

Study Review Protocol and Supplement to the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* Version 5.1 (November 2025)

A Publication of the National Center for Education Evaluation at IES

The What Works Clearinghouse (WWC) Study Review Protocol accompanies the [WWC Procedures and Standards Handbook, Version 5.0](#) and guides reviews of studies by the WWC. The WWC uses this protocol to review all studies, including those cited as evidence for U.S. Department of Education grant competitions, studies that were funded by the Department, and studies identified for systematic reviews of evidence based on a search of the research literature in a particular topic area. When the Study Review Protocol is used to review studies for systematic reviews, an accompanying [topic area synthesis protocol](#) will provide criteria for the literature search; guidance on how to identify and prioritize relevant studies for review and inclusion in evidence synthesis products; and guidance on intervention, sample, and outcome eligibility criteria for the synthesis (see [page 10](#) in the *Handbook*).

This protocol updates the December 2024 [Study Review Protocol, Version 5.1](#). It describes a pilot process that integrates artificial intelligence (AI) tools, such as large language models, into the WWC review process. This version of the protocol retains clarifications to outcome domains and formalized guidance and clarifications to the *Handbook* provided with the December 2024 *Study Review Protocol, Version 5.1*. No other substantive changes have been made. Review teams should use this *Study Review Protocol, Version 5.1*, applying either the pilot or the standard review process, for all new reviews conducted under Version 5.0 standards.

Contents

ELIGIBLE OUTCOME DOMAINS.....	2
Academic Readiness, Knowledge, or Skills (Preschool through Postsecondary)	3
General Achievement	3
Language Arts	4
Mathematics	5
Science	6
Other	6
Preschool through Grade 12 Progress and Completion	7
Postsecondary Progress and Success	7
Workforce Outcomes	8
Social, Emotional, Behavioral, and Mental Health Outcomes (Preschool through Postsecondary)	9
School and Family Outcomes and Educational Opportunity (Preschool through Postsecondary)	10
School Leader Outcomes	12
Teacher Outcomes	12
Other	13
MAIN VERSUS SUPPLEMENTAL FINDINGS ...	13
BASELINE EQUIVALENCE.....	16
OTHER.....	17

WWC products may synthesize findings from studies reviewed under the Version 5.0 and Version 5.1 *Study Review Protocols* if the outcome domains are substantively the same across the versions. A crosswalk of outcome domains can be found on the [WWC website](#). Studies that were previously reviewed under *Study Review Protocol, Version 5.0* should not be re-reviewed under the *Study Review Protocol, Version 5.1* unless there is a specific justification made by review team leadership and approved by IES. If a study was previously reviewed under the Version 4.1 (*Study Review Protocol, Version 1.0 or Version 4.1*) or Version 5.0 standards and needs to be reviewed under the *Study Review Protocol, Version 5.1*, a simplified review process can occur where only one reviewer and one reconciler review the study.

The *Handbook* specifies the [eligibility criteria](#) for studies of educational interventions to be reviewed by the WWC. The criteria include eligible [research reports](#), eligible [research designs](#), eligible [populations](#), eligible [interventions](#), and eligible [outcomes](#). Findings that are not publicly available and/or are not included in a complete manuscript are not eligible for review. Author queries may obtain other information that is necessary for a review but is not publicly available, including unadjusted means.

ELIGIBLE OUTCOME DOMAINS

The list of outcome domains below defines the domains that are eligible for review by the WWC and describes the types of outcome measures that would fall within each domain. The list of outcome domains is organized by topical categories and alphabetized within categories.

Several considerations apply across multiple outcome domains:

- **Measure independence.** Outcome measure independence is assessed for outcomes within domains marked with an asterisk (*) that relate to literacy, mathematics, or broad measures of student achievement. The [WWC website](#) lists the measures that the WWC considers independent. Independent measures are reviewed on a quarterly basis based on nominations sent to the [WWC Help Desk](#).
- **Course performance versus course credits.** Course grades—including pass/fail—and grade point averages (GPAs) are always reviewed under the Course Performance domains. The number of credits earned are reviewed in Progressing in PK-12 Education, Progressing in Developmental Education, or Progressing in College domains.
- **Course performance outcomes when intervention targets course type.** For interventions that aim to increase the number of students taking specific courses, such as an intervention that aims to increase the number or rigor of STEM courses taken, course performance outcomes for the set of courses do not meet the WWC's requirement of consistent data collection for the intervention and comparison conditions (p. 28 of the *Handbook*). These outcomes would not meet WWC standards.
- **Measures contributing to course performance.** For any outcome domain, scores below a course grade, such as quiz and exam scores, are not eligible for review because the content or scoring might differ across educators or classes. Measures below a course grade are eligible for review when all students take the same assessment, such as measures created by researchers or

states/districts (but not schools), such as a district end-of-unit exam. Assessments that individual schools or teachers create are not eligible for review.

- **Single item and other very narrow measures.** Single item measures and other very narrow measures of constructs would generally not meet the face validity requirement of the WWC outcome measure standards. As indicated in the *Handbook* (p. 14), measures must appear to measure what they claim. Review team leadership may determine that a single item or narrow measure has face validity if it, individually, adequately represents the breadth of the construct.
- **Other face validity considerations.** Some domains note that measures should reflect practically meaningful benchmarks within the study context. This principle generally applies across all domains. If a measure is based on arbitrary thresholds without practical meaning, the measure does not meet the face validity requirement of the WWC outcome measure standards.

Academic Readiness, Knowledge, or Skills (Preschool through Postsecondary)

General Achievement

Academic Achievement—PK-12*: Academic measures in preschool through grade 12 based on broad standardized student assessments across multiple subjects, including at least two of literacy, mathematics, science, and social studies. This domain includes standardized achievement tests for students in preschool through grade 12 not falling into one of the subject-specific achievement domains, such as the ACT, SAT, and state-mandated tests.

Academic Achievement—Postsecondary*: Academic measures based on broad standardized student assessments across multiple subjects, including at least two of literacy, mathematics, science, and social studies. This domain includes standardized achievement tests for postsecondary students, such as the LSAT, MCAT, GRE, and GMAT.

Cognition: Measures of skills and abilities students use to obtain and process knowledge or conceptual understanding, including abstract reasoning, concept formation, critical thinking, flexible thinking, organization, planning and prioritization, general problem solving, logical thinking, memory, metacognition, spatial ability, symbolic learning, and IQ. Specific subdomains of executive functioning fall within the Cognition, Student Behavior, and Mental Health domains. These outcomes should be reviewed under the category best reflected by the subdomain(s) of executive function being measured in the study.

Course Performance—Elementary: Course grades and GPAs for students in grades K-5. Grades may be on a pass versus fail, satisfactory versus unsatisfactory, or other scale not equivalent to a numerical or alphabetic grade. Studies that use course performance across multiple grade levels may be included in this domain if grades are calculated according to the same scale. As indicated in the *Handbook* (p. 27), when different schools or districts are included in the study, the same scale of GPA must be used across sites.

Course Performance—Postsecondary: Course grades—including pass or fail—and GPAs from one or more postsecondary courses. Performance in developmental and nondevelopmental college courses are eligible under this domain. As indicated in the *Handbook* (p. 27), when different institutions are included in the study, the same scale of GPA must be used across sites.

Course Performance—Secondary: Course grades—including pass or fail—and GPAs for students in grades 6-12. GPAs may include grades for dual enrollment courses. As indicated in the *Handbook* (p. 27), when different schools or districts are included in the study, the same scale of GPA must be used across sites.

Language Arts

The Expressive Communication and Receptive Communication include outcomes in both the English language and student home language. Other literacy domains include only outcomes in English except for outcomes that fall under the Proficiency in a Language Other than English domain.

Expressive Communication*: Communicating words or ideas using developmentally appropriate spoken language; assistive devices, including picture-based alternatives; sign language and communication gestures; or nonverbal cues. This includes the length and complexity of communication and rates of communication. Communication can be in any language.

Literacy Achievement*: Content in two or more distinct literacy domains: Phonics and Alphabetics, Reading Comprehension, Reading Fluency, Vocabulary, Writing Conventions, and Writing Quality. Outcomes limited to the Vocabulary and Reading Comprehension domains are reviewed under Reading Comprehension; outcomes limited to multiple writing domains are reviewed under Writing Quality; and outcomes including either Expressive or Receptive Communication domains are reviewed under Proficiency in the English Language.

Literary Analysis: Analysis of a literary work—including prose, poetry, and drama—from various periods and countries to understand the author’s meaning; analyze a work’s structures, style, and themes; understand the use of symbolism, figurative language, imagery, and tone; and write expository, analytical, or argumentative essays to analyze or interpret the text.

Phonics*: Letter identification, phonological awareness (including phonemic awareness), encoding/spelling, decoding, and print awareness.

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 literacy domains: Phonics and Alphabetics, Reading Comprehension, Reading Fluency, Vocabulary, Writing Conventions, and Writing Quality. Measures of acquiring the English language for both English learners and native English speakers may also fall under this domain in early childhood education.

Proficiency in a Language Other Than English*: Ability to speak, comprehend, read, or write in a language other than English. Outcomes that focus only on Receptive Communication or Expressive Communication should be reviewed under those domains. Outcomes that include components of both

Receptive Communication and Expressive Communication in a language other than English should be reviewed under this domain.

Reading Comprehension*: Understanding the meaning of written texts or passages. When measures include outcomes in this domain and the Vocabulary domain, review the measures under this domain. This domain does not include tests of content knowledge.

Reading Fluency*: Reading words and text accurately, automatically, and with expression. This domain includes measures of fluency of single words or word lists.

Receptive Communication*: The ability to follow, process, and understand spoken language, sign language, facial expressions, body language, or nonverbal cues. This domain includes measures that assess a learner's ability to demonstrate comprehension after listening to a passage, instructions, or other spoken language. Communication can be in any language.

Vocabulary*: Understanding the meanings of words or pictures using receptive or expressive vocabulary, whether oral, written, or nonverbal (i.e., using a communication device, pictures, sign language, and gesturing).

Writing Conventions: Using rules of standard language, such as word usage, syntax/sentence structure, grammar, morphology/word inflections, language mechanics/capitalization and punctuation, handwriting quality, and spelling. When spelling skills are assessed on writing samples, they are included in this domain.

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

Mathematics

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

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 include functions, complex numbers, vectors, and matrices. Trigonometry topics are also eligible under this domain, including 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.

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

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.

Mathematics Achievement*: Content in two or more of the mathematics domains. Also included in this domain are tests of mathematical understanding, procedures, and problem solving. General mathematics measures in early childhood education are eligible in this domain.

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

Science

Life Sciences (Biology, Environmental, and Health 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; biodiversity; how the physical environment and human activities affect living things; and the anatomy, physiology, nutrition, and health of animals and humans.

Physical Sciences (Astronomy, Chemistry, Earth and Space Sciences, Geology, and Physics): 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; electromagnetic radiation; 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; and paleontology.

Science Achievement: Content in two or more of the science domains. Also included in this domain is general knowledge of science concepts, such as forming hypotheses and making predictions, controlling variables, participating in scientific argumentation, and planning and conducting observations and experiments.

Other

Social Sciences: Outcomes in social science disciplines, such as one or more of history, social studies, anthropology, civics, economics, geography, history, psychology, sociology, and world cultures.

Technology and Engineering: 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.

Visual and Performing Arts: Knowledge, skills, and creativity 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.

Preschool through Grade 12 Progress and Completion

For all domains under this category, measures should reflect practically meaningful benchmarks within the study context. If benchmarks are arbitrary, the measure does not meet the face validity requirement of the WWC's outcome measure standards.

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 or HiSET exam.

Progressing in Grades 1-12 Education: Number of credits the student has earned (note: credits attempted is not an eligible outcome), earning credit for a particular required course or courses, whether the student was promoted to the next grade, or highest grade completed. This domain includes preparedness for postsecondary education including whether the student has met specific minimum coursework requirements for entry and whether the student has met the requirements to enroll in credit-bearing (nondevelopmental) college courses. This domain also includes outcomes related to receiving a satisfactory score to earn college credit (e.g., a score of 3 or higher on an Advanced Placement exam, passing a dual enrollment course). When measures include components of the Progressing in Grades 1-12 Education and Progressing in Preschool-K Education domains, review the measure under this domain.

Progressing in Preschool-K Education: Participation or completion of a preschool, prekindergarten, or kindergarten program.

School Attendance: Attendance outcomes including attendance rates, absenteeism, tardiness at school, and truancy. Examples include the number or proportion of days in attendance and measures of excessive or chronic absenteeism. Measures can use administrative data or well-established surveys.

Staying in School: Whether the student has dropped out of preschool through grade 12 education or remained enrolled.

Postsecondary Progress and Success

For all domains under this category, measures should reflect practically meaningful benchmarks within the study context. If benchmarks are arbitrary, the measure does not meet the face validity requirement of the WWC's outcome measure standards.

Applying for College: Applying to a postsecondary institution, including the number or selectivity of admitted institutions. When measures include components of Applying for College and Applying for Financial Aid domains, review the measures under this domain.

Applying for Financial Aid: Measures related to applying for financial aid, including completing the [Free Application for Federal Student Aid](#) (FAFSA), applying for scholarships, and obtaining work study positions. When measures include components of Applying for College and Applying for Financial Aid domains, review the measures under the Applying for College domain.

College Degree Attainment: Completion of a postsecondary degree, certificate, or program. See the related Industry-Recognized Credential, Certificate, or License domain in the Workforce Outcomes category.

College Enrollment: First-time enrollment in a postsecondary institution, such as enrollment in college, enrollment by institution type, full-time versus part-time enrollment, and immediate versus delayed enrollment.

Progressing in College: Progress toward the completion of a postsecondary degree, certificate, or program, such as number of college-level credits earned (note: credits attempted is not an eligible outcome), number of terms of continuous enrollment, re-enrollment, and postsecondary retention. This domain includes remaining in good academic standing, transferring into a graduate program prior to finishing a bachelor's degree, and transfer to a 4-year institution. Measures cannot be based on intentions (e.g., intentions to enroll). Non-college-level credits, such as developmental credits, are not eligible under this domain, unless the developmental credits are required to complete a degree.

Progressing in Developmental Education: Completing required developmental education coursework or completing the first college-level course in which remediation was needed.

Time to Completion: The amount of time to complete a degree or credential and on-time completion. When outcomes overlap with outcomes in the Progressing in College domain, use Progressing in College.

Workforce Outcomes

Earnings: Income received from work. Earnings must be defined for those not employed as well those employed. Reviewers should document the specific time periods when the outcomes were collected. Measures can use administrative data or well-established surveys.

Employability Skills: Skills that employers value in all employees, regardless of field, such as the 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, participating in unpaid internships, and interviewing skills are also included in this domain.

Employment: Indicator of any paid employment, including paid internships; number of months, quarters, or years employed; or number of hours worked in an average week. Employment must be

defined for those not employed as well those employed. Reviewers should document the specific time periods when the outcomes were collected. Measures can use administrative data or well-established surveys.

Industry-Recognized Credential, Certificate, or License: Industry-recognized certificates are sometimes completed outside of a college or university setting and are widely recognized in the field. 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.

Technical Skills Proficiency: Technical skills at the occupation level, measured by standardized assessments aligned with industry-recognized standards, such as the NCLEX, NCEES Principles and Practice of Engineering (PE) exam, PRAXIS, and National Counselor Examination. Measures must be recognized by the industry to be eligible for review in this domain.

Social, Emotional, Behavioral, and Mental Health Outcomes (Preschool through Postsecondary)

Specific subdomains of executive functioning fall within the Cognition, Student Behavior, and Mental Health domains. These outcomes should be reviewed under the category best reflected by the subdomain of executive function being measured in the study.

Academic Dispositions: Indicators that are focused on student attitudes towards academics or attitudes towards participation in school activities (rather than observed behaviors). Outcomes in this domain include academic growth mindset, academic motivation, academic or subject-specific self-efficacy, self-efficacy to perform or achieve, academic engagement, academic interest, and academic grit. Measures may be based on self-report or results from a clinical scale. Measures are included in this domain if they reflect attitudes towards learning, as opposed to observable behaviors (Student Behavior), mental well-being (Mental Health), or schoolwide environment (School Climate).

Mental Health: Indicators of attributes or characteristics that cannot be observed directly and reflect a student's emotional status and psychological well-being, and both positive and negative thoughts and feelings not tied specifically to academics. Outcomes in this domain include constructs such as anxiety, depression, and loneliness, as well as emotional regulation, happiness, self-esteem, positive identity development, and overall adjustment. Measures may be based on a self-report or results from a clinical scale. Measures of current mood are not eligible under this domain. When measures include components of Mental Health and Academic Dispositions domains, review the measures under this domain.

Prosocial Behavior: Observable outcomes such as developing age-appropriate social skills, social and emotional competencies, friendship development, and measures of peer acceptance and rejection. This domain also includes initiating or responding to joint attention, joint engagement, play initiations, and peer imitations which involve aspects of social communication. This domain also includes respecting and empathizing with others. Measures may be based on a self-report, observation, or

results from a clinical scale and should be based on behaviors. When measures include outcomes in this domain and the Student Behavior domain, review the measures under the Student Behavior domain. When measures of Prosocial Behavior and Expressive or Receptive Communication overlap, review under Prosocial Behavior.

Problem Behavior: Measures capturing reductions in problem behaviors such as skipping school, violence, fighting, lying, stealing, elopement, running away from home or school (separate from elopement), bullying, cheating, or vandalism. Measures may be based on a self-report, observation, or results from a clinical scale and should be based on behaviors instead of thoughts. This domain focuses on more serious behaviors. This differs from Student Behavior, which focuses on less serious behaviors that do not usually violate school rules, and Student Discipline, which focuses on official school records for violations. When measures include components of Student Behavior and Problem Behavior, review the measures under this domain.

Student Behavior: Observable behaviors that conform or fail to conform to developmentally appropriate behavioral norms, rules, or expectations. The types of behaviors in this domain are those that are expected to be observed in an educational setting. Examples of positive outcomes in this domain include paying attention, self-regulation, impulse control, time on task, and independent play. Examples of negative outcomes in this domain include behavioral inhibition, disruptive or impulsive behaviors, interrupting others, teasing, and elopement. Measures may be based on a self-report, observation, or results from a clinical scale but are based on behaviors instead of thoughts. When outcomes include components of Student Behavior and Prosocial Behavior, Mental Health, or Academic Dispositions, review the outcomes under Student Behavior. When outcomes include components of Student Behavior and Problem Behavior, review the outcomes under Problem Behavior.

Student Discipline: Documented school records of arrests, suspensions, number of office referrals, or expulsion from school that may result from student behaviors or other factors. This domain includes disciplinary incidents and referrals to the judicial system, except for truancy. When outcomes include components of Student Behavior and Student Discipline domains, review the outcomes under this domain.

School and Family Outcomes and Educational Opportunity (Preschool through Postsecondary)

Access to Educational Opportunities: The numbers or percentages of students who have access to educational opportunities from preschool to postsecondary that can influence whether a student remains on track for college or career readiness where access means the opportunity is made available to students. This domain includes access to career development courses and access to courses needed for postsecondary success, including STEM courses, Algebra 1, advanced courses, Advanced Placement courses, International Baccalaureate courses, and dual enrollment courses. This domain also includes access to universal preschool; mental health supports; gifted and talented programs; qualified, experienced, demographically representative or diverse, and/or effective teachers; and resources

needed for learning, such as Wi-Fi. Participation in these opportunities should be reviewed under the Participation in Educational Opportunities domain, and successful completion should be reviewed under one of the Progressing in Education domains.

Accountability Metrics: Outcomes that relate to improvements in school accountability metrics and that include measures with an academic component, such as proficiency rates or school accountability ratings. Student-level proficiency rates should also be reviewed under this domain, and student scale scores aggregated to the school level should be reviewed under other domains.

Compositional Change: The effect of the intervention on the demographics of the grade, school, or district. This domain includes changes to the racial or ethnic composition of students, proportion of students with disabilities, proportion of English learners, proportion of students with low socio-economic status, and overall number of students. Outcomes in this domain can be reviewed as supplemental findings only.

Family Engagement: Actions and attitudes reflecting the extent to which families are engaged in supporting student success in school. Examples include families participating in school events and family perceptions about their engagement, as opposed to school efforts to engage families, which would fall under the School Climate domain. Measures of teacher-initiated communication with families should be reviewed in the School Climate domain. If measures include outcomes that fall under this domain and the School Climate domain, review under the School Climate domain.

Participation in Educational Opportunities: The numbers or percentages of students who participated in educational opportunities from preschool to postsecondary that can influence whether a student remains on track for college and career readiness, including participation in career development courses and enrollment in STEM courses, Algebra 1, advanced courses, dual enrollment courses, universal preschool, and gifted and talented programs. This domain also includes assignment to qualified, experienced, demographically representative or diverse, and/or effective teachers; and receipt of resources needed for learning, such as Wi-Fi, or receipt of mental health supports. Measures that reflect successful completion of courses should instead be reviewed under one of the Progressing in Education domains.

School Climate: Observations or assessments of the schoolwide or postsecondary institution environment or culture, as distinct from one's own behavior, such as the quality of social interactions, safety, engagement in school, sense of belonging within the school environment, staff cohesion, teacher-student relationships, family-teacher communication, perceived fit, and the prevalence of bullying at the school. This domain includes efforts made by educators to reach out to families and actions taken by educators to improve the school climate. Measures may be based on administrative data, self-reports, or observations.

School Leader Outcomes

School Leader Retention: Includes outcomes that measure the percentage of school leaders who return to work as a school leader, either in general, in particular settings such as economically disadvantaged districts, or in the same school, district, or state, from year to year.

School Leader Well-Being: Includes outcomes that measure school leader satisfaction; burnout; perceived ability to do one's job; intentions to continue leading; staff support, support from the district and school board; strength and quality of professional relationships; workload; job-related stress and anxiety, including emotional exhaustion and/or regulation; and general and physical well-being. Ability to do one's job also includes confidence and self-efficacy with respect to being able to lead generally and to lead specific instructional or professional practices; perceived ability to set instructional direction; perceived ability to improve student outcomes; access to adequate financial, human, and material resources, including access to data and technology.

School Leadership Practice: Includes outcomes that measure the quality of leadership ability demonstrated by the school leader. This includes the ability to recruit and retain staff, obtain resources for the school, engage families, and other skills related to the job. This is measured by rubrics assessed by supervisors or from surveys of school staff, families, or students.

Teacher Outcomes

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

Teacher Practice: Quality of instruction provided by teachers and their application of developmentally appropriate knowledge of content and/or pedagogy, as demonstrated by their actions in the classroom. This domain includes attempts to promote a positive and/or culturally responsive classroom environment and influence problematic student behavior by responding to student actions with consequences or rewards. Outcomes can be based on classroom observation rubrics assessed by school principals, supervisors, or trained evaluators or using student or family surveys. Teacher certification exam scores are reviewed under the Technical Skills Proficiency domain.

Teacher Retention: Includes outcomes that measure the percentage of teachers who return to work as a teacher, either in general, in particular settings such as special education or economically disadvantaged districts, or in the same school, district, or state, from year to year.

Teacher Well-Being: Includes outcomes that measure teacher satisfaction, burnout, perceived ability to do one's job; intentions to continue teaching; perceived leadership support and/or collegial support; perceived support for collaboration and professional development; workload; job-related stress and anxiety, including emotional exhaustion and/or regulation; feelings of safety; and general and physical well-being. Ability to do one's job also includes confidence and self-efficacy with respect to teaching in general and to specific types of instructional or professional practices and the perceived ability to improve student outcomes.

Other

Civic, Social, and Economic Participation: Participation in activities and demonstration of high-level skills beneficial for functioning within society. Example skills include information literacy and financial literacy. Activities include voting or registering to vote, volunteering, enrolling in health insurance, preparing tax returns, engaging in community gardening, and participating in other community activities or organizations. Measures could include records, assessments, or self-report. Basic life skills are included in the Functional Skills domain.

Functional Skills: Skills needed to participate in developmentally appropriate routines and activities in the home or in community settings and support independent living. This domain includes outcomes such as demonstrating the skills needed for a specific task, dressing, preparing and eating food, hygiene, cleaning, organizing, crossing the street, making a deposit or withdrawal from a bank, purchasing items, or using various forms of transportation. Measures may be based on a self-report, observation, or results from a clinical scale. Employability skills are in separate domains. This domain is most relevant for interventions in special education or early childhood education.

Motor Skills: The body's ability to use its muscles with control. This domain includes gross motor skills, such as the ability to sit, stand, walk, run, and jump, and fine motor skills, such as the ability to write, type, and cut with scissors. Measures that contain components of both Motor Skills and Functional Skills should be reviewed in the Functional Skills domain.

Physical Health and Nutrition: Measures of physical fitness, illnesses, vaccination rates, and other relevant outcomes. Measures can include administrative data, clinical scales, or well-established surveys.

Public Health and Safety: Outcomes that relate to community-level public health and safety factors. Measures can include rates of substance use disorders or other disease; unintended pregnancies; intimate partner violence; self-harm activities; and recidivism. Measures can include administrative data, clinical scales, or well-established surveys.

MAIN VERSUS SUPPLEMENTAL FINDINGS

As described in the *Handbook*, the WWC distinguishes *main* findings from *supplemental* findings ([pp. 128-129](#)). Main findings play a key role in determining a study's research rating and effectiveness rating for interventions. For example, effectiveness ratings and evidence tiers for interventions are determined based on main findings only (see *Handbook* pp. 130-134). The criteria below for determining main versus supplemental findings should be considered collectively, that is, if a finding is determined to be supplemental for any of the criteria below, the finding should be reviewed as supplemental.

Composite versus Subscale Measures. Main findings include those measured using the composite—not subscale—scores, unless only a subscale was administered in the study, or the

composite is not eligible for review. If composite and subscale scores are reported, subscale measures may only be reviewed as supplemental findings under the same domain as the composite.

However, if the composite falls under the Academic Achievement-K-12 or Academic Achievement-Postsecondary domain and the subscale is a standardized assessment under the Mathematics Achievement, Literacy Achievement, Science Achievement, Social Sciences, or Reading Comprehension domains, such as the SAT math, the subscale should be reviewed in the subject-specific domain and may also be reviewed as the main finding instead of the composite if the goal of the intervention is to affect outcomes in the subject-specific domain.

Dichotomous versus Continuous Measures. As indicated in the *Handbook* (p. 129), when both a continuous version and a dichotomized version of the same outcome are available, the WWC will treat the continuous version as a main finding and the dichotomized version as a supplemental finding. A dichotomous measure can be reviewed as a main finding if no corresponding continuous measure is reported.

In some cases, outcome measures are converted into multiple dichotomous outcomes and analyzed separately. For example, a study may measure whether students enroll in a 4-year college, 2-year college, or any college, and reports findings for each. In this case, reviewers should determine the purpose of the intervention when determining the main finding versus supplemental findings. Reviewers might classify enrolling in any college as the main finding and review other findings as supplemental. Alternatively, if the goal of the intervention was enrollment in a 4-year institution, reviewers could classify enrollment in a 4-year institution as the main finding and other findings as supplemental. Reviewers should not review as main findings multiple dichotomous outcomes that were converted from a single outcome in the same domain. Reviewers should first determine which dichotomous outcome should be the main finding according to the purpose of the intervention.

The *Handbook* (p. 27) also indicates that dichotomized measures must preserve the natural ordering of the latent variable. This means that if a continuous variable is recoded into multiple dichotomous variables, then it does not meet the face validity requirement of the WWC outcome measure standards. The only exception is if the dichotomous variable representing the greatest or lowest value of the variable has practical meaning, for example, an indicator of below proficient or an indicator of proficient/allowed.

Independent versus Nonindependent Measures. For literacy- and mathematics-related domains and domains related to broad student achievement marked with an asterisk (*) in the list of outcome domains, only findings from independent measures may be designated as main findings. Findings in these domains from nonindependent measures will be reviewed as supplemental. A list of known independent measures in literacy- and mathematics-related domains is available on the [WWC Website](#). This list is not exhaustive and will be updated on a quarterly basis as new measures are added. If a study was reviewed using a nonindependent measure that later becomes classified as independent, the review team may republish the review to reflect the new status.

As stated in the *Handbook* (p. 28), a measure will be considered nonindependent if either it was developed by study authors and is not documented as in use by non-overlapping research teams and apart from the intervention, or if it was developed by the intervention's developers. Measures created by intervention developers are considered independent when implemented with an intervention created by a different developer. When determining independence, the WWC looks for at least two studies that use the measure that were conducted by non-overlapping research teams that do not contain an intervention's developer.

Intent-to-Treat versus Treatment-on-the-Treated. The WWC generally reviews intent-to-treat findings as main findings and treatment-on-the-treated findings as supplemental findings. Review teams have flexibility in determining whether to review the treatment-on-the-treated findings as supplemental findings unless they are required for the purpose of the review.

No Main Findings Meet Standards. Review teams should review all eligible supplemental findings in a manuscript, even if no main findings meet WWC standards. However, if no main findings meet WWC standards, the review team does not need to conduct author queries for supplemental findings. If the review team needs to conduct an author query to determine whether the main findings meet WWC standards, it is recommended to also conduct author queries for supplemental findings at the same time.

Subgroup Findings. The WWC classifies the findings for the full analytic sample as the main findings. Generally, for studies where at least one main finding meets WWC standards, the WWC also reviews the following subgroup findings as supplemental if they are reported in a study: (i) students with disabilities or developmental delays; (ii) students at risk of low performance in academics or problem behaviors according to a standardized baseline measure; (iii) dual language learners, English learners, or non-native English speakers; (iv) racial or ethnic groups; and (v) economically disadvantaged students. Reviewers should report these subgroups as they are reported in the study as long as they are conceptually similar to the previously listed subgroups. For example, a subgroup based on low parental educational attainment should be reviewed if there are no subgroups based on economic disadvantage. Racial and ethnic subgroups should be reported as similar to the subgroups in the study as possible. For example, if a study reports race by census category, each racial group should be reported as a separate supplemental finding. And if a study reports findings for only one racial subgroup, the review should report the subgroup and does not need to conduct an author query for additional subgroups. Subgroups that represent less than 15% of the analytic sample do not need to be reviewed.

Timing of Outcome Measures. Sometimes studies present findings at different time periods. For most outcome domains, the time period closest to the end of the intervention, and no more than 1 year following the conclusion of the intervention, is considered the main finding, and all other follow-up time periods are considered to be supplemental.

Exceptions to this rule are the Earnings, Employment, College Degree Attainment, and Industry-Recognized Credential domains. In these cases, main findings are those measured at least 1 year after

the conclusion of the intervention or at the expected graduation date, and findings near the end of the intervention and other follow-up findings are supplemental. Review teams may justify selecting main findings more than 1 year after the conclusion of the intervention based on the goal of the intervention.

Other. The WWC may review other supplemental findings that are relevant to the purposes of the WWC's review. The WWC generally does not review sensitivity analyses as described in the *Handbook* (p. 130).

BASELINE EQUIVALENCE

Demographic Baseline Measures. When demonstrating equivalence on the sample demographic characteristics listed on page 55 of the *Handbook*, the reviewers should assess baseline equivalence for two baseline measures. For categorical demographic measures, baseline equivalence should be reported for categories that account for at least 15% of the study sample in both the baseline and outcome periods. For example, if a study reported 55% White, 35% Black, 4% Hispanic, 3% Asian, and 3% other, baseline equivalence is only needed for White and Black students. If more than two acceptable baseline measures are reported, review team leadership may choose which baseline measures are the most relevant to the study.

GPA as Baseline Measure in Postsecondary Studies. As stated in the *Handbook* (p. 27), GPA is an allowable measure for demonstrating baseline equivalence as a broad, standardized measure of student academic readiness, knowledge, and skills if GPA is calculated in the same manner for the intervention and comparison groups. For postsecondary studies, GPA is an allowable measure for demonstrating baseline equivalence if the GPA is on the same scale, even if it is not calculated exactly the same way for all students. There are no changes in the requirement that GPA must be calculated in the same manner for the intervention and comparison groups for non-postsecondary studies.

Multiple Baseline Measures in Same Domain. This section provides guidance for reviewing main findings that report multiple baseline measures for the same outcome domain and for the same analytic sample when baseline equivalence is required. A baseline difference on any main finding in the domain that is larger than 0.25 standard deviations means that all main findings in the outcome domain fail to satisfy the baseline equivalence standard, as specified in the *WWC Procedures and Standards Handbook, Version 5.0* (p. 56). The standard does not apply to supplemental findings. Review teams should consult with the WWC Contractor Help Desk and their Contracting Officer's Representative (COR) if they believe that the baseline measures within the same domain are not highly related, and if the Contractor Help Desk and the COR agree that the baseline measures are unlikely to be highly related, then the review team can assess baseline equivalence on a finding-by-finding basis.

For some study designs, the *Handbook, Version 5.0* requires a statistical adjustment of a baseline measure(s) to satisfy the baseline equivalence standard. The *Handbook* does not require multiple statistical adjustments if there are multiple baseline measures and multiple main findings in the same domain with the same analytic sample.

OTHER

COVID-19 Related Flexibilities. For studies conducted between March 2020 and July 2022, the WWC allows for additional flexibilities to account for the global pandemic and related school closures. The WWC will allow cluster-level baseline equivalence to be established using data from Fall 2019 or the 2018-2019 school year if baseline data collection occurred between March 2020 and July 2022 and was affected by COVID-19, regardless of the time elapsed between the baseline and outcome data collection.

In addition, if outcome data were collected between March 2020 and July 2022, the study can be reviewed using the optimistic boundary, even if the *Handbook* suggests that the cautious attrition boundary should be used. Studies reviewed using this flexibility will have a disposition rating that reflects these flexibilities.

Design Comparable Effect Size (D-CES) for Cluster Outcomes. The WWC will calculate a Design Comparable Effect Size (D-CES) for cluster outcomes in single case design studies for cluster-level outcomes if they are based on measurements of individual outcomes aggregated to the group level. For example, the WWC will calculate a D-CES when an outcome is based on the percentage of students in a class exhibiting disruptive behavior, so long as each student's behavior was systematically observed. However, the WWC will not calculate a D-CES for cluster-level outcomes when measurements are based on scans of the cluster without a fixed method for individually observing each student in the cluster (e.g., if the observer scans the entire classroom at once but does not systematically observe each student). The WWC will also not calculate a D-CES if measurements are based on small groups of students within the cluster instead of individual students.

Reliability. To clarify the reliability of outcome measure requirement in the *Handbook* (p. 27), reliability may be established using only one metric of internal consistency, test-retest reliability, inter-rater reliability, or inter-rater agreement. If a study reports multiple reliability metrics for a measure, and at least one metric meets the WWC reliability standards, then the measure meets the WWC reliability standard. However, the reliability metric should be appropriate for the measure. For example, reliability for teacher observations should be determined using inter-rater metrics.

Direction of Effect Sizes. Favorable outcomes should be represented by positive effect sizes, which may require reviewers to select the option to reverse the effect size direction for some domains in the Online Study Review Guide.

Pilot Review Process. The Version 5.0 Standards describe a review process in which three certified reviewers—two reviewers and a reconciler—are needed to complete most WWC reviews. Under the pilot review process, the WWC will assign two reviewers to each group design study review: a lead reviewer and a deputy reviewer. The lead reviewer will be an experienced reviewer who will be responsible for an accurate and thorough review. The pilot will assess whether this streamlined review process enables the WWC to review more efficiently while maintaining quality. The review process will proceed as follows:

1. The lead and deputy reviewers read the study independently, taking notes on preliminary review decisions, identifying the eligible and ineligible findings in the study, and identifying any missing information that may need to be requested from the study authors.
2. The two reviewers meet to discuss their notes and agree on how to address any differences to correctly apply the standards to the study. If the two reviewers cannot agree on how to apply the standards, they will ask a senior member of the review team, such as a methodologist, for help resolving the issue.
3. The deputy reviewer completes a formal review of the study in the Online Study Review Guide, noting how any information that needs to be requested from the study author could affect the review. The deputy reviewer will also prepare a draft of the specific questions for the study authors.
4. The lead reviewer will examine the review and materials for the author query and then work with the deputy reviewer to finalize the author query, if needed. The lead reviewer may request that a senior member of the review team review the author query materials before they are sent.
5. After the WWC receives a response to the query or the typical two-week response period has elapsed, the lead reviewer may ask the deputy reviewer to update the review with the new information. The reviewers may elect to send an additional author query.
6. The lead reviewer finalizes the review in the Online Study Review Guide, consulting with the deputy reviewer.

Review teams have the discretion to adapt these procedures as needed to support efficiency. Studies using regression discontinuity designs, single case designs, or randomized controlled trials that estimate complier average causal effects will be reviewed using the standard procedures described in the Version 5.0 Standards.

Use of AI to Support Study Reviews. Used appropriately, AI tools have the potential to improve the accuracy, completeness, and efficiency of WWC reviews. Under this review protocol, the WWC may use AI in (1) conducting literature searches; (2) screening studies to identify those eligible for review; (3) extracting information from studies, such as outcome measure names, sample sizes, and findings; (4) summarizing the study context and the implementation of the intervention and comparison conditions; (5) applying evidence standards; and (6) preparing requests to study authors. In integrating AI, the WWC will (1) use tools well aligned to a specific purpose or need; (2) consider and address data privacy, security, intellectual property, and copyright concerns; (3) and employ human oversight and review in all phases of the review process. In particular, it is essential that trained and certified human reviewers continue to conduct WWC study reviews. In addition, to support accurate reviews, all WWC reviews of studies that meet standards must continue to undergo peer review by an independent certified reviewer.