WWC EVIDENCE REVIEW PROTOCOL FOR SCIENCE INTERVENTIONS, Version 2.0

Topic Area Focus

This What Works Clearinghouse (WWC) review focuses on science interventions designed to impact student achievement, including curriculum-based interventions, instructional techniques, and products designed to deliver content and monitor student progress. Systematic reviews of evidence in Science address the following questions:

- Which interventions are effective in increasing students' learning of science content, processes, and skills?
- Are some interventions more effective for certain types of students, particularly students who are members of traditionally underserved populations in science education?

In this review, a *science intervention* is defined as a replicable instructional program, practice, or strategy that clearly delineates science learning goals for students and is designed to directly affect student science achievement.

Outcomes that fall in the science achievement domain are those related to science content and skills, commonly described as what students should know and be able to do. Science content varies across curricula and grade levels, but it generally includes life science, Earth/space science, and physical science. Science processes and skills are the application of the learning of this content, as well as a demonstrated understanding of science concepts and procedures and science inquiry.

ELIGIBILITY CRITERIA AND EVIDENCE STANDARDS

Populations to be Included

The WWC Science area reviews include science curricula, techniques, and products that were developed for students in grades K–12 (ages 5–18).

If students with learning disabilities comprise more than 50% of the sample, a study will be eligible for review in the Students with Learning Disabilities topic area review rather than the Science area.

The study must have been conducted in the United States or any other country that used English language science curriculum materials available for use in the United States.

An intervention's effectiveness could vary by subgroups of student or school characteristics. Whether a study examines effects on subgroups does not affect the inclusion of the study for review or the rating given to the study. However, WWC reports will present in an appendix

findings for subgroups of interest, provided that the subgroups are equivalent with regard to pretest scores and grade level. Student characteristics of interest for this review include baseline science achievement, grade, gender, socioeconomic status, racial/ethnic breakdown, percentage of English as a second-language students, percentage of bicultural students, and underserved population status (as provided by study authors). School settings of interest for this review include location of the schools involved, homogeneous groupings of students, school type (public, private, religious), school SES (e.g., Title I school), average class size (small, medium, large), average teacher characteristics (e.g., teacher education and experience), Urban-Centric Locale Codes (city, suburb, town, rural) that could moderate effects, and school size.

Types of Interventions to be Included

The interventions included are determined after an exhaustive search of the published and unpublished literature by the Science review team, as well as a review of nominations submitted to the WWC. Only research on interventions that are replicable is reviewed.

> *Replicable.* For an intervention to be considered replicable, it must be either a "branded" intervention or an unbranded intervention, practice, or strategy that meets the following conditions:
>
> (1) the intervention is described in general terms, such as process(es) and/or skill(s) being targeted, approach to enhancing the process(es) and/or skill(s), targeted population, unit of delivery of the intervention (i.e., whole group, small group, or individual student), and medium/media of delivery of the intervention (i.e., teacher-led instruction or software),
>
> (2) the duration of the intervention is described, and
>
> (3) the characteristics of the individuals who are expected to deliver the intervention are described.

Examples of possible interventions include textbooks and textbook series, software programs, hands-on kit-based science materials, and other educational technology that serves as the basis for well-defined curricula, as well as instructional practices and strategies, such as university summer programs for young scholars, math-science partnership programs, or student trips to museums to participate in scientific programs and activities. Although the Science review will focus primarily on studies of interventions used during the regular school day, studies of interventions used during afterschool programs will also be eligible for review.

Types of Research Studies to be Included

The study must have become publicly available in January 1990 or later and obtained by the WWC prior to drafting the intervention report.

The design must be an empirical study, using quantitative methods and inferential statistics. Eligible designs include well-conducted randomized controlled trials (RCTs), quasi-experimental designs (QEDs) with matching or equating of student samples on a baseline student-level measure, single-case designs (SCDs), and regression discontinuity designs (RDs).

Types of Outcomes to be Included

Science achievement domain. Outcomes that fall in the science achievement domain are those related to science content, practices, and skills, commonly described as what students should know and be able to do. Science content varies somewhat across curricula and grade levels, but it generally is covered by the three major constructs: life science, Earth/space science, and physical science. Science practices include identifying science principles, using science principles, using science inquiry, and using technological design (NAEP Science Framework, 2009)[1]. Science skills are demonstrated by the application of students' understanding of the science content, as well as an understanding of science concepts, inquiry, procedures, and problem solving. These include reasoning and proof, making connections, oral and written communication, and use of scientific representation.

The study needs to include at least one type of relevant science achievement outcome. Findings in WWC reviews report only on these outcomes.

> *Relevant outcomes* are measures of science achievement, including standardized, nationally-normed achievement tests that are appropriate for K–12 students, standardized state or local tests of science achievement, and research-based or locally developed tests or instruments that assess students' science concepts or skills.

The study must include at least one outcome measure that demonstrates sufficient reliability or face validity.

> *Reliability* for group-design studies will be assessed using the following standards determined by the WWC: internal consistency (minimum of 0.50), temporal stability/test-retest reliability (minimum of 0.40), or inter-rater reliability (minimum of 0.50).

A study will be rated based only on those measures (if any) that are not overaligned.

> *Overalignment* occurs with outcome measures that are more closely aligned to one of the research groups (intervention or comparison) than the other and could bias a study's results. For instance, if the outcome measure assesses science achievement using some of the same materials included in the intervention (such as specific problems), it is considered to be overaligned with the intervention. In these situations, the intervention group may have an advantage over the comparison group, and the size of the intervention's measured effects may be incorrect. A science intervention may have an immediate effect as well as a longer-term effect on student science achievement. Thus, outcomes measured at the end of an intervention, as well as those measured any time thereafter, are included**.** The end-of-intervention outcome measure will be used to determine the overall impact of the intervention. Delayed measures taken several months or years after an intervention may be useful because they may provide strong evidence for an intervention's longer-term effectiveness.

---

1 U.S. Department of Education, National Assessment Governing Board. (2008, September). Science Framework for the 2009 National Assessment of Educational Progress. Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf.

Design Ratings

Sample attrition is a key factor in determining the WWC rating for RCTs. Baseline equivalence on measures of the outcome variable or factors correlated with the outcome measure is a key factor in determining the WWC rating for QEDs and RCTs with high attrition.

Attrition in RCTs. The WWC considers both the overall sample attrition rate and the differential in sample attrition between the intervention and comparison groups, as both contribute to the potential bias of the estimated effect of an intervention. The WWC has established conservative and liberal standards for acceptable levels of attrition. The conservative standards are applied in cases where the principal investigator (PI) has reason to believe that much of the attrition is endogenous to the intervention reviewed—for example, high school students choosing whether or not to participate in a drop-out prevention program. The liberal standards are applied in cases where the PI has reason to believe that much of the attrition is exogenous to the intervention reviewed (e.g., in cases where movement of young children in and out of school districts due to family mobility). Attrition rates are based on the number of sample cases used in the analysis sample with measured, as opposed to imputed, values of the outcome measures. Table 1 presents the maximum difference in the attrition rate for the treatment and comparison group that is acceptable for a given level of overall sample attrition. The empirical basis for these thresholds is described in Appendix A of the *WWC Procedures and Standards Handbook, Version 2.0*.

Studies based on cluster random assignment designs must meet attrition standards for both the study sample units that were assigned to treatment or control group status (e.g., schools or districts) and the study sample units for analysis (e.g., typically, students). In applying the attrition standards to the subcluster level (e.g., students), the denominator for the attrition calculation includes only sample members in the clusters that remained in the study sample.

RCTs with combinations of overall and differential attrition rates that exceed the applicable threshold, based on the applicable standard, must demonstrate baseline equivalence of the analysis sample, or, if non-equivalence falls within the allowable range, statistically control for the nonequivalence, in order to receive the second-highest rating: *meets WWC evidence standards with reservations*. See the Baseline Equivalence section for more details.

Table 1: Attrition Standards for Randomized Controlled Trials

Highest Level of Differential Attrition Allowable to Meet the Attrition Standard Under the Liberal Attrition Standard

Overall Attrition	Allowable Differential Attrition	Overall Attrition	Allowable Differential Attrition
0	10.0	34	7.4
1	10.1	35	7.2
2	10.2	36	7.0
3	10.3	37	6.7
4	10.4	38	6.5
5	10.5	39	6.3
6	10.7	40	6.0
7	10.8	41	5.8
8	10.9	42	5.6
9	10.9	43	5.3
10	10.9	44	5.1
11	10.9	45	4.9
12	10.9	46	4.6
13	10.8	47	4.4
14	10.8	48	4.2
15	10.7	49	3.9
16	10.6	50	3.7
17	10.5	51	3.5
18	10.3	52	3.2
19	10.2	53	3.0
20	10.0	54	2.8
21	9.9	55	2.6
22	9.7	56	2.3
23	9.5	57	2.1
24	9.4	58	1.9
25	9.2	59	1.6
26	9.0	60	1.4
27	8.8	61	1.1
28	8.6	62	0.9
29	8.4	63	0.7
30	8.2	64	0.5
31	8.0	65	0.3
32	7.8	66	0.0
33	7.6	67	-

Baseline Equivalence. RCTs with high attrition and all QEDs must demonstrate baseline (that is, pre-intervention) equivalence between the intervention and comparison groups in the analysis sample in order to receive the rating of *meets WWC evidence standards with reservations*. Baseline equivalence is examined on measures of the outcomes or baseline measures that are expected to be highly correlated with these outcomes. For the Science review, these variables are a pretest of an acceptable outcome measure and grade level.[2] When a pretest of a science achievement outcome measure is not available, any one of mathematics, reading achievement, or literacy outcome measures can be used instead.[3] When multiple alternative outcomes are available (for example, reading achievement and math scores), baseline equivalence should be demonstrated for each outcome measure.

Groups are considered equivalent if the reported differences in mean baseline characteristics of the groups are less than or equal to 5% of the pooled standard deviation in the sample. If this is the case, the equivalence standard is met, and the study can receive a rating of *meets WWC evidence standards with reservations.* Statistical significance of the difference in means is not considered.

If differences are greater than 5% and less than or equal to 25% of the pooled standard deviation in the sample, the study findings must be based on analytic models that control for the individual-level baseline characteristic(s) on which the groups differ in order to receive a rating of *meets WWC evidence standards with reservations*. Otherwise, the study is rated *does not meet WWC evidence standards*.

Studies with baseline differences greater than 25% of the pooled standard deviation do not meet the baseline equivalence standard, regardless of whether or not the impacts are estimated using models that control for baseline characteristics. The study is rated *does not meet WWC evidence standards.*

Finally, when there is evidence that the populations being compared are drawn from very different settings (such as rural versus urban, or high-SES versus low-SES), these settings may be deemed too dissimilar to provide an adequate comparison. In these cases, the study is rated *does not meet WWC evidence standards*.[4]

---

2 Baseline science achievement tends to be highly correlated with other characteristics that can moderate effects and, therefore, tends to be a useful measure for assessing baseline equivalence.

3 There is an ample correlation between the math and science scores in the United States. At the eighth grade, for example, correlation coefficients range from 0.61 to 0.78 (Wang, 2005). Similarly, the ACT Science Reasoning test correlates 0.76 or 0.75 with each of the other ACT scores: Reading, English, and Mathematics (Dorans, 1999). The English language arts/literacy and math scores are utilized most often to determine equivalence across schools, classrooms, students, and/or districts, since most states, districts, and/or schools do not test science every year.

4 The Science review team also will examine other baseline characteristics (when available) to assess baseline equivalence of studies. These characteristics include, but are not limited to, gender, race/ethnicity, percentage of English as a second-language students, measures of underserved population status, tracking level, special education, school location, and average class size. The provision of all such information, however, is not a requirement of the review.

Statistical and Analytical Issues

RCT studies with low attrition do not need to use statistical controls in the analysis, although statistical adjustment for well-implemented RCTs is permissible and can help generate more precise effect size estimates. For RCTs, the effect size estimates will be adjusted for differences in pre-intervention characteristics at baseline (if available) using a difference-in-differences method if the authors did not adjust for pretest (see Appendix B of the *Handbook, Version 2.0*). Beyond the pre-intervention characteristics required by the equivalence standard, statistical adjustment can be made for other measures in the analysis as well, although they are not required.

For the WWC review, the preference is to report on and calculate effect sizes for post-intervention means adjusted for the pre-intervention measure. If a study reports both unadjusted and adjusted post-intervention means, the WWC review will report the adjusted means and unadjusted standard deviations.

The statistical significance of group differences will be recalculated if (a) the study authors did not calculate statistical significance, (b) the study authors did not account for clustering when there was a mismatch between the unit of assignment and unit of analysis, or (c) the study authors did not account for multiple comparisons when appropriate. Otherwise, the review team will accept the calculations provided in the study.

When a misaligned analysis is reported (i.e., the unit of analysis in the study is not the same as the unit of assignment), the statistical significance of the effect sizes computed by the WWC will incorporate a statistical adjustment for clustering. The default intra-class correlation used for the Science review is 0.20. For an explanation about the clustering correction, see Appendix C of the *Handbook, Version 2.0*.

When multiple comparisons are made (i.e., multiple outcome measures are assessed within an outcome domain in one study) and not accounted for by the authors, the WWC accounts for this multiplicity by adjusting the reported statistical significance of the effect using the Benjamini-Hochberg correction. See Appendix D of the *Handbook, Version 2.0* for the formulas the WWC uses to adjust for multiple comparisons.

All standards apply to overall findings as well as analyses of subsamples.

LITERATURE SEARCH METHODOLOGY

The literature search strategy for the WWC Science review has two components. First, the review team will conduct a search to identify interventions with studies that may be eligible for review. Second, the team will conduct focused intervention searches to ensure that all potentially eligible studies of the selected interventions are identified. Both search types are described below.

Keyword Search

The primary objective of the keyword search is to identify interventions with potentially eligible studies and assess the likely extent of studies on each intervention so that interventions can be prioritized for review. The focus will be on breadth rather than depth. The keywords are meant to capture literature that falls within the scope of the protocol. Given the objective stated above, targeted outcomes and study design terms will be included to focus the search on identifying literature that will support an intervention report. The keyword list (TBD) is followed by a list of databases that are searched.

The core list of electronic databases that are searched includes the following:

- ***ERIC.*** Funded by the U.S. Department of Education (ED), ERIC is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All ED publications are included in its inventory.

- ***PsycINFO.*** PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in the field of psychology. Journal coverage, which dates back to the 1800s, includes international material selected from more than 1,700 periodicals in more than 30 languages. More than 60,000 records are added each year.

- ***Campbell Collaboration.*** C2-SPECTR (Social, Psychological, Educational, and Criminological Trials Register) is a registry of more than 10,000 randomized and possibly randomized trials in education, social work and welfare, and criminal justice.

- ***Dissertation Abstracts.*** As described by Dialog, Dissertation Abstracts is a definitive subject, title, and author guide to virtually every American dissertation accepted at an accredited institution since 1861. Selected master's theses have been included since 1962. In addition, since 1988, the database has included citations for dissertations from 50 British universities that have been collected by and filmed at the British Document Supply Centre. Beginning with DAIC Volume 49, Number 2 (Spring 1988), citations and abstracts from Section C, Worldwide Dissertations (formerly European Dissertations), have been included in the file. Abstracts are included for doctoral records from July 1980 (Dissertation Abstracts International, Volume 41, Number 1) to the present. Abstracts are included for master's theses from spring 1988 (Masters Abstracts, Volume 26, Number 1) to the present.

- ***Academic Search Premier.*** This multidisciplinary database provides full text for more than 4,500 journals, including full text for more than 3,700 peer-reviewed titles. PDF backfiles to 1975 or further are available for well over 100 journals, and searchable cited references are provided for more than 1,000 titles.

- ***EconLit.*** EconLit, the American Economic Association's electronic database, is the world's foremost source of references to economic literature. The database contains more than 785,000 records from 1969 to the present. EconLit covers virtually every area related to economics.

- ***Business Source Corporate.*** Contains full text from nearly 3,000 quality business and economics magazines and journals (including full text of many only abstracted in other sources we search). Information in this database dates as far back as 1965.

- ***SocINDEX with Full Text.*** SocINDEX with Full Text is the world's most comprehensive and highest quality sociology research database. The database features more than 1,986,000 records with subject headings from a 19,600+ term sociological thesaurus designed by subject experts and expert lexicographers. SocINDEX with Full Text contains full text for 708 journals dating back to 1908. This database also includes full text for more than 780 books and monographs, and full text for 9,333 conference papers.

- ***EJS E-Journals.*** Electronic journals (E-Journals) from EBSCO host® provide article-level access for thousands of E-Journals through EBSCO's Electronic Journal Service (EJS). This resource covers journals to which MPR subscribes.

- ***Education Research Complete***. Education Research Complete is the definitive online resource for education research. Topics covered include all levels of education from early childhood to higher education, and all educational specialties, such as multilingual education, health education, and testing. Education Research Complete provides indexing and abstracts for more than 1,840 journals, as well as full text for more than 950 journals and full text for more than 81 books and monographs and for numerous education-related conference papers.

- ***WorldCat***. WorldCat is the world's largest network of library content and services and allows users to simultaneously search the catalogs of more than 10,000 libraries, containing more than 1.2 billion books, dissertations, articles, CDs, and other media.

- ***Google Scholar***. Google Scholar provides a simple way to broadly search for scholarly literature. From one place, users can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts, and articles from academic publishers, professional societies, preprint repositories, universities, and other scholarly organizations.

In addition to the keyword search in databases, the review team seeks to identify other relevant studies through the following approaches:

- Public submissions of materials via the WWC website or directly to WWC staff.

- Solicitations made to key researchers by the review team.

- Checking websites summarizing research on programs in science, prior literature reviews, and research syntheses (i.e., using the reference lists of prior reviews and research syntheses to make sure key studies have not been omitted).

- Searches of the websites of all the developers of relevant interventions or practices for any research or implementation reports.

- Searches of the websites of the following think tanks, research centers, and associations:

  ABT Associates
  Alliance for Excellent Education
  American Association for the Advancement of Science
  American Association of Physics Teachers
  American Enterprise Institute
  American Institutes for Research (AIR) Appalachian Education Laboratory (Edvantia)
  Best Evidence Encyclopedia (BEE)
  Broad Foundation (Education)
  Brookings Institution Carnegie Corporation
  Center for Comprehensive School Reform and Improvement
  Center for Data-Driven Reform in Education (CDDRE) at Johns Hopkins University
  Center for Research and Exploration in Space Science and Technology (CRESST)
  Center for Research and Reform in Education (CRRE) at Johns Hopkins University
  Center for Research in Educational Policy (CREP)
  Center for Social Organization of Schools at Johns Hopkins University
  Center on Education Policy
  Center on Instruction
  Chapin Hall Center for Children
  Consortium for Policy Research in Education (CPRE) at the University of Wisconsin-Madison
  Congressional Research Service
  Government Accountability Office
  Harvard Graduate School of Education
  Heritage Foundation
  Hoover Institution
  Horizon Research Inc.
  Inverness Research
  Institute for Higher Education Policy
  Institute for Public Policy and Social Research (IPPSR)
  Johns Hopkins University School of Education
  Learning Point Associates
  Mathematica Policy Research

MDRC
Mid-Continent Research for Education and Learning
National Association for Bilingual Education (NABE)
National Association of State Boards of Education
National Center on Secondary Education and Transition
National College Access Network
National Dropout Prevention Centers
National Governors' Association
National Science Foundation (NSF)
National Science Resources Center (NSRC)
National Science Teachers' Association (NSTA)
Pacific Resources for Education and Learning (PREL)
Pathways to College Network
Public Education Network
Public Policy Research Institute at Texas A&M University Public/Private Ventures (PPV)
Rand Corporation
Regional Educational Laboratories (RELs)
Southwest Educational Development Laboratory (SEDL)
SRI
The Education Resources Institute
The University of California, Los Angeles (UCLA)
Thomas B. Fordham Institute Urban Institute
U.S. Department of Education (includes Institute of Education Sciences)
Wisconsin Center for Education Research (WCER)
WestEd

References resulting from these searches will be screened and sorted by intervention.

Intervention Search

The primary objective of the intervention search is to identify ALL effectiveness studies conducted for a specific intervention identified in the keyword search, as well as any that the keyword search did not identify. The strategy for the search is as follows:

- If the intervention was reviewed under different WWC topic areas, re-review all references against the protocol for this topic area.
- Conduct standard library searches of the intervention name.[5]
- Scan references to identify possible synonyms for the intervention in the literature and conduct standard library searches of these terms.

5 A standard library search consists of searching titles and abstracts in each of the databases described above.

- Once potentially eligible studies are identified, request full text and review the reference lists to cross-check search results. Similarly, review relevant literature reviews. Revise search terms as needed.
- Identify seminal researchers associated with the intervention. Conduct full-text searches of the researcher name combined with the intervention name.
- Identify seminal studies of the intervention and conduct searches of the associated citation.
- Contact the intervention's developer for a list of known research on the intervention.

All references resulting from these searches will be screened for eligibility.

References

Dorans, N. J. (1999). *Correspondences between ACT™ and SAT® I scores.* College Board Report No. 99-1 ETS RR No. 99-2. College Entrance Examination Board, New York.

Wang, J. (2005). Relationship between mathematics and science achievement at the 8th grade. *International Journal of Science and Mathematics Education, 5*, 1–17.