REVIEW PROTOCOL FOR TEACHER EXCELLENCE VERSION 4.0 (MAY 2019)

This protocol guides the review of research that informs the What Works Clearinghouse (WWC) intervention reports in the Teacher Excellence subtopic area. The protocol is used in conjunction with the *WWC Procedures Handbook (version 4.0)* and the *WWC Standards Handbook (version 4.0).*

PURPOSE STATEMENT

Research has shown that teacher effectiveness is the most important school-based factor that influences student outcomes, including student achievement. Studies have also shown that there is substantial variation in teacher effectiveness, such that some teachers improve student outcomes at a much faster rate than others. The strong influence of teachers on student outcomes, as well as the variation in teacher effectiveness, has led to the creation of numerous programs designed to help teachers improve student outcomes.

This review focuses on interventions aimed at making teachers more effective at improving the achievement of students in grades PK–12. These interventions are targeted at adults who are considering teaching; undergoing teacher preparation; or already employed in the teaching profession. The interventions may be delivered in a variety of educational and training settings. Although the interventions are delivered to teachers (or potential teachers), this review focuses on outcomes for both students and teachers. Student outcomes include measures of achievement, progression in school, social-emotional learning, and behavior. Teacher outcomes include those that research has shown are related to these student outcomes, including teacher retention and instructional practice.

The following research questions guide this review:

- Which teacher-focused interventions improve achievement, progression in school, social-emotional learning, and behavior for students in grades PK–12?
- Which teacher-focused interventions improve outcomes for teachers that are related to these student outcomes?

KEY DEFINITIONS

Teacher. An adult who is employed by a school or school district to provide instruction to students in grades PK–12.

This definition *includes* the following groups of individuals:

- Those who provide students with at least 50% of instruction in a subject (that is, are teachers of record), regardless of whether those students are general education students, students with special needs, or a combination of general and special education students
- Long-term substitutes, defined for this review as individuals who fill in for particular teachers for more than half the intervention period

This definition *excludes* the following groups of individuals:

- Those who provide instruction outside of school hours (for example, tutors)
- Those whose primary role is administrative or supervisory (for example, principals, deans, and superintendents)
- Those providing instruction to individuals outside of grades PK–12 (for example, college lecturers)
- Those who provide non-instructional support to students (for example, nurses, school psychologists, speech language pathologists)
- Short-term substitutes, defined for this review as individuals who fill in for particular teachers for half the intervention period or less
- Teaching aides or paraprofessionals who provide students with less than 50% of instruction in a subject

Teacher candidates. Individuals who are participating in a teacher pre-service training program. These individuals may be in a traditional teacher preparation program at a college or university, or using an alternative route to become certified. Individuals are no longer teacher candidates when they complete their training program and become certified as a teacher.

Pre-service training. Training of prospective teachers as part of becoming certified. Generally, this occurs before being hired by a school/district and becoming solely responsible for a classroom or becoming the teacher of record. Pre-service training can involve coursework completed toward becoming certified, as well as student-teaching or other practice-based experiences.

CATEGORIES OF RELEVANT RESEARCH

The review team identified and defined five categories of research examining interventions aimed at making teachers more effective at improving student outcomes. Those categories are:

- ***Teacher preparation:*** Studies of programs that train individuals to serve as teachers, including both traditional programs based in college or university schools of education and nontraditional (or alternative) teacher training programs.
- ***Teacher induction:*** Studies of programs that provide specialized training targeting the needs of novice teachers, including those in their first 3 years of service.
- ***Teacher evaluation:*** Studies that examine processes or systems used to determine teacher performance for the purposes of identifying professional development needs, providing formative feedback to teachers, setting compensation, or making other personnel decisions.
- ***Teacher compensation:*** Studies of programs or systems that provide monetary incentives to teachers for improving student academic achievement, teaching performance, or related school outcomes; taking on additional professional responsibilities; demonstrating particular knowledge or skills; or filling hard-to-staff

teaching positions. Such systems might include bonuses or salary structures that differ from a traditional teacher salary schedule.

- ***Teacher professional development:*** Studies of programs that provide training addressing the needs of teachers, including increasing their content knowledge (about the academic subjects they teach), pedagogical content knowledge (about how students learn), and the ability to analyze student work and achievement and to use this analysis to adjust instructional strategies, assessments, or materials. This definition of professional development shares some elements of a more narrow and specific definition under the Every Student Succeeds Act (ESSA), but the set of activities that can be considered professional development under this review is broader than those identified by the ESSA definition.

ELIGIBILITY CRITERIA

Eligible Populations

Studies that include teachers or teacher candidates can be eligible for reviews under this protocol. Additionally, studies that include students who are on track to develop grade- (or age-) appropriate skills, students who are at risk for academic difficulties, and students with disabilities can be eligible for reviews under this protocol. In particular, the students in an eligible sample may be students classified as English learners or receiving special education services. In this review, the following populations are of interest:

- ***Grade range.*** Teachers must provide instruction to students in grades PK–12.
- ***Location.*** Teachers must be employed by schools located within the United States, its territories or tribal entities, or at U.S. military bases overseas.

Eligibility of Findings from Multiple Analyses in a Study

This review follows the guidance in the *WWC Procedures Handbook* (in Chapter VI: Reporting on Findings) regarding reporting on findings from subgroups, multiple analyses that use composite or subscale scores, or different time periods. In particular, the WWC reports findings from all eligible analyses that meet standards, split into main and supplemental findings. The rating of effectiveness for an intervention is based on the main findings. Other eligible findings that meet standards can be included in supplemental appendices to the intervention report. For each outcome measure, and among those findings that meet WWC design standards, the WWC uses the following criteria to designate one finding or set of findings as the main finding: (1) includes the full sample; (2) uses the most aggregate measure of the outcome measure (rather than individual subscales); and (3) is measured at a time specified by the protocol.

Under this review, findings for the subgroups listed in Exhibit 1 are eligible to be reported in supplemental appendices to the intervention report. Findings for other subgroups are not eligible for review (unless designated as the main finding based on the criteria above).

Exhibit 1. Subgroups of Interest to the Teacher Excellence Topic Area

Characteristics of:		
Teachers	**Students**	**Classrooms or schools**
• Experience • Certification status or credentials • Demographic characteristics • Subject area	• Special education status • English learner status • Economically disadvantaged (for example, free or reduced-price lunch status) • Grade • Low-achieving (as defined in the study based on an eligible outcome measure) • Race/ethnicity • Gender	• Location of the instructional setting (for example, urban, suburban, rural) • School governance (for example, traditional public, charter, private, religious) • Economically disadvantaged (for example, Title I status or percentage of students with free or reduced-price lunch status) • Accountability rating (as defined by an educational agency such as the state or district)

For this review, measures obtained at the end of an intervention, as well as any time thereafter, are admissible. When reported, this review will classify findings for outcomes administered immediately after the intervention (for example, outcomes administered after the third year of a 3-year intervention is completed) as main findings because these findings are most prevalent in the studies reviewed under this topic area. Measures occurring several months or years after the intervention may also provide strong evidence for an intervention's effectiveness. Intermediate outcome measures that reflect partial exposure to an intervention can also provide useful information about the intervention's effectiveness. Therefore, follow-up and intermediate findings, when available and appropriate, may be reported in supplemental appendices to the intervention report.

Main findings for this review will include eligible teacher outcomes in the domains listed below in the section for eligible outcome measures. These teacher outcomes include those that research has shown are related to student outcomes. Findings for additional teacher outcomes that may have less evidence of relationships with student outcomes are also eligible for review as supplemental findings. Outcomes that could be reported in supplemental appendices include measures of content knowledge, pedagogical content knowledge, summative accountability measures, and other measures used to make high-stakes decisions. Appendix A contains additional details about how the WWC will review these supplemental findings against evidence standards.

While the above rules will guide how main and supplemental findings are identified, review team leadership has discretion to identify main and supplemental findings after considering additional factors about the findings under review, such as the prevalence of findings across implementation levels and the design of the intervention.

Eligible Interventions

Only interventions that are replicable (can be reproduced in another setting) are eligible for review. The following characteristics of an intervention must be documented to reliably reproduce the intervention with different participants, in other settings, and at other times:

- Intervention description: skills being targeted, approach to enhancing the skill(s) (for example, strategies, activities, and materials), unit of delivery of the intervention (for example, whole group, individual), medium/media of delivery (for example, teacher-led instruction or software), and targeted population (for example, teachers or teacher candidates);
- Intervention duration and intensity; and
- Description of individuals delivering or administering the intervention and any significant training they received that is essential to their role.

In this review, the following types of interventions may be included:

- ***Practices.*** A *practice* is typically an action taken by teachers as they plan, implement, or evaluate instruction. The practice must be clearly described and commonly understood in the field and literature. For example, instructional scaffolding is a teacher-enacted practice of providing strong support to students when introducing topics and concepts, but then gradually withdrawing that support so students can use/integrate the new concepts independently. An intervention that provides professional development to support such a practice is of interest to this review. An intervention that trains teachers to support a practice within a specific content area and is focused on improving outcomes in that same area may be eligible for review under both this protocol and another WWC review protocol. For example, professional development to train teachers to use a new approach to support students to graph equations might also be reviewed under a math topic area review protocol.
- ***Policies.*** A *policy* is a named condition, system, or set of formal rules that affects teachers. The policy must be commonly understood in the field and literature. Policies may be set by federal, state, or local governments or by the organization providing services. Policies may focus on changing teachers' behaviors or motivation. Examples of teacher-focused policies of interest include:
    - Financial incentives for effective teaching, or
    - Professional development requirements for renewal of certification.
- ***Programs.*** A *program* is a system of training supports that aim to improve the performance of teachers. Examples of interest include well-defined teacher preparation and professional development programs. Programs that bundle other substantive activities with teacher-focused activities, such as school turnaround strategies that seek to improve several aspects of the whole school, are not eligible under this review.

Both “branded” and “non-branded” interventions will be reviewed. Branded interventions are commercial or published programs and products that may possess any of the following characteristics:

- An external developer who provides technical assistance (for example, instructions/guidance on the implementation of the intervention) or sells or distributes the intervention
- Trademark or copyright

Eligible Research

The *WWC Procedures Handbook* discusses the types of research reviewed by the WWC in Section II: Developing the Review Protocol and Section III: Identifying Relevant Literature. In this review, the following additional parameters define the scope of research studies to be included:

- ***Topic.*** The study must examine the effect of teacher preparation, teacher induction, teacher evaluation, teacher compensation, or teacher professional development on student academic achievement, progression in school, or social-emotional learning/behavior outcomes, or teacher behaviors or knowledge linked to these student outcomes.
- ***Time frame.*** For new intervention reports, the study must have been released within the 20 years preceding the year of the review (for example, in 1999 or later for reviews occurring in 2019). For updated intervention reports, the study must have been released since the original interventions report’s literature search start date (for example, if the original report used 1989 literature search start date, the updated report will continue using the same date). Studies must be publicly available (accessible online or available through a publication, such as a journal) at the time of the original or updated literature search.
- ***Sample.*** The study sample must meet the requirements described in the “Eligible Populations” section above.
- ***Language.*** The study must be available in English to be included in the review.
- ***Location.*** The study must include teachers working in the United States, its territories, or tribal entities, or at U.S. military bases overseas.

Eligible Outcome Measures

This review includes outcome measures in multiple domains (Exhibits 2 and 3). In the first column of these tables, we organize the various student and teacher domains into groups for reference, and we list the domains in the second column with descriptions. Unlike the domains themselves, which can influence how the outcomes are reviewed and reported on in intervention reports, the purpose of the groups is only to support presentation. Eligible outcome measures in any domain must be administered in English.

Exhibit 2. Eligible Student Outcome Domains

Group	Outcome domain name and description
Student achievement	***General literacy achievement***—includes outcomes in the following areas: language development, alphabetics (including phonemic and phonological awareness, letter identification, print awareness, and phonics), foundational reading (word reading, fluency and/or accuracy in reading connected text, vocabulary, reading comprehension), general reading, measures of English language conventions (for example, grammar), writing, and general English language arts achievement (such as a standardized test covering an array of language arts topics).
	General mathematics achievement—includes outcomes in the following areas: basic number concepts; number operations; patterns and classification; measurement; understanding of different subjects within mathematics, including algebra, arithmetic, calculus, geometry, probability, statistics, and trigonometry; understanding of concepts and procedures; understanding of word problems and applications; and general math achievement (such as a standardized test covering an array of mathematics topics).
	General science achievement—includes outcomes in the following areas: science facts, and the capacity to use the tools, procedures, inquiry, nature of science, argumentation in science, and reasoning processes of science. This includes subjects such as biology, chemistry, and earth science.
	General social studies achievement—includes outcomes in social studies subdisciplines, such as civics, economics, geography, history, and world cultures.
	General achievement—includes a general measure of student academic achievement, only to be documented if study authors do not distinguish students' achievement in specific areas (for example, math, reading). Examples include composite scores from state assessments that represent a combination of reading and math scores.
	English language proficiency—includes outcomes in the areas of *vocabulary, oral language, listening comprehension, and grammar*. This domain is only for use with majority English learner samples. For other samples, the outcome measures that would otherwise be eligible within this domain should be reviewed in another student achievement domain. *Vocabulary* includes understanding the meanings of words (receptive vocabulary) and using words appropriately (expressive vocabulary). It includes the understanding of *academic language*, which is the language used for formal discourse in academic disciplines such as mathematics, literature, economics, science, and history. Terms that cross disciplines (such as "in contrast," "permutation," "enable," "facilitate," and "comprehensive") are part

Group	Outcome domain name and description
Student achievement *(continued)*	of academic language, but a measure of understanding terms that are unique to one of math, science, or social studies (such as "hypotenuse," "thermodynamics," or "angular momentum") should be reviewed under a student achievement domain specific to that subject (for example, "hypotenuse" would be part of the general mathematics achievement domain). *Oral language* includes listening and speaking skills. *Listening comprehension* refers to understanding spoken language. *Grammar* refers to the appropriate use of language (spoken or written) in terms of syntax (sentence structure) or morphology (word inflections).
Student progress in school	***Staying in school***—includes outcomes that measure whether the student has dropped out of school and the number of days the student was enrolled in school.
	Progression in school—includes outcomes that assess the number of high school course credits the student has earned, whether the student was promoted to the next grade, and the highest grade the student has completed.
	Completing school—includes outcomes that measure whether the student has earned a high school diploma or GED or whether he or she has graduated from a high school.
Social-emotional learning/behavior	***Student social interaction***—includes student behaviors that primarily involve interactions with others, or reflect attempts at social interactions. Examples include observed or perceived peer rejection; isolation; victimization; actions characterized as bullying (including physical, relational, cyber); sexual harassment; other aggressive behavior to others; specific social skills, such as social awareness of context and others, and interpersonal relationship skills; and other measures of behavior that are intended to either benefit (sometimes called prosocial behavior) or harm/hurt others.
	Observed individual behavior—includes observed or recordable student behaviors that primarily reflect individual choices and have individual consequences for the student (sometimes referred to as externalizing behaviors). A key differentiating factor between outcomes in this domain and outcomes in the student social interaction domain are that outcomes in this domain do not require interactions with others. Examples include delinquent behaviors, such as lying or stealing; impulsivity; arrests; substance abuse; ratings of adaptive functioning; and degree of self-control/self-regulation/self-management. Eligible behaviors in this domain can occur in or away from school. However, behaviors that are solely situated in a school context are not eligible in this domain, but may instead be eligible measures in the school engagement domain, below.

Group	Outcome domain name and description
Social-emotional learning/behavior *(continued)*	***Student emotional status***—includes student behaviors and self-ratings that are primarily focused inward and reflect a student's emotional state (sometimes referred to as emotional or internalizing behaviors). Examples of emotional status measures include self-awareness of thoughts, feelings and behavior; thought disorders; emotion regulation; depression; and overall adjustment/well-being.
	Student engagement in school—includes behaviors that are typically only observed during school and often reflect school connectedness. This outcome domain includes outcome measures that might otherwise be eligible within the student social interaction domain, the observed individual behavior domain, or the student emotional status domain, except that the specific behavior assessed by the measure would only be observed in a school setting. Examples include school attendance; school suspensions; cheating on a test; disrupting class; truancy or absences; following school rules; coming to class prepared; staying on task during a class assignment; and participating in school activities. Engagement is also demonstrated when students indicate they put effort into being successful in school.

Additional notes on student outcome domains:

- Eligible student outcomes may be measured at the student or cluster level (for example, classroom, teacher, or school). For example, the percentage of grade 12 students in a school who graduate is an eligible school-level outcome in the student progression domain. Similarly, student gain scores aggregated to the teacher or school level are eligible for review. Also, scores from a teacher or school effectiveness model can be eligible for review if, for example, the results are based on an average or median of adjusted student test scores, including value-added measures and median student growth percentiles.
- Eligible measures of social-emotional learning/behavior may be based on administrative records, student self-reported measures and surveys, a diagnosis or classification, an assessment scale, observations by educators or trained staff, or parental reports. However, measures based on teachers' observations or self-reports are not eligible when teachers who provide these types of data participated in the intervention. This is because teachers might be influenced by knowing their study condition. As an example, the Internalizing Problems Subscale of the Behavior Assessment Scale for Children (BASC) includes teacher and parent reports, and student self-reports. The teacher reports would only be eligible if completed by someone besides the teacher participating in the intervention. However, parent reports and student self-reports would potentially be eligible.
- Outcomes measuring student health, nutrition, or course grades, and teacher reports of student proficiency, are not eligible outcome measures.

Exhibit 3. Eligible Teacher Outcome Domains

Group	Outcome domain name and description
Instruction	***Instructional practice***—includes outcome measures that reflect the quality of instruction provided by teachers and their application of content knowledge or pedagogical content knowledge as demonstrated by their actions in the classroom. These measures can be based on rubrics assessed by school principals, supervisors, or trained evaluators, or based on surveys administered to students. Eligible assessments of the quality of teacher instruction must satisfy a validity requirement described below. Those known to satisfy this requirement include, but are not limited to: • Charlotte Danielson's Framework for Teaching (FFT) • Classroom Assessment Scoring System (CLASS) • Protocol for Language Arts Teaching Observations (PLATO) • Mathematical Quality of Instruction (MQI, predicting mathematics achievement) • Tripod • UTeach Teacher Observation Protocol (UTOP)
Teacher behavior	***Teacher attendance—***includes outcomes that indicate the number (or percentage) of eligible work days for which the teacher is present.
	Teacher retention at 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. Measures of retention in particular instructional settings (for example, percentage of teachers in urban settings who return to teach in an urban setting) are also included in this domain.
	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.

Additional notes on measures of instruction

- To be eligible in this domain, the outcome must measure what teachers do in the classroom or how their actions are perceived by students. Outcomes that measure teachers' acquisition of content knowledge or pedagogical content knowledge (such as the Praxis) are not eligible to be reviewed as main findings because they are not known to have as consistent and strong associations with student outcomes compared to those for measures of instruction. However, findings for content knowledge or pedagogical content knowledge may be eligible to be reviewed as supplemental findings and reported in appendices to the intervention report.
- As described in the section on outcome measure requirements below, eligible measures in the instructional practice domain must satisfy a validity requirement with a statistical relationship between the outcome measure and student achievement, progression in school, or social-emotional learning/behavior. All of the measures of instruction named above meet this validity requirement for student achievement based on evidence reported in the *Measures of Effective Teaching* study (Kane & Staiger 2012).

Additional notes on teacher retention measures:

- This review focuses on outcomes that assess whether or not teachers return to their school, their school district, their state, or their profession from year to year in their same role (for example, teachers who return as teachers). When reported, this review will consider a measure of retention in the same role as the main outcome measure, but also eligible for review are outcomes that count teachers who move to other instructional positions (such as a teacher who returns as a principal or instructional specialist) as remaining in their school, school district, or state.
- More detailed mobility outcomes generally will not be reviewed because they either are (1) captured by the key, commonly measured retention outcomes of interest or (2) may not be defined consistently across studies. For example, an indicator for moving to another specific high school in the school district would be captured by a broader outcome that measures whether a teacher returned to teach in the same school district, and the interpretation of the move to another high school might depend on the characteristics of that high school.
- An eligible teacher retention outcome must be measured based on teachers' actual movement from a teaching position, not expected movement. For example, teacher ratings on whether they expect to return to their positions are not eligible for review.
- Other measures of teacher tenure or professional achievement, or of teacher satisfaction or plans, are not eligible for review within these outcome domains.

For composite outcomes that include components that do not fall in a single eligible domain (for example, a school performance score that is based on student achievement and school climate), eligibility of the composite outcome will be determined as follows:

- if the study reports that 75% or more of the components of the composite outcome are eligible outcomes, the composite outcome is eligible;
- if the study reports that less than 50% of the components are eligible, the composite outcome is ineligible; and,
- if the study reports that at least 50% but less than 75% of the components are eligible, the review team leadership has the discretion to determine whether or not the composite outcome is eligible, based on the properties of the individual measures.

EVIDENCE STANDARDS

Eligible studies are assessed against WWC evidence standards, as described in the *WWC Procedures Handbook* Section IV: Screening Studies and Section V: Reviewing Studies, as well as the *WWC Standards Handbook.*

Sample Attrition

The *WWC Standards Handbook* discusses the sample attrition standards used by the WWC in the following sections:

- Step 2 of the WWC review process for individual-level group design studies in Section II.A—"Sample Attrition: Is the combination of overall and differential attrition high?"
- Step 1 of the WWC review process for cluster-level group design studies in Section II.B—"Is the study a cluster RCT with low cluster-level attrition?"
- Step 3 of the WWC review process for cluster-level group design studies in Section II.B—"Is there a risk of bias due to non-response of individuals?"
- Section 3 of the WWC standards for reviewing complier average causal effect estimates in Section II.D—"Calculating attrition when rating CACE estimates"
- Standard 2 of the WWC standards for reviewing regression discontinuity designs (RDDs) in Section III.C

This review uses the ***optimistic*** boundary for attrition. In the *WWC Standards Handbook*, Figure II.2 illustrates the attrition boundary and Table II.1 reports attrition levels that define high and low attrition. Based on the choice of the boundary, the study review guide calculates attrition and whether it is high or low.

This choice of boundary was based on the assumption that most attrition in studies of teacher training, evaluation, and compensation was due to factors that were not strongly related to intervention status. For example, most attrition in these teacher-focused interventions results from factors that are unrelated to intervention group membership, such as teachers' absence on the day of observations or parent mobility and students' absence on the days that assessments are conducted.

Joiners in Cluster Randomized Controlled Trials (RCTs)

The WWC defines a *joiner* as any individual (student, teacher, or school leader) who enters a cluster (for example, a school or classroom) after the results of random assignment are known to any person who could influence a student's placement into a cluster (for example, parents, students, teachers, principals, or other school staff). The presence of joiners in an analytic sample has the potential to introduce bias into estimates of an intervention's effectiveness.

In some cases, joiners who enter clusters relatively early in the study period have less potential to introduce bias than those who enter later. Therefore, the WWC sometimes differentiates between *early joiners* and *late joiners.* For this review protocol, we will consider a student to be an *early joiner* if they enter a cluster in the 6 weeks after the results of random assignment are known, or, in cases where random assignment occurred during the summer, 6 weeks after the start of the school year. *Late joiners* are those that enter clusters after the end of the early period.

This review protocol specifies the following rules:

a. In cluster RCTs where the unit of assignment is a classroom or another group defined within a school (such as a group of classrooms or a small group of students within a classroom), all joiners pose a risk of bias. This is because classroom rosters are often determined by school administrators who might assign students to classrooms based on knowledge of the intervention. Additionally, students or parents might influence their assignment to clusters (for example, classrooms) because they may have a specific preference for or against the intervention. Therefore, a study that includes at least one such joiner in the analytic sample does not limit the risk of bias from joiners.

b. In cluster RCTs where the unit of assignment is a school or a group of schools (such as a district), whether joiners pose a risk of bias depends on whether the intervention is expected to influence school enrollment or placement decisions.

- If the intervention may affect enrollment or placement decisions, then ***all joiners pose a risk of bias.*** A study of such an intervention that includes one or more joiners in the analytic sample ***does not limit the risk of bias from joiners.***
- If it is unlikely that the intervention affects enrollment or placement decisions (such as a low-profile teacher induction or professional development program), then ***only late joiners pose a risk of bias.*** A study of such an intervention that includes at least one late joiner in the analytic sample ***does not limit the risk of bias from joiners.***

For this review, the default assumption is that the interventions being examined with assignment at the school level or higher are unlikely to affect enrollment or placement decisions; however, review team leadership has discretion to revise this assessment.

Additionally, the typical scenarios the WWC encounters in cluster RCTs for this topic area are described above, but we cannot anticipate all scenarios. When an intervention and unit of assignment in a cluster RCT do not fall into a category described above, review team leadership has discretion to make a decision on which joiners pose a risk of bias.

BASELINE EQUIVALENCE

If the study design is an RCT or RDD with high levels of attrition or a quasi-experimental design (QED), the study must satisfy the baseline equivalence requirement for the analytic intervention and comparison groups. The *WWC Standards Handbook* discusses how authors must satisfy the baseline equivalence requirement in:

- Step 3 of the WWC review process for individual-level group design studies in Section II.A—"Baseline Equivalence: Is equivalence established at baseline for the groups in the analytic sample?"
- Steps 4 and 7 of the WWC review process for cluster-level group design studies in Section II.B—"Does the study establish equivalence of individuals at baseline for groups in the analytic sample?" and "Does the study establish equivalence of clusters at baseline for groups in the analytic sample?", respectively.
- Section 5 of the *WWC Standards* for reviewing complier average causal effect (CACE) estimates in Section II.D—"Procedures for rating CACE estimates when attrition is high"
- Standard 3 of the *WWC Standards* for reviewing RDDs in Section III.C

This review assesses baseline equivalence within each domain and analytic sample. In particular:

- If baseline differences *for a given analytic sample* exceed 0.25 standard deviations for any pre-intervention measure within a domain (or any acceptable alternative pre-intervention measure for the domain when no pre-intervention measure is available from within the domain, as described in the section on baseline equivalence of individuals below), study findings using this analytic sample will not meet WWC group design standards within this domain. However, findings using different analytic samples and outcomes in different domains may still be eligible to meet WWC group design standards.
- When the baseline difference for a pre-intervention measure is in the statistical adjustment range (that is, it is between 0.05 and 0.25 standard deviations), the adjustment must be made only in the analysis of the associated outcome measure. For example, if A, B, and C are available as pre- and post-intervention measures all within one domain, and the pre-intervention difference in B requires statistical adjustment, only the analysis of outcome B must adjust for B.

In addition to the pre-intervention measures that are required for satisfying the baseline equivalence requirement, other sample characteristics such as student age and grade level, may be associated with the outcome. A large baseline difference on these characteristics could be evidence that the intervention and comparison groups are not sufficiently comparable for the purposes of the review. When differences in student age or grade level are larger than 0.25 standard deviations, the study will be rated *Does Not Meet WWC Design Standards*. If the study does not report these characteristics, but describes a study sample that gives the reviewer reason to question the magnitude of the differences on these characteristics, the review team leadership

has the discretion to conduct an author query to obtain information on the similarity of the groups based on age and grade level.

1. Baseline equivalence of individuals

For studies that must satisfy baseline equivalence of individuals, including cluster-level assignment studies being reviewed for evidence of effects on individuals, the baseline equivalence requirement must be satisfied for the analytic intervention and comparison groups on one of the following pre-intervention (or baseline) characteristics:

- A pre-intervention measure within the same domain; or,
- If a pre-intervention measure is not available, one or more acceptable alternative pre-intervention measures, as explained below and summarized in Exhibit 4.

Acceptable pre-intervention measures for student achievement outcomes. For outcomes in the five student achievement domains, studies must show that the groups are equivalent on an acceptable pre-intervention measure of student achievement. A pre-intervention measure in the same subject as the outcome is preferred; however, if a same-subject pretest is not available, a pre-intervention measure of general achievement (for example, a combined mathematics and reading score) is acceptable. In addition, a pre-intervention measure of mathematics or literacy achievement can be used to establish baseline equivalence for a science achievement outcome, and a pre-intervention measure of literacy achievement can be used to establish baseline equivalence for a social studies achievement outcome. However, for outcomes in the English language proficiency domain, the pre-intervention measure must also be an eligible measure of English language proficiency.

Acceptable pre-intervention measures for student progress in school outcomes. For outcomes in the staying in school, progression in school, and completing school domains, studies must show that groups are equivalent on the following set of characteristics that are correlated with student progression.

- Grade level, measured at baseline; **AND**
- One of the following measures of student performance **or** social-emotional learning/behavior (only one measure must satisfy baseline equivalence, even if the study reports on more than one):
  - Standardized test scores (if the study reports standardized test scores, the review team will use them to assess baseline equivalence, rather than another measure from this list),
  - Whether behind in grade level (could be measured by age among students in the same grade),
  - Any eligible measure in a social-emotional learning/behavior domain (such as prevalence of school behavior or discipline issues, or rate of school attendance), **or**
  - Grade point average (GPA); **AND**

- One of the following (only one measure must satisfy baseline equivalence, even if the study reports on more than one):
  - Student race/ethnicity, **or**
  - A measure of degree of disadvantage (for example, free or reduced-price lunch status, poverty status, family income, English learner status, special education status, or disability status); **AND**,
- **If the unit of assignment is the school,** a school-level measure of the student progression outcome.

Acceptable pre-intervention measures for social-emotional learning/behavior domains. For outcomes in the four social-emotional learning/behavior domains, studies must show that the groups are equivalent on an acceptable pre-intervention measure in the same domain. A pre-intervention measure of the outcome is preferred; however, if a pre-intervention measure of the outcome is not available, any eligible pre-intervention measure in the same domain would be acceptable. In addition, an eligible student achievement pretest measure together with a pre-intervention measure from another social-emotional learning/behavior domain can be used to establish baseline equivalence (for example, baseline equivalence for the BASC outcome in the social-emotional domain can be established using pre-intervention standardized math scores and student attendance from the school engagement domain).

Acceptable pre-intervention measures for instructional practice and teacher attendance. For outcomes in the instructional practice and teacher attendance domains, studies must show that the groups are equivalent on a pre-intervention measure of the outcome or a related measure in the same domain. Because no teacher or school characteristics have been demonstrated to be highly related to teacher performance outcomes or attendance, measures of teacher or school characteristics are not acceptable pre-intervention measures of teacher performance or teacher attendance.

Acceptable pre-intervention measures for teacher retention outcomes. For outcomes in the four teacher retention domains, studies must show that groups are equivalent on the following set of characteristics that are correlated with teacher retention.

- Average years of teacher experience or the experience categories used in the study (for example, a "novice teachers" experience category, unless all of the teachers in the study fall into that category). Equivalence must be satisfied at the teacher level. **AND**
- One of the following measures of performance **or** social-emotional learning/behavior of the teachers' students (only one measure must satisfy baseline equivalence, even if the study reports on more than one):
  - Standardized test scores (if the study reports standardized test scores, the review team will use them to assess baseline equivalence, rather than another measure from this list),
  - Whether behind in grade level (could be measured by age among students in the same grade),

- Any eligible measure in a social-emotional learning/behavior domain (such as prevalence of school behavior or discipline issues, or rate of school attendance), **or**
- GPA.

Equivalence may be satisfied at the student or teacher level. **AND**

- One of the following measures of the characteristics of the teachers' students (only one measure must satisfy baseline equivalence, even if the study reports on more than one):
  - Student race/ethnicity **or**
  - A measure of degree of disadvantage (for example, free or reduced-price lunch status, poverty status, family income, English learner status, special education status, or disability status).

  Equivalence may be satisfied at the student, teacher, or school level. **AND**

- **If the unit of assignment is the school,** a school-level measure of the outcome. For example, if the outcome is teacher retention in the profession and the unit of assignment is the school, equivalence must also be demonstrated on a baseline measure of the percentage of teachers in the school who returned to the teaching profession.

For all measures used to establish baseline equivalence for a teacher retention outcome, except years of teacher experience, equivalence must be satisfied for the base period used in defining retention. For example, if the outcome is teacher retention from the first year of teaching into what would be the third year of teaching, the study must show that the students taught by intervention and comparison teachers during their first year of teaching were equivalent on academic performance.

Exhibit 4. Acceptable Pre-Intervention Measures by Outcome Domain

Outcome domain	Acceptable pre-intervention measures
• General literacy achievement • General mathematics achievement • General science achievement • General social studies achievement • General achievement • English language proficiency	Achievement in the same subject as the outcome OR Achievement in the general achievement domain (except English language proficiency) OR If the outcome is science achievement: general mathematics **or** general literacy achievement OR If the outcome is social studies achievement: general literacy achievement

Outcome domain	Acceptable pre-intervention measures
• Staying in school • Progression in school • Completing school	Grade level AND Student performance **or** social-emotional learning/behavior[a] AND Student race/ethnicity **or** degree of disadvantage[a] **AND, if the unit of assignment is the school**, a school-level measure of the student progression outcome
• Student social interaction • Observed individual behavior • Student emotional status • Student engagement in school	The same measure as the outcome OR Another measure from the same domain as the outcome OR An eligible measure of student achievement **and** a measure from another social-emotional learning/behavior domain
• Instructional practice • Teacher attendance	The same measure as the outcome OR Another measure from the same domain as the outcome
• Teacher retention at the school • Teacher retention in the school district • Teacher retention in the state • Teacher retention in the profession	Average years of teaching experience **or** the experience categories used in the study AND Student performance **or** social-emotional learning/behavior[a] AND Student race/ethnicity **or** degree of disadvantage[a] **AND, if the unit of assignment is the school,** a school-level measure of the teacher retention outcome

[a] See pages 15–17 for examples of acceptable measures of student performance, social-emotional learning/behavior, and degree of disadvantage; requirements for the number of these measures that must satisfy baseline equivalence; and requirements for the timing and level of measurement.

2. Baseline equivalence of clusters

Assessing equivalence of clusters

In general, considerations for satisfying baseline equivalence of individuals also apply to satisfying baseline equivalence of clusters. In particular, baseline equivalence of clusters in the intervention and comparison groups must be satisfied by the same baseline measures listed above

for assessing baseline equivalence of individuals, and the same statistical adjustment requirements apply.

Acceptable samples for demonstrating baseline equivalence of clusters

For this review, any of the following three samples can be used to satisfy the baseline equivalence requirement for the analytic sample of clusters (provided the data are representative of the individuals who were in the clusters at the time the baseline data were collected).

a. The analytic sample of individuals from any pre-intervention time period.

b. Individuals from the same cohort as the individuals in the analytic sample, within the same clusters. The baseline data may be obtained at the time that clusters were assigned to conditions or during the year prior to when clusters were assigned to conditions.

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

If authors provide baseline information at multiple time periods, a reviewer should assess baseline equivalence using the information collected at the latest period before the start of the intervention. If authors provide baseline information for multiple samples, a reviewer should assess baseline equivalence using the sample listed first in the list above—that is, (a) should be used if available, then (b), and then (c). If authors provide baseline information for multiple samples across multiple time periods, the reviewer should consult review team leadership to determine which information to prioritize.

When a study examines the effectiveness of an intervention in multiple time periods, the sample used to satisfy baseline equivalence of clusters in the base period (for example, the school year after random assignment) also satisfies baseline equivalence of clusters in the later time periods (for example, 2 years after random assignment), so long as the outcome data are representative of the individuals in the clusters.

Outcome Measure Requirements

In this review, the validity requirements for measures in the teacher instructional practice domain are more stringent than those specified in the *WWC Standards Handbook* (in Section IV.A: Outcome Requirements and Reporting). For teacher instruction outcomes to meet the validity requirement for this topic area, a statistical relationship must be evident between the outcome and student achievement, progression in school, or social-emotional learning/behavior. The evidence of the statistical relationship may come from another study using the same teacher instruction or leadership practice outcome.

Statistical Adjustments

The *WWC Procedures Handbook* discusses the types of adjustments made by the WWC in Section VI: Reporting on Findings. For "mismatched" analysis (that is, when a study assigns units at the cluster level but conducts analysis at the individual level), this topic area uses the WWC default intra-class correlation coefficients of 0.20 for all student achievement outcomes

(including eligible outcomes in the five student achievement domains) and 0.10 for all behavior outcomes (including eligible outcomes in all other domains), unless a study-reported intra-class correlation coefficient is available.

Eligible Study Designs

Studies that use group designs (RCTs and QEDs), RDDs, or single-case designs (SCDs) are eligible for review using the appropriate standards or pilot standards.

PROCEDURES FOR CONDUCTING THE LITERATURE SEARCH

The *WWC Procedures Handbook,* discusses the procedures for conducting a literature search in Section III: Identifying Relevant Literature and Appendix B: Policies for Searching Studies for Review. This review will use a quick literature search process to identify research on a limited number of interventions that may be of most interest to decision makers, rather than using a broad keyword search on the full topic area to identify interventions. In the first step of this process, content experts identify and recommend interventions with a large body of causal evidence likely to be of interest to decision makers. This review will identify additional interventions that may be the focus of WWC-reviewed studies that are not already the subject of up-to-date WWC intervention reports.

After identifying these interventions, the second step of the process is to conduct intervention-specific literature searches, using the intervention name, to identify all publications on each intervention. This review may refine the potential scope of this search by including additional search terms. For example, for searches likely to produce a large number of results (such as for performance bonus programs), this review narrowed the search by including terms describing eligible study designs.

In a third step, each citation gathered through this search process undergoes a screening process to determine whether the study meets the eligibility criteria established in the review protocol. This screening process is described in Chapter IV of the *WWC Procedures Handbook.* Finally, the interventions are prioritized for review based on the quantity and quality of eligible studies of the intervention. This prioritization process is described in Appendix A of the *WWC Procedures Handbook.*

Additional Sources

Literature reviews for this topic area involve searching the websites and electronic databases listed in Appendix B of the *WWC Procedures Handbook.* In addition to those listed, this review searched the following electronic database:

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

In addition to those listed in the *WWC Procedures Handbook*, Appendix B, this review searched the following websites:

- American Association of Colleges for Teacher Education
- Bill & Melinda Gates Foundation
- Center for Teaching Quality
- Center on Great Teachers and Leaders
- Consortium for Policy Research in Education
- Education Development Center
- Institute of Education Sciences
- National Association for Alternative Certification
- National Center for Analysis of Longitudinal Data in Education Research (CALDER)
- National Center on Performance Incentives
- National Council on Teacher Quality
- RTI International
- University of Chicago Consortium on Chicago School Research
- Westat
- WestEd

APPENDIX A. EVIDENCE STANDARDS FOR SUPPLEMENTAL TEACHER OUTCOMES

Some teacher outcomes that will not be reported on as main findings in intervention reports under this protocol may be of interest to policy makers. These additional outcomes could include content knowledge, pedagogical content knowledge, summative accountability measures, and other measures used to make high-stakes decisions. Findings for these outcomes are eligible as supplemental findings that may be included in appendices to the intervention report if they meet WWC design standards.

Findings for these supplemental teacher outcomes will be reviewed according to the same evidence standards described above for outcomes within the eligible domains, including the rules for establishing baseline equivalence on pages 14–15. Those rules should be applied to the following acceptable baseline equivalence measures for supplemental teacher outcomes.

Acceptable pre-intervention measures for supplemental teacher outcomes. For supplemental teacher outcomes, studies must show that the groups are equivalent on a pre-intervention measure of the outcome or a highly related measure. The review team leadership has the discretion to determine whether a measure is sufficiently related to the outcome of interest for the measure to be used to establish baseline equivalence.