance and chronic absenteeism; see *Attendance Outcomes*).
- Family Engagement.

Measures must:

- Be standardized or otherwise program-independent; if a developer- or researcher-created measure is used, the study must provide evidence of reliability/validity and demonstrate that the measure's design, selection, and scoring are unlikely to advantage the intervention group through over-alignment or other sources of bias. The study should also provide a rationale explaining why standardized or program-independent measures were not available, appropriate, or sufficient for assessing the intervention's targeted outcomes.
- Not be intrinsic to, or over-aligned with, the treatment content (i.e., not primarily a test of material uniquely taught/practiced in the intervention).
- Not be individually administered by interested parties: Individually administered measures given by the student's teacher or tutor solely for the purposes of the study or the intervention are never accepted. It is acceptable, however, to use a measure that is administered by teachers to students in intervention and comparison classrooms as part of the school's or district's regular assessment practices.
- Address developmentally appropriate skills in a domain that is plausibly targeted by the program's goals and theory of action. Outcomes should not be drawn from measures intended for substantially different age/grade levels (e.g., administering early-elementary phonics assessments to high school students) and should align with the program domain (e.g., a reading program evaluated on reading outcomes rather than unrelated math outcomes).

B. Social-Emotional Learning (SEL) Outcomes

Procedures are somewhat different for SEL studies because there are many quite diverse measures used in such studies. Although they are subject to the same requirements as above, outcomes in any one of four categories are accepted.

Social-Emotional Learning outcomes are organized into four domains:

1. Academic.
2. Problem Behaviors.
3. Social Relationships.
4. Emotional Well-Being.

Results from multiple SEL outcomes are aggregated and reported at the domain level as described in the sections on *Pooling and Aggregation Procedures* and *Program Classification and Website Presentation*.

C. Attendance Outcomes

There are two approved metrics in the Attendance category:

- Average Daily Attendance.
- Chronic Absenteeism (defined as missing >10% of school days each year).

Results from multiple Attendance outcomes are aggregated and reported at the domain level as described in the sections on *Pooling and Aggregation Procedures* and *Program Classification and Website Presentation*.

D. Qualitative Evidence (Supplemental Only)

While Evidence for ESSA program pages primarily describe the results of quantitative research studies measuring a program's impact on student achievement, some eligible studies also include qualitative findings on consumer experiences and perceptions during and after implementation. Potential users of a program can find meaningful information in these findings about the actual experience of implementing a program, including perceived strengths or challenges encountered by those who implemented the program. Therefore, if a program's quantitative research study meets Evidence for ESSA inclusion standards, qualitative findings (including survey, interview, or focus group data) within the same research paper will also be included if they meet eligibility criteria.

These criteria are:

- The qualitative data provide information that will assist in decision-making regarding product adoption or implementation.
- The qualitative research assessing stakeholder satisfaction must be conducted by an independent third party (e.g., school district or external evaluator), not by the intervention owner or developer.
- Surveys must have at least 20 respondents, and interviews or focus groups must have at least 10, and appear to reflect representativeness of the target group (e.g., teachers, students, parents).
- The study must describe the selection and engagement of respondents, the satisfaction instrument (survey, interview, and/or focus group) with enough detail to assess face validity (i.e., whether questions were appropriate/unbiased), and the method of analysis.
- The study describes the qualitative research component sample size, response rates, and findings comprehensively rather than selectively.
- For comparison analyses (e.g., studies comparing change in perceptions over time or studies comparing intervention and control group perceptions), the study must meet these criteria for all groups or time points in order for their outcomes to be considered for inclusion in Evidence for ESSA.

****Note: Evidence-strength determinations are based solely on *quantitative* impact evidence.**

VI. Statistical Standards

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.

Determination of Statistical Significance

1. **Student-level randomization → ANCOVA or equivalent**
   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. **Clustered designs → HLM or multilevel modeling**
   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.

3. **Recalculation if clustering ignored**
   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.

4. **Reanalysis ignoring clustering for possible Promising classification**
   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.

5. **Grade-level rules**
   If outcomes are analyzed by pooling across grade levels, the reported statistical significance is used. If outcomes are analyzed separately by grade level, a statistically significant positive result in at least one grade level qualifies as a positive effect, provided the grade-level analysis independently meets all other study requirements (e.g., baseline equivalence, sample size) for the relevant evidence tier.

6. **Multi-site and multi-cohort treated as mini-studies**
   In studies that have multiple randomized sites (states/districts), significant results from one state/district may qualify a program for a rating, but effect sizes will be averaged across sites in our reporting of the overall study and its outcomes. If overall effects are not significant, each site will be treated as a separate "mini-study" that must meet the criteria outlined in this document (including sample size requirements for Strong and Moderate). The same approach will be applied in studies with multiple cohorts, that is, distinct participant groups recruited and/or randomized in different waves or time periods

(e.g., separate school years, or separate grade-level cohorts randomized and analyzed as distinct cohorts). For example, if a researcher conducts a multi-site study of a literacy intervention in Kansas and Iowa, each site can serve as a "mini-study" if the results are not significant overall. If in this instance, the results from Kansas independently meet criteria for Moderate, while the results from Iowa do not yield significant positive or negative effects, the Kansas "mini-study" analysis will qualify the study for Moderate status, though reported results will include all results and the aggregate effect size for the intervention.

7. **Universal vs. at-risk subgroup handling**
   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 be treated as its own "mini-study" that must meet the criteria outlined in this document (including sample size requirements for Strong and Moderate).

8. **No overriding negative effects**
   For a program to receive a rating of Strong, Moderate, or Promising, statistically significant positive effects must not be overridden by statistically significant negative effects within the same outcome domain. This rule applies:
   - Across outcomes within a study
   - Across grade levels within a study
   - Across sites within a multi-site study
   - Across cohorts within a study
   - Across studies of a program

VII. Effect Size Calculation Standards

A. Approved Formulas

- Glass' delta/Cohen's d / Hedges' g acceptable.
- No adjusted SD denominators: Standard deviations adjusted for pretests or other covariates may not be used as the denominator of the effect size formula.
- No gain-score SDs: SDs of gain scores may not be used. Only unadjusted SDs are acceptable.

B. Effect Size Conversions

- Pre–post difference-in-differences for effect-size estimation: When a study reports unadjusted pretest and posttest group differences but does not report an adjusted posttest impact estimate, we may compute an effect size using a change-score difference approach: $ES_{post} - ES_{pre}$.
- Lipsey & Wilson procedures: Lipsey & Wilson (2001) 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.

C. Directionality Rules

- Reverse negative behavior outcomes: 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).
- Positive always indicates desirable direction for SEL and Attendance: 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.

VIII. Pooling and Aggregation Procedures

A. Pooling Effect Sizes Within Single Studies

- Using overall/composite scores vs. subscales: The study-level effect size for an outcome domain (e.g., reading, math) is generally based on the overall/composite test score for that domain when it is reported (e.g., a state assessment's overall mathematics score or an overall standardized-test composite). If the study reports only subscale scores and does not provide a domain composite, we will compute a domain effect size by combining the relevant subscales (e.g., combining Algebra and Geometry subscale impacts into a single math-domain effect size).
- When combining multiple subscales (or multiple measures within a domain), each outcome is weighted by the analytic sample size for that outcome.

B. Aggregating Effect Sizes Across Multiple Studies

- **One effect size per program**: A program's rating is based on results from one or more qualifying studies/outcomes. On the Evidence for ESSA website, we report a single average effect size for each program by averaging the effect sizes from all eligible studies that meet inclusion criteria, regardless of statistical significance.

- **Inverse variance weighting**: After a single effect size is computed for each study within a subject, then effect sizes are averaged across studies, weighted by sample size using an inverse variance procedure.

C. Pooling Studies to Meet Sample Size Requirements

One of the requirements to receive a rating of either Strong or Moderate is having a total sample size of at least 350 students. 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 or Moderate, the following requirements must be met:

- **Conditions for Strong evidence pooling**:
  - **Study Design Requirements**: all pooled studies must meet the research design requirements for Strong with the exception of sample size.
  - **Statistical Significance:** all pooled studies that meet the research design requirements for Strong must each have statistically significant effects on their own; OR have a statistically significant effect across the pooled studies, which is calculated using a fixed-effects meta-analysis.

- **Conditions for Moderate evidence pooling**
  - **Study Design Requirements:** All pooled studies must meet the research design requirements for either Moderate or Strong with the exception of sample size.
  - **Statistical Significance**: All pooled studies that meet the research design requirements for either Moderate or Strong must each have statistically significant effects on their own; OR have a statistically significant effect across the pooled studies, which is calculated using a fixed effects meta-analysis.

D. Aggregation of SEL and Attendance Outcomes

Social-emotional learning (SEL) and attendance studies often include multiple outcomes measuring different aspects of student functioning. To provide a meaningful summary while preserving distinctions among outcome types, Evidence for ESSA aggregates outcomes within domains rather than across all SEL or attendance measures combined.

SEL and Attendance domains are described in Sections V.B and V.C.

When multiple outcomes are reported within a domain or attendance metric, effect sizes are aggregated using the procedures described above. The resulting average effect size represents the overall impact of the program within that domain or attendance metric.

Evidence levels (Strong, Moderate, or Promising) are determined separately from effect-size aggregation. A domain or attendance metric is assigned the highest evidence level achieved by any qualifying outcome within that domain. For example, if one outcome within the Social Relationships domain meets the criteria for Strong evidence and another outcome within the same domain meets only the criteria for Promising evidence, the Social Relationships domain is classified as Strong.

Because domain-level effect sizes are calculated using all eligible outcomes within a domain, it is possible for a domain to achieve a high evidence rating while having a relatively small average

effect size. This may occur when one or more outcomes show statistically significant positive effects while other outcomes in the same domain show smaller, null, or negative effects.

IX. Program Classification and Website Presentation

A. Placement Within Categories

On the Evidence for ESSA website, programs are categorized as Strong, Moderate, or Promising in accord with the foregoing standards. 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. Collective sample size across all qualifying studies.
3. Weighted mean effect size across all qualifying studies.

B. Presentation of Social-Emotional Learning and Attendance Results

Because social-emotional learning (SEL) and attendance programs often target multiple distinct outcomes, results are presented by domain rather than as a single overall program rating.

SEL and Attendance domains are described in Sections V.B and V.C.

The evidence level and average effect size reported for each domain or attendance metric are determined using the aggregation procedures described in *Aggregation of SEL and Attendance Outcomes*. When a program has evidence in multiple domains, each domain is displayed separately to allow users to identify the specific areas in which evidence of effectiveness has been demonstrated.

C. Badged Programs 

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

X. Updating, Re-Review, and Evidence Revisions

- **Ongoing acceptance of new studies:** Researchers, program developers, and others continually submit new studies for review by emailing us at evidenceforessa@jhu.edu. Studies do not have to be published in order to be considered, however they do need to include the information outlined in our inclusion criteria. When we receive new studies about programs that are already posted, this study evidence is evaluated and then incorporated into a program's page as appropriate.

- **Periodic re-review:** For all programs included on the website, we record the date and details of each review and periodically search for new evidence. Programs and their associated studies may be reclassified when Evidence for ESSA standards are updated or when additional studies are identified..
- **Version transparency:** We post our current Standards and Procedures on the website, as well as previous versions.
- We encourage users to notify us at evidenceforessa@jhu.edu if they find errors or inconsistencies; these may be from prior standards that we are in the process of updating.
- Appeals: If applicants disagree with one of our findings, we encourage them to let us know so that we can address their concerns.

For Additional Information

Susan Davis, Managing Editor, Evidence for ESSA
evidenceforessa@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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications, Inc.

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.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). (2022). *What Works Clearinghouse procedures and standards handbook* (Version 5.0). https://ies.ed.gov/ncee/wwc/Handbooks