Evidence for ESSA: Standards and Procedures

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). In consultation with our Technical Work Group (TWG), we developed policies to apply the ESSA standards to the evidence available for all programs currently available to schools in the U.S.

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. Strong, moderate, and promising categories are defined as follows (in brief):

1. Strong: At least one randomized, well-conducted study showing significant positive student outcomes, and no studies showing significant negative outcomes.
2. Moderate: At least one quasi-experimental (i.e., matched), well-conducted study showing significant positive student outcomes, and no studies showing significant negative outcomes.
3. Promising: At least one correlational, well-conducted study with controls for inputs showing significant positive student outcomes, and no studies showing significant negative outcomes.

Procedures

Finding Eligible Studies

A comprehensive search of the literature by topic is carried out through a multi-step process that includes an electronic database search, a web-based search of educational research sites and educational publishers' websites, an ancestral 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.

Inclusion Criteria for Studies

1. Studies must be of programs available today to schools in the U.S.
2. Studies have to have been carried out from 1990 to the present, and from 2000 to the present if they evaluated technology approaches.
3. Studies have to have compared experimental groups to control groups. Either random assignment to conditions or matched, quasi-experimental assignment based on pre-specified schools, classes, or students had to be used. After-the-fact (post hoc) matching is not acceptable, and comparisons to norming groups, pre-post comparisons, or other non-experimental comparisons are not accepted. Comparisons of two equally innovative approaches, without a control group representing ordinary practice, are not accepted.

4. Studies have to provide pretest data to establish initial equivalence. On achievement measures, the average pretest difference could not exceed 25% of a standard deviation. Studies have to establish equivalence before the experiment, and also equivalence at pretest for the remaining sample after attrition at the end of the study.
5. Studies' dependent variable(s) have to include a quantitative measure of academic achievement. The measure could be a standardized test or a test created by test developers not involved with the research, but tests made by 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 are not accepted.
6. Study durations have to be at least 12 weeks, from program inception to posttest.
7. Studies have to have at least 2 teachers and 30 students per treatment. Also, at least two schools per treatment are required when randomization/matching of students/teachers takes place outside of a single school.
8. 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. Also, if attrition causes the pretests of the final sample to differ by more than 25% of a standard deviation, the study is rejected.
9. Studies have to have used 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 is 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 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 methods accounting for clustering.

    a. If HLM was not used, we use a formula in the What Works Clearinghouse Standards 3.0 document that recalculates statistical significance accounting for clustering.

b. 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.

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 (or a pooled SD if the control group SD is not available):

$$ES = \frac{X_E - X_C}{SD_C}$$

Standard deviations already 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, odds ratios, and other statistics.

Pooling Effect Sizes

Effect sizes are pooled at the study level and the program level, to find the average effect for a program.

At the study level, the overall effect size is generally the reading total or math total. For example, if the study reports GRADE/GMADE or PARCC, we would calculate total reading or math. Otherwise, if only separate subscales are reported, we combine appropriate measures.

For main measures (e.g., state test total, GRADE, GMADE, comprehension, computations, concepts), effect sizes are included at full value. For measures that are less central, such as vocabulary, these effect sizes are weighted half as much as the more central measures. Measures given at grade levels far from usual (e.g., phonics measures in secondary schools) are not accepted. Table 1 summarizes our treatment of reading measures.

Table 1
Treatment of Types of Reading Outcome Measures*

Measure	PreK	K	Grades 1-2	Grades 3-5	Grades 6-12
Total Reading	Full	Full	Full	Full	Full
Comprehension	Full	Full	Full	Full	Full
Vocabulary (*not* PPVT)	Half	Half	Half	Half	Half
Phonics (Word Attack, Word ID)	Full	Full	Full	Half	Zero
Fluency	Full	Full	Full	Half	Half
Phonological Awareness (e.g., CTOPP)	Full	Half	Zero	Zero	Zero
Language Arts (as a reading measure)	Zero	Zero	Zero	Zero	Zero

*Individually administered measures given by the teacher are never accepted.

Acceptable tests include, for example: GMRT, GRADE, GORT, Woodcock-Johnson, CST, CAT, MAP, ERDA, STAR, ITBS, Terra Nova, CTBS, SAT, iSAT, ISTEP, SDRT, DRP, ETS, NWEA, TOSREC, TERA, Durrell, WIAT, DIBELS, AIMSweb, and state standardized tests

A single effect size is computed for each study, and then effect sizes are averaged across studies, weighted by sample size using an inverse variance procedure.

Definitions of "Strong," "Moderate," and "Promising" ESSA Categories

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

Strong- A program is placed in "strong" if it has a statistically significant positive effect on at least one major measure (e.g., state test or national standardized test) analyzed at the proper level of clustering (class/school or student). Programs with one significantly positive study are not listed as "strong" if there is also at least one study with a significantly negative effect.

Moderate- 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 quasi-experiments (i.e., matched studies).

Promising – Programs with at least one correlational study with controls for inputs may be placed in the "promising" category. Also, programs that would have qualified for "strong" or "moderate" but did not qualify because they failed to account for clustering (but did obtain significantly positive outcomes at the student level) may qualify for "promising" if there are no significant negative effects.

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 place programs within ESSA evidence categories according to an algorithm that emphasizes the following:

1. Weighted mean effect size, across all qualifying studies.
2. Number and quality of studies.
3. Collective sample size across all qualifying studies.
   The exact algorithm for each qualifying study is as follows: (ES x 100) (2 if randomized; 1.5 if quasi) (1 + SQRT [N/100]).

Badged Studies

In the "strong" category, we put a "badge" on studies with particularly strong evidence. These are programs that meet the following criteria:

a) At least two studies meeting the "strong"category.
b) Weighted mean effect size across all qualifying studies of +0.20 or more, or two studies each of which has an effect size of at least +0.20.

Additional information

Robert E. Slavin
Director, Center for Research and Reform in Education, Johns Hopkins University
rslavin@jhu.edu or 410-616-2310