

Study Review Protocol

Version 4.1 (October 2024)

What Works
Clearinghouse™

NCEE
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation at IES

The What Works Clearinghouse (WWC) Study Review Protocol guides the reviews of studies by the WWC. This protocol is used in conjunction with the *WWC Standards Handbook, Version 4.1* and the *WWC Procedures Handbook, Version 4.1*.

PURPOSE STATEMENT

WWC reviews provide education practitioners and policymakers with timely and objective assessments of the quality of research on a particular topic.

The WWC uses this protocol to review two categories of studies. The first category of studies includes those cited as evidence for U.S. Department of Education grant competitions, studies that were funded by the Department, and studies that have received significant media attention. The second category of studies includes those identified for systematic reviews of evidence based on a search of the research literature in a particular topic area. In those instances, an accompanying topic area review protocol will provide the research questions, the parameters for the selection of studies, the prioritization of studies for review, and information about how findings will be synthesized.

ELIGIBILITY CRITERIA

Eligible Populations

To be eligible for review under this protocol, the study must examine the effectiveness of an intervention administered to:

- Students and other learners in early intervention programs for infants and toddlers; in preschool education programs for children ages 3 through 5; or in elementary, secondary, postsecondary, or adult education programs. In this protocol, “student” and “learner” may be used interchangeably.
- Teachers, school leaders, other educators, or home- or school-based service providers.

There are no minimum sample size, or composition, requirements for studies reviewed with this protocol.

Eligible Interventions

The study must examine an educationally relevant or school-based intervention. The WWC defines the term “intervention” broadly, and this term may comprise education practices, products, policies, and programs. Therefore, the following types of interventions may be included, which are not mutually exclusive:

- **Practices.** Education practices are discrete, clearly defined activities focused on improving student learning and related outcomes. Practices may be used with a broad and diverse range of participants to address a wide range of learning goals; or they may be targeted to address a specific learning goal, skill, or population. An example of a practice is pre-teaching new vocabulary.

- **Products.** Education products are “branded” or commercial interventions such as curricula or software. Products may be used as the primary instructional tool in the classroom or to supplement classroom material with differentiated instruction, remediation, or enrichment. Products may possess a trademark or copyright and are generally supported by a developer who provides technical assistance and sells or distributes the intervention.
- **Policies.** Education policies involve structural changes that are intended to directly or indirectly improve student outcomes. Examples of education policies include modifying the academic calendar and changing the number of credits required for graduation.
- **Programs.** Education programs are combinations of practices, products, or policies. For example, a charter school program may combine teacher practices with policies regarding school uniforms and total days of instruction.

Where appropriate and feasible given the information provided in the study, reviewers should code the delivery method for students for each eligible intervention according to the following categorizations, which may not be mutually exclusive:

- **Individual.** The intervention is implemented one on one with an individual student.
- **Small group.** The intervention is implemented with groups of less than half of the classroom of students.
- **Whole class.** The intervention is implemented with entire classes of students.
- **School.** The intervention is implemented at the school level and includes components beyond those implemented in classrooms.

Finally, there is no minimum implementation duration requirement, nor must the intervention be “branded.”

Eligible Research

The *WWC Standards Handbook, Version 4.1* lists and describes the types of research designs that are eligible to be reviewed by the WWC. To be eligible for WWC review, studies must adhere to the following additional parameters:

- **Timeframe.** Studies must have been released or made public within the 20 years preceding the year of the review—for example, in 2001 or later for reviews being conducted in 2021.
- **Language.** The study must be available in English to be included in the review.
- **Location.** The majority of the study’s analytic sample must include students in the United States, in its territories or tribal entities, at U.S. military bases overseas, or in Organisation for Economic Cooperation and Development member countries in which English is the primary or most commonly used language – that is, Australia, Canada, Ireland, New Zealand, or the United Kingdom.

Eligible Findings

The *WWC Procedures Handbook, Version 4.1* (p. 20) distinguishes *main* findings from *supplementary* findings. Reviewers using this protocol should assess all eligible main findings in a study. Main findings include those measured for the full analytic sample using composite—not subscale—scores at the time period closest to the end of the intervention, except for outcomes in the Employment and Earnings domains and any outcome domain under Educational Attainment. For these domains, main findings under this protocol are those measured at least one year after the intervention.¹

The review of supplementary findings may depend on the purpose of the review. In general, in addition to main findings, reviewers should code the following supplementary findings if they are reported in the study:

- All eligible outcomes for the full analytic sample not represented in the main findings but consisting of findings on subscale scores or findings at additional follow-up time periods.
- Subgroup findings, measured at the conclusion of the intervention or any follow-up time period, for:
 - (i) students with disabilities or developmental delays;
 - (ii) students at risk of low performance in academics

¹ For these domains, findings that meet WWC standards at the earliest time point at least one year after the intervention has concluded should be recorded as the main findings. All other findings in these domains, including the findings near the end of the intervention and other follow-up findings, should be reviewed as supplementary.

or behavior according to a standardized baseline measure; and (iii) dual language learners, English learners, or non-native English speakers.

Supplementary findings in studies in which no main findings meet WWC standards, and analyses characterized by the study authors as “sensitivity analyses,” should not be reviewed under this protocol.

Eligible Outcomes

The Eligible Outcome Domains table (Table 1) delineates outcomes eligible for study reviews. The studies reviewed under this protocol cover a diverse set of interventions that examine a large set of outcomes. The outcome domains identified in the protocol are drawn from prior topic area protocols with minor adjustments, and include outcomes that are of interest to the education community and consumers of the WWC².

Measures of Academic Readiness, Knowledge, or Skills³

Algebra: Ability to solve, graph, or write equations, systems of equations, and inequalities, as well as functions, exponents, polynomials, factoring, and quadratic equations.

Alphabets: Letter identification, phonemic awareness, phonics, phonological awareness, and print awareness for the English language.

Calculus and Precalculus: Differential calculus—concerning rates of change and slopes of curves—or integral calculus, concerning accumulation of quantities and the areas under and between curves. Precalculus topics such as functions, complex numbers, vectors, and matrices, and trigonometry topics such as trigonometric functions, angular formulas and relationships, and the unit circle. Note that right triangle trigonometry—such as the Pythagorean theorem—would fall under the Geometry and Measurement domain.

Cognition: The process through which an individual obtains and processes knowledge or conceptual understanding, including abstract reasoning, concept formation, critical thinking, executive function skills, general problem solving, logical thinking, memory, metacognition, spatial ability, symbolic learning, and IQ.

Data Analysis, Statistics, and Probability. The act of collecting, organizing, and displaying data to answer questions, as well as statistical methods to analyze data, making inferences and predictions based on data, and probability.

Earth/Space Sciences: The structures, properties, and materials of Earth; tectonics; Earth’s place in the solar system and the universe; changes in Earth over time; water, weather, and climate; energy in Earth systems; bio-geology.

Expressive Communication: Communicating words or ideas using developmentally appropriate spoken English, assistive devices, sign language, or non-verbal cues.

Academic Achievement: Academic measures based on student assessments across multiple subjects, including at least two of literacy, mathematics, science, and social studies. For students in grades 6–12, this domain also includes standardized achievement tests such as the ACT and SAT and state-mandated tests not falling into one of the subject-specific general achievement domains, a final grade in a single course, or student grade point average (GPA). Measures of academic achievement below the course grade level, such as exam scores and quizzes, are not eligible for review.

General Literacy Achievement: Content in two or more distinct English literacy domains: Alphabets, Reading Fluency, Vocabulary, Reading Comprehension, Writing Conventions, Writing Productivity, and Writing Quality. Note,

² Where possible, reviewers should code in the notes to their reviews (a) the full name and acronym of the outcome measures, and (b) the sources of measures, such as administrative data; study-administered assessments; or surveys of students, parents, or educators. The review team may obtain this information from the text of the study—including the citations for the measures—as well as website searches, the WWC study database, or reviewer resources.

³ Where possible, reviewers should code in the notes to their reviews (a) the full name and acronym of the outcome measures, and (b) the sources of measures, such as administrative data; study-administered assessments; or surveys of students, parents, or educators. The review team may obtain this information from the text of the study—including the citations for the measures—as well as website searches, the WWC study database, or reviewer resources.

however, that outcomes limited to the Vocabulary and Reading Comprehension domains are reviewed under the domain of Reading Comprehension, and outcomes limited to multiple Writing domains are reviewed under the domain of Writing Quality.

General Mathematics Achievement: Content in two or more of the mathematics domains: Algebra; Data Analysis, Statistics, and Probability; Geometry and Measurement; Numbers and Operations; Calculus; Trigonometry/Precalculus. Also included in this domain are tests of mathematical understanding, procedures, and problem solving, as well as mathematics measures in early childhood education.

General Science Achievement: Content in two or more of the science domains: Life Sciences, Physical Sciences, or Earth/Space Sciences. Also included in this domain is general knowledge of science practices, such as forming hypotheses and making predictions, control of variables, and planning and conducting observations and experiments.

General Social Studies Achievement: Outcomes in social studies disciplines, such as one or more of civics, economics, geography, history, and world cultures.

Geometry and Measurement: Two-dimensional and three-dimensional geometric shapes and understanding properties, composition, and geometric relationships, including visualization, spatial reasoning, and geometric modeling, as well as understanding the attributes, units, systems, and processes of measurement, and applying techniques, tools, and formulas to determine measurements. Right triangle trigonometry—such as the Pythagorean theorem—also falls under this domain.

Life Sciences: The structures and functions of living things at different scales; growth, development, and reproduction of organisms; information processing and behavior in organisms; matter and energy transfer in living things and ecosystems; inheritance of and variation in traits; natural selection and adaptation; evidence of common ancestry; and biodiversity.

Numbers and Operations: Understanding numbers and integers, such as subitizing, estimation, number order, number combinations, counting, comparisons, operations, computing fluently, representing fractions and ratios, and understanding the base-ten number system and fractions.

Physical Sciences: The properties of matter and changes in matter; force, motion, and interactions of forces; energy and energy transfer and conservation; relationship between energy and forces; properties of waves; and electromagnetic radiation.

Postsecondary Academic Achievement: Mastering academic content, as measured by the final exam score in a single college-level course, GPA in college-level courses, or scores on standardized professional exams such as the GRE not falling into one of the subject-specific general achievement domains. Developmental and non-developmental college courses are eligible under this domain. Measures of academic achievement below the final exam level, such as mid-term exams and quizzes, are not eligible under this domain. Technical skills proficiency is a separate domain.

Proficiency in the English Language: Content in the Expressive Communication domain or the Receptive Communication domain combined with each other, or with one of the English literacy domains: Alphabets, Reading Fluency, Vocabulary, Reading Comprehension, Writing Conventions, Writing Productivity, and Writing Quality. Literacy measures in early childhood education also fall under this domain.

Proficiency in a Language Other Than English: Ability to speak, comprehend, read, or write in a language other than English.

Reading Comprehension: Understanding the meaning of written texts or passages in English, which may be combined with receptive or expressive vocabulary in the Vocabulary domain. This domain does not include tests of content knowledge.

Reading Fluency: Reading English words and text accurately, automatically, and with expression.
Receptive Communication The ability to follow, process, and understand spoken English, sign language, facial expressions, body language, or non-verbal cues. This domain includes measures that assess a learner's ability to respond to comprehension questions after listening to a passage that has been read to them, to provide an accurate retelling of a passage, or to follow prompts or instructions.

Technical Skills Proficiency: Technical skills at the occupation level, measured by assessments aligned with industry-recognized standards.

Technology and Engineering Literacy: The capacity to use, understand, and evaluate technology and technological principles and strategies needed to develop solutions and achieve goals. This domain includes computer science, information technology, technology and society, engineering design, maintenance and troubleshooting, computers and software, networking systems and protocols, computational thinking, and digital devices. Technical skills proficiency at the occupation level is a separate domain.

Visual and Performing Arts: Knowledge and skills in dance, music, theater, or the visual arts, such as painting, drawing, printmaking, sculpture, folk art, decorative arts, photography, video, film, computer imaging, graphic design, industrial design, and architecture.

Vocabulary: Understanding the meanings of written English words using receptive vocabulary or expressive vocabulary, whether oral or written.

Writing Conventions: Using rules of standard English language, such as word usage, syntax/sentence structure, grammar, morphology/word inflections, language mechanics/ capitalization and punctuation, handwriting quality, or spelling. When spelling skills are assessed on writing samples, they are included in this domain; otherwise, they are included in the Alphabetic domain. Measures that include writing conventions and writing productivity should be included in the General Literacy Achievement domain.

Writing Productivity: Measures of English writing quantity, such as counts of written words or sentences. Measures that include writing conventions and writing productivity should be included in the General Literacy Achievement domain.

Writing Quality: Writing effective, clear, well-organized text in English, such as narrative, informative, persuasive, or creative writing, including poetry. This domain includes measures of writing quality combined with measures in either the Writing Conventions or Writing Productivity domains.

Educational Attainment

College Degree Attainment: Completion of a postsecondary degree, certificate, or program. Industry-recognized certificates are in a separate domain.

High School Completion: Whether the student has earned a high school diploma or a recognized equivalent, such as obtaining passing scores on a state-approved high school equivalency test or passing the GED.

Industry-Recognized Credential, Certificate, or License: Examples of ways completion of an industry program might be operationally defined include certificate completion rates, non-degree award receipt rates, and certifications from third-party licensing or credentialing bodies.

Educational Progress⁴

College Enrollment: Applying to, enrolling in, or attending a postsecondary institution, such as actual enrollment in college, number or selectivity of admitted institutions, enrollment by institution type, full-time vs. part-time enrollment, and immediate vs. delayed enrollment.

School Attendance and Progress

College Readiness: Student progress on preparedness to enter postsecondary education, such as meeting specific

⁴ For these domains, “postsecondary institution” includes universities, colleges, community colleges, and trade schools, and “secondary school” pertains to students in grades 6–12.

minimum coursework requirements for entry into postsecondary institutions, completing other required key milestones, such as applying for college or for student financial aid, and completing college requirements in high school, such as dual enrollment or Advanced Placement (AP) courses.

Progressing in College: Progress toward the completion of a postsecondary degree, certificate, or program, such as number of college-level credits earned, number of terms of continuous enrollment, and enrolled vs. did not enroll the next semester. Note that college credits completed during high school and non-college level credits, such as developmental credits, are not eligible under this domain.

Progressing in Developmental College Courses: Completed required developmental coursework, grades earned in developmental courses, or completed vs. did not complete the first college-level course in which remediation was needed.

Progressing in School (Secondary School): Number of Carnegie units or credits the student has earned, whether the student was promoted to the next grade, or highest grade completed.

School Attendance: Elementary or secondary school attendance rates, absenteeism, or tardiness at school, such as the number or proportion of days in attendance, measures of excessive or chronic absenteeism, or referrals for truancy.

Staying in Secondary School: Whether the student has dropped out of secondary school or is still enrolled.

Social-Emotional Outcomes, Behaviors, and Skills

Functional Skills: Skills needed to participate in developmentally appropriate routines and activities in the home or in community settings and support independent living. This could include outcomes such as demonstrating gross or fine motor skills, dressing, preparing and eating food, hygiene, cleaning, organizing, crossing the street, using a bank, purchasing items, or using various forms of transportation. Measures may be based on a self-report, observation, or results from an assessment scale. Academic and employability skills are in separate domains. This domain is most relevant for interventions in special education, early childhood education, or adult education.

Intrapersonal Competencies: Mental health indicators that are primarily focused inward and reflect a student's emotional status and psychological well-being, and that include internalizing behaviors and both negative and positive feelings. Outcomes in this domain include: anxiety, depression, loneliness, and thought disorders, as well as emotional regulation, happiness, motivation, self-concept, self-efficacy, self-regulation, and overall adjustment. Measures may be based on a self-report, educator observation, or results from an assessment scale.

Student Behavior: Observable behaviors that conform or fail to conform to developmentally appropriate behavioral norms, rules, or expectations within school or in the community. Examples of positive and prosocial outcomes in this domain include: participating in class or in extracurricular activities, paying attention, respecting others, staying on task, and other measures of interpersonal engagement, social functioning, or time management. Negative outcomes and externalizing behaviors in this domain include: disruptive or impulsive behaviors such as interrupting others, peer rejection, and physical or verbal aggression; and problem behaviors such as cheating, lying, stealing, substance abuse, or vandalism. Measures reflect behaviors rather than feelings and may be based on a self-report, observation, or results from an assessment scale. School Attendance and Student Discipline are separate domains, and School Climate focuses on the schoolwide environment rather than a student's own behavior.

Student Discipline: Documented consequences such as arrests, suspensions, number of office referrals, or expulsion from school that may result from student behaviors or other factors. This domain includes disciplinary incidents at the student, teacher or classroom, and school levels. Disproportionate disciplinary incidents in schools and across different student subgroups is included in the School Equity domain.

School Environment Outcomes

School Climate: Observations or assessments of the schoolwide environment or culture, as distinct from one's own behavior, such as the quality of interpersonal relationships within the school, the sense of belonging or safety, engagement in school, and the prevalence of bullying at the school. Measures may be based on extant data, self-reports, observations, or results from an assessment scale.

School Equity: The extent of equity within classrooms or schools, such as perceptions of equitable treatment of students by school staff, or disproportionality in disciplinary actions or in identification for special education or in access to advanced courses. Measures may be based on extant data, self-reports, observations, or results from an assessment scale. School Equity is distinguished from School Climate in that it focuses primarily on the disparate impact of school policies and practices. Number of disciplinary incidents – without regard for their disproportionately – would fall under the Student Discipline domain.

Labor Market Outcomes

Earnings: Income received from work. Earnings must be defined for those not employed as well those employed. Reviewers should also denote the time periods when the outcomes were collected in the notes.

Employment: Indicator of any paid employment, number of months or quarters employed, or number of hours worked in an average week. Reviewers should also denote the time periods when the outcomes were collected in the notes.

General Employability Skills: Skills that employers value in all employees, regardless of field, such as critical thinking, ability to work in teams, communication with coworkers or supervisors, self-management and initiative, and problem solving in work settings. Resume development, completing job applications, and interviewing skills are also included in this domain. Technical Skills Proficiency at the occupation level is a separate domain.

Teacher Outcomes

Teacher Attendance: Includes outcomes that indicate the number or percentage of eligible work days for which the teacher is present.

Teacher Practice: Quality of instruction provided by teachers and their application of content knowledge or pedagogical knowledge as demonstrated by their actions in the classroom, including attempts to promote a positive classroom environment and influence problematic student behavior by responding to students' actions with consequences or rewards. Outcomes can be based on classroom observation rubrics assessed by school principals, supervisors, or trained evaluators or student surveys.

Teacher Retention in the Profession: Includes outcomes that measure the percentage of teachers who return to work as a teacher from year to year, regardless of location.

Teacher Retention in the School: Includes outcomes that measure the percentage of teachers who return to work as a teacher in the same school from year to year.

Teacher Retention in the School District: Includes outcomes that measure the percentage of teachers who return to work as a teacher in the same school district from year to year.

Teacher Retention in the State: Includes outcomes that measure the percentage of teachers who return to work as a teacher in the same state from year to year, or who return to teach in certain kinds of settings, such as in special education or in economically disadvantaged districts, within the state.

School Leader Outcomes

School Leadership Practice: Includes outcomes that measure the quality of leadership ability demonstrated by the school leader, as measured by rubrics assessed by supervisors or from surveys of school staff, parents, or students.

School Leader Retention in the Profession: Includes outcomes that measure the percentage of school leaders who return to work as a school leader from year to year, regardless of location.

School Leader Retention in the School: Includes outcomes that measure the percentage of school leaders who return to work as a school leader in the same school from year to year.

School Leader Retention in the School District: Includes outcomes that measure the percentage of school leaders

who return to work as a school leader in the same school district from year to year.

School Leader Retention in the State: Includes outcomes that measure the percentage of school leaders who return to work as a school leader in the same state from year to year, or who return to lead schools in certain kinds of settings, such as in economically disadvantaged districts within the same state.

IMPLEMENTATION OF THE WWC STANDARDS

The *WWC Standards Handbook, Version 4.1* and the *WWC Procedures Handbook, Version 4.1* allow review teams some specific types of discretion in completing study reviews. These areas of discretion are available to review teams unless specifically disallowed below.

Outcome Measure Standards

This protocol uses the outcome measurement standards as outlined in the *WWC Standards Handbook, Version 4.1* (pp. 83–85). Outcome measures must demonstrate face validity and meet minimum score reliability requirements, must not be overlapped with the intervention, and should be collected in the same manner for both intervention and comparison groups. This protocol also requires a more stringent internal reliability criterion of 0.60 for measures. Reliability information is not required for: standardized tests that have established administration and scoring procedures often documented in technical manuals; outcomes using administrative data such as graduation from high school, school enrollment, and grade retention; or grade-point average, if measured from administrative data or calculated using a specified formula.

Reliability for Outcome Measures When Reliability Is Unreported

Reviewers may encounter outcome measures that fall within an eligible domain but do not have reliability information. The *WWC Standards Handbook, Version 4.1* (p. 84) suggests that some outcome measures might be deemed reliable even if reliability information is unreported. Review teams may determine that some measures may be acceptable if they can be scored by a single coder with low error, such as a multiple-choice test or counts of words spelled correctly. Reviewers should consult with review team leadership if they encounter a study with an outcome that might fall into this category. Any judgments made should be documented in the review.

Attrition Threshold

This protocol uses the optimistic boundary for attrition as a default. This boundary was selected because it was the most common boundary used across prior WWC review protocols, suggesting that in typical education studies, attrition is not thought to be strongly related to intervention status. Review team leadership, in consultation with a content expert, may choose the cautious boundary if it is justified for a particular study review, as in the case of a study with suspected endogenous attrition. Examples where the cautious boundary should be used include dropout prevention or school choice programs, programs delivered outside of regular school hours, elective or selective courses such as Advanced Placement courses, or other interventions intended to influence school retention. If the cautious boundary is used, a rationale should be documented in the study review guide.

Assessing Baseline Equivalence

For student outcomes, baseline equivalence of the intervention and comparison groups in the analytic sample may be demonstrated either on a pretest in the same domain as the outcome, or on a pretest in a broader domain in the same content area and for which the outcome domain is a subset. For example, for an outcome measure in the Algebra domain, baseline equivalence could be established using a measure of General Mathematics Achievement, but for an outcome measure in the General Mathematics Achievement domain, baseline equivalence could not be established using only a measure in the Algebra domain.

If an appropriate pretest in the same domain as the outcome—or in a broader domain in the same content area encompassing the outcome—was not given or is not available for certain outcomes, then baseline equivalence of the intervention and comparison groups in the analytic sample should be demonstrated on a broad measure of student academic readiness, knowledge, or skills within one of the following domains: Cognition, Academic Achievement,

General Literacy Achievement, General Mathematics Achievement, or Postsecondary Academic Achievement.⁵ Note that broad measures within the Academic Achievement domain include both standardized tests and student GPA in grades 6–12, and broad measures within the Postsecondary Academic Achievement domain include both standardized tests and student GPA in college-level courses.

Additional considerations for assessing baseline equivalence are:

- For studies of interventions prior to kindergarten and of adult education interventions, equivalence also needs to be demonstrated for the age of the learners in the intervention and comparison groups.
- For teacher and school leader outcomes, baseline equivalence should be established with acceptable preintervention measures:
 - For the Teacher Practice, Teacher Attendance, and School Leadership Practice outcome domains, the acceptable baseline measures are the same measure as the outcome or another measure from the same domain as the outcome.
 - For the teacher or school leader retention in the school, district, state, or profession outcome domains, the acceptable baseline measures are the average years of teaching or school leader experience or the experience categories used in the study, AND, a broad measure of student academic readiness, knowledge, or skills, as defined above, **AND, for teacher retention if the unit of assignment is the school**, a school-level measure of the teacher retention outcome.

Baseline Equivalence for Cluster Design Studies

If the baseline equivalence of individuals in the intervention and comparison groups cannot be established because individuals with outcomes data are not the same as those with baseline data, the following two types of samples can be used to satisfy the baseline equivalence requirement for the analytic sample of intervention and comparison clusters, provided the review team leadership determines that the baseline data are representative of the individuals who were in the clusters at the time the follow-up data were collected, consistent with the *WWC Standards Handbook, Version 4.1* (pp. 26–27):

- Individuals from the same cohort, within the same clusters as the individuals in the analytic sample.
- Individuals from the previous cohort, within the same grades and clusters as individuals in the analytic sample.

If baseline data are provided at multiple time points, reviewers should report data that are from individuals from the same cohort, within the same clusters as the individuals in the analytic sample.

Adjusting for Baseline Differences

If a pretest is available for an outcome and the difference between conditions is shown to be within the range that both allows and requires statistical adjustment, an adjustment on an acceptable pretest is needed in order for the study to be rated *Meets WWC Standards With Reservations*. Acceptable pretests are specified in the section above, *Assessing Baseline Equivalence*.

Applying a Difference-in-Differences Adjustment

The WWC will apply a difference-in-differences adjustment to all eligible main and supplementary findings, as applicable, regardless of whether the study must satisfy the baseline equivalence requirement, provided the required information is obtained from the study or the study authors. If this adjustment is needed to satisfy equivalence for baseline differences between 0.05 to 0.25 standard deviations, then the baseline and outcome measures must be the same and have a correlation of 0.60 or higher. If the sample correlation is not available, review teams can impute an empirically-based correlation but must provide empirical evidence of the assumed correlation in the study review guide.

Adjusting Multiple Outcomes in the Same Domain

This section provides guidance for reviewing studies that report multiple baseline measures for the same outcome domain and for the same analytic sample. For main findings, when the review team considers the baseline measures

⁵ Baseline measures in the Proficiency in the English Language domain are acceptable for early childhood education studies.

highly correlated, then a baseline difference on any measure in the domain that is larger than 0.25 standard deviations means that all measures in the outcome domain for the same analytic sample do not meet WWC standards, as specified in the *WWC Standards Handbook, Version 4.1* (p. 17). Also in this case, when a baseline difference on any measure in the domain is within the range requiring a statistical adjustment, the analyses for *all* findings in the domain for the same analytic sample must adjust for the baseline measure that is in the range requiring a statistical adjustment, as specified in the *WWC Standards Handbook, Version 4.1* (p. 17). For supplementary findings and measures in a broad outcome domain such as Student Behavior that are not considered highly correlated by the review team, review teams may assess baseline equivalence and require statistical adjustments on an outcome-by-outcome basis as needed for the review. Baseline measures affect baseline equivalence and required statistical adjustments only for findings using the same analytic sample.

Additional Considerations for Studies With Cluster-Level Assignment

This section provides additional guidance for cluster-level assignment studies, including: the clustering adjustment in mismatched analyses; the general risk of bias posed by joiners, and how this definition differs at different levels of assignment; and the definition of cluster representativeness.

Clustering Adjustment in Mismatched Analyses

When there is a mismatch in the level of assignment and the level of analysis, the WWC will conduct an adjustment of the significance level to account for this mismatch if not accounted for in the study. This computation requires an estimate of the proportion of variance in the outcome at the level of analysis that is explained by the assignment level, an estimate known as the intraclass correlation (ICC). Study-reported ICC values will be used when available; otherwise, 0.20 will be used for achievement outcomes and 0.10 will be used for all other outcomes, as specified in the *WWC Procedures Handbook, Version 4.1* (p. 20).

Risk of Bias From Joiners

The WWC defines a *joiner* as any individual who enters a cluster after the results of random assignment are known to any individual who could plausibly influence individuals' placement into a cluster, such as decisions by parents, students, teachers, or principals. Reviewers should assume *any* joiners in the analytic sample pose a risk of bias if the unit of assignment is *smaller* than a school—such as teachers, classrooms, or small groups of students within classrooms—or if the intervention is a dropout prevention or school choice program, a program delivered outside of regular school hours, an elective or selective course such as an Advanced Placement course, or any other intervention intended to influence school retention. Otherwise, unless the review team documents concerns about the intervention influencing the composition of the analytic sample, reviewers should assume *no* joiners pose a risk of bias to the corresponding finding.

Cluster Representativeness

If a study has cluster-level assignment and uses administrative data, then its analytic sample is deemed representative of its clusters unless review team leadership determines that relevant individuals are excluded in a systematic way. If the latter, reviewers should implement the approach outlined in the *WWC Standards Handbook, Version 4.1* (pp. 26–27) and should calculate the individual-level nonresponse using the full population of individuals in clusters at follow-up as the denominator and the analytic sample as the numerator. The boundary for this calculation, whether cautious or optimistic, should be congruent with the boundary selected to assess attrition.

Additional Considerations for Single-Case Design Studies

This section provides additional guidance for single-case design (SCD) studies, including: SCD studies with fewer than three data points in a phase, and reviews of multiple probe designs.

SCD Studies With Fewer Than Three Data Points per Phase

As noted in the *WWC Standards Handbook, Version 4.1* (p. 80, footnote 23), there may be occasions when fewer than three data points in a phase would not automatically require the study to be rated as Does Not Meet WWC SCD Standards. In these circumstances, content experts will consider the intervention and outcomes to determine whether an exception should be made. Possible exceptions include:

- Interventions for severe problem behavior, such as aggression and self-injury, in which extended initial

baseline or reversal conditions pose serious ethical and procedural concerns.

- Studies that include “zero baseline” behaviors or skills for which there is no logical reason to conceive that further assessment would yield other than zero baseline performance—for example, when a child is asked to provide a verbal label for an object and consistently provides no response to the request because the child has little to no language skills.

Multiple Probe Designs

The WWC Standards Handbook, Version 4.1 (p. 81) includes three criteria that multiple probe designs must meet in order to be rated Meets WWC SCD Standards With Reservations or Meets WWC SCD Standards Without Reservations. These criteria are required because some baseline data points are intentionally missing. The review of SCD studies under this review protocol allows for an exception to the third multiple probe criterion, as described below.

- The third multiple probe criterion requires that “each case not receiving the intervention must have a probe point in a session where another case either (a) first receives the intervention or (b) reaches the prespecified intervention criterion.” However, studies reviewed under this topic area can be rated *Meets WWC SCD Standards With Reservations*, even if they do not meet this requirement.
- Cases must still continue to have baseline data for at least one session after the intervention is administered to preceding cases—as this is a requirement for all multiple baseline designs—and must meet the other two multiple probe criteria specified in the *WWC Standards Handbook, Version 4.1*. Failure to meet these criteria will result in a study rating of *Does Not Meet WWC SCD Standards*.