WWC EVIDENCE REVIEW PROTOCOL FOR TEACHER TRAINING, EVALUATION, AND COMPENSATION INTERVENTIONS
VERSION 3.0[1] (MARCH 2014)

Topic Area Focus

What Works Clearinghouse (WWC) reviews in the Teacher Training, Evaluation, and Compensation topic area focus on interventions aimed at making teachers more effective at improving the academic achievement of students in grades PK–12. Interventions may include teacher preparation programs, teacher induction programs, teacher professional development programs, teacher evaluation systems, and teacher compensation systems. The populations on which these interventions focus are adults who are considering teaching, undergoing teacher preparation, or are already employed in the teaching profession. The review will focus on student achievement and those teacher-level outcomes that have been shown to be related to student achievement (e.g., teacher retention).

Systematic reviews of evidence in this topic area address the following questions:

- Which teacher-focused interventions improve the academic achievement of students in grades PK–12?

- Which teacher-focused interventions improve those teacher outcomes that are associated with student achievement?

Key Definitions

In-service training. Training experiences provided to individuals while they are employed as teachers of record by a school or district.

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

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 certification. Individuals are no longer teacher candidates when they complete their training program and become fully certified as a teacher.

Pre-service training. Training of prospective teachers as part of obtaining teacher certification. Generally, this occurs before being hired by a school/district and becoming responsible for a

[1] This protocol is aligned with the *WWC Procedures and Standards Handbook* (version 3.0).

classroom or becoming the "teacher of record." Pre-service training can involve coursework completed toward certification, as well as student-teaching or other practice-based experiences. In cases where teachers within a study are described in more detail, the decision rules for classifying people as teachers are:

Teachers

- Those who provide students with more than 50% of instruction on a subject (i.e., teachers of record)
- Individuals providing instruction to students, 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 (i.e., fill in for a particular teacher for more than half the period between pretest and posttest)

Non-teachers

- Those who provide instruction outside of school hours (i.e., tutors)
- Those whose primary role is administrative (e.g., principal, dean, superintendent)
- Those providing instruction to individuals outside of grades PK–12 (e.g., college lecturers)
- Those who provide non-instructional support to students (e.g., nurse, school psychologist, speech language pathologists)
- Short-term substitutes (i.e., spend less than half of the period between pretest and posttest filling in for a particular teacher)

Questions regarding whether a particular study or group within a study involves teachers or non-teachers should be directed to the Teacher Training, Evaluation, and Compensation topic area review leadership team.

Populations to be Included

In this review, the following populations are of interest:

- Teachers who are studied must be employed by schools located within the United States, its territories, or tribal entities.
- Studies involving teachers of any and all student subgroups are eligible for review.

Subgroups of interest. An intervention's effectiveness could vary by subgroups defined by characteristics of teachers or students in the population or by characteristics of the interventions' settings. In studies that present analyses of the subgroups specified below, the subgroup analyses may also be eligible for review in the Teacher Training, Evaluation, and Compensation topic area. We will present findings for subgroups of interest in an appendix, provided the subgroup analyses meet topic area criteria and design standards with or without reservations. Potential subgroups of interest for this review include:

- Characteristics of teachers
    - Experience
    - Certification status
    - Demographic characteristics
- Characteristics of students
    - Special education status
    - English learner (EL) status
    - Grade levels
- Characteristics of interventions' settings
    - Location of teaching setting (e.g., urban, suburban, or rural)
    - School type (elementary, middle, high)
    - School size

Types of Interventions to be Included

This review will include practices, programs, and policies (referred to as *interventions* throughout the protocol) that fall within the following categories:

Categories of Teacher Training, Evaluation, and Compensation-related Research

- ***Teacher Preparation:*** Studies of programs that train individuals to enter the teaching profession, including both traditional programs based in college or university schools of education and alternative route programs.

- ***Teacher Induction:*** Studies that examine programs that provide novice teachers with direction and support during the first 3 years of their careers.

- ***Teacher Evaluation:*** Studies that examine processes or systems used to determine teacher effectiveness in the classroom.

- ***Teacher Compensation:*** Studies of programs or systems that provide incentives to teachers for taking particular teaching positions or for demonstrating effectiveness in improving academic achievement. Such systems might include salary structure, bonuses, or pensions.

- ***Teacher Professional Development:*** Studies of programs that provide in-service training to presently employed teachers.

Types of Interventions

Interventions in this topic area can take the form of practices, programs, or policies which are designed to provide teachers with new knowledge, skills, or motivation to improve student achievement.

Practices. The review will include both general and targeted practices. A *general practice* is a specific approach to interacting with students or materials in classrooms. A *targeted practice* is a named approach to promote the development of a subset of students in the classroom (e.g., English language learners). Both general and targeted practices must be clearly described and commonly understood in the field and literature. An example of a teacher-quality related practice is professional development to support instructional scaffolding, which is a 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. Practices are typically actions that teachers take as they plan, implement, or evaluate instruction.

Policies. A *policy* is a named condition, system, or set of formal rules that affect 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-quality related policies 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 effectiveness of teachers. Well-defined teacher preparation programs and well-defined professional development sequences within a district are examples of programs.

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 pre-service training;
  - Provides in-service training;
  - Provides technical assistance (e.g., instructions/guidance on the implementation of the intervention); or
  - Sells or distributes the intervention.
- Replicability: packaged or otherwise available for distribution/use beyond a single site
- Trademark or copyright

Interventions that are implemented through professional development with teachers but focus on a single academic subject and include a curriculum will not be covered under this protocol. Such interventions may be more appropriately reviewed under the protocol for that academic subject (e.g., adolescent literacy, math).

Elements of Intervention Replicability

All reviewed policies, practices, and programs must be replicable (i.e., can be implemented by those other than the developers of the approach). 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 targeted, approach to enhancing the skill(s) (e.g., strategies, activities, and materials), unit of delivery of the intervention (e.g., whole group, individual), medium/media of delivery, and targeted population
- Intervention duration and intensity
- Description of individuals delivering or administering the intervention

Types of Research Studies to be Included

In this review, the following additional parameters define the types of research studies to be included:

- ***Topic relevance.*** Studies will be relevant if they meet two criteria:

    1. They must examine a program, practice, or policy on which the review focuses; and

    2. They must examine the effect of an intervention on student academic outcomes or teacher behaviors that are linked to student academic outcomes.

- ***Time frame relevance.*** To be included in this topic area, studies generally must have been released or made public in 1993 or later.

- ***Sample relevance.*** The sample must include teachers who provide instruction to students in grades PK through 12.

- ***Language relevance.*** The study should be written in English. Studies can look at impacts on teachers providing foreign language instruction, so long as the study is conducted in the focal geographic location and at least 50% of students are native English speakers.

- ***Study location relevance.*** Studies must be conducted with teachers working in the United States, its territories, or tribal entities.

- ***Study design relevance.*** Studies must be empirical, using quantitative methods and inferential statistical analysis, and must take the form of a randomized controlled trial (RCT) or use a regression-discontinuity design (RD), a quasi-experimental design (QED), or a single-case experimental design (SCD).

- ***Outcome relevance.*** The study can focus on student achievement or teacher outcomes that have been shown to be related to student achievement. Studies must include at least one outcome measure in at least one of the following domains: student general academic achievement; student academic achievement in mathematics, reading, science, or social studies; student progression; teacher attendance; quality of teacher instruction; or teacher turnover. These outcomes are described in more detail in the next section.

Outcome Measures

Student-level Outcome Domains

Mathematics achievement. Includes outcomes in the following areas: understanding of different subjects within mathematics, including algebra, calculus, geometry, trigonometry; understanding

of concepts and procedures; understanding of word problems and applications; and general math achievement (i.e., a standardized test covering a full array of mathematics topics).

Reading achievement. Includes outcomes in the following areas: foundational reading (word reading, fluency and/or accuracy in reading connected text, vocabulary, reading comprehension), English/language arts, and general reading.

Science achievement. Includes outcomes in any of the physical or life science disciplines, such as biology, chemistry, general science, and physics.

Social studies achievement. Includes outcomes in social studies subdisciplines, such as civics, geography, history, and world cultures.

General student achievement. Includes a general measure of student achievement, only to be documented if study authors do not distinguish students' academic achievement in specific areas (e.g., math, reading). Examples include composite scores from state assessments that represent a combination of reading and math scores.

Course grades, teacher reports of proficiency in the different subject areas, and assessments unrelated to academic achievement are not eligible outcome measures.

Student progression. Includes measures of progression. Specific constructs within progression are *student promotion* (e.g., students' advancement to next grade level) and *graduation* (e.g., students' completion of the PK–12 education system).

Teacher-level Outcome Domains

Teacher instruction. Includes outcomes that reflect the quality of teachers' instruction. As discussed below, only outcomes that are shown in that study or other studies to be correlated with improved student performance will be deemed eligible.

Examples of eligible assessments of the quality of teacher outcomes include:[2]

- Charlotte Danielson's Framework for Teaching (FFT)
- Classroom Assessment Scoring System (CLASS)
- Protocol for Language Arts Teaching Instruction (PLATO)
- Mathematical Quality of Instruction (MQI, predicting mathematics achievement)
- UTeach Teacher Observation Protocol (UTOP)

[2] All of these measures have been validated through the MET study (Kane & Staiger, 2012).

Teacher attendance. Includes outcomes that indicate the number (percent) of eligible work days for which the teacher is present.

Teacher turnover at the school. Includes outcomes that reflect whether teachers keep their teaching positions within the same school from year to year. Although teachers' decisions to give up their positions may reflect their dissatisfaction with their current employment, different types of turnover can reflect different types of motivation. These turnover types will be considered separate constructs within the turnover domain.

Constructs include:

- ***Attrition from teaching.*** Occurs if teachers decide to leave the teaching profession altogether. Their departure from the profession may be voluntary (i.e., retirement, resignation to raise a family) or involuntary (i.e., getting fired/counseled out of teaching).
- ***Mobility or "lateral mobility."*** Occurs when teachers voluntarily leave a teaching position in one school to take a similar position in another school.
- ***Staff reductions.*** Occurs when some teachers lose their positions involuntarily when economic conditions necessitate staffing cuts as a cost-savings measure. Such turnover may not reflect teachers' dissatisfaction with his/her teaching position, and so represents an additional type of turnover that we may want to distinguish at a later date.
- ***Change to other positions in education or "vertical mobility."*** Occurs when teachers decide to take administrative positions (e.g., principal, curriculum director).

In situations when study authors present findings for overall turnover and findings for specific types of turnover (i.e., the constructs listed above), we will document in the Data tab of the SRG both the overall domain finding (turnover in the aggregate) and the findings for each construct. The intervention report will focus on the aggregate measure, but we may include findings for different types of turnover in an appendix of the report.

Value-added scores. Conceptually, a value-added score represents the average deviation of the students' achievement test scores of a given teacher or school from their predicted scores. Predicted scores can be based on average student growth (i.e., the average learning trajectory of students at each grade level) or the students' scores at baseline and then adjusted for other factors. Computationally, value-added scores most often are created using regression models or hierarchical/multi-level models that include students' prior academic performance and other variables that may be related to achievement but are outside of the teacher's or school's control (e.g., student eligibility for free or reduced-price lunch).

- Value-added measures are held to the same reliability standards as other measures. It is the authors' responsibility to report the reliability estimates for their value-added measure.

- Studies of school-level interventions may analyze schoolwide value-added scores. These value-added scores are also held to the same reliability standards as other outcomes.

Overalignment of outcome measures. A study's rating will be based only on those measures that are not overaligned. Overalignment occurs when outcome measures are more closely aligned to one of the research groups (intervention or comparison) than the other and could bias a study's results. For instance, if the outcome measure involves teachers' instructional practices as reflected in an observation rubric with which only intervention teachers have familiarity, and the lesson being observed was specifically used during the intervention, then this might be grounds for considering the outcome measure to be overaligned. In these situations, the intervention group may have an unfair advantage over the comparison group, and the effect size is not a fair indication of the intervention's effects.

For this topic area, teachers are the recipients of the intervention. Because teachers become the intermediary between the intervention and the student, it is highly unlikely that a student outcome could be judged to be overaligned.

Reliability and validity of outcome measures. Measures of the outcome of interest should demonstrate adequate reliability and validity. For both student outcomes and teacher outcomes, reliability (internal consistency, temporal stability/test-retest reliability, and inter-rater reliability) will be assessed using the following standards:

- Internal consistency: minimum of 0.60;
- Temporal stability/test-retest reliability: minimum of 0.40;
- Inter-rater reliability: minimum of 0.50 (percent agreement, correlation, Kappa).

If the reliability of each outcome measure is not specified in the research article, data from the test or scale publisher or other sources, including an author query, may be used to establish the reliability of an outcome measure. If reliability cannot be determined with the given information from authors or from the research literature, then the lead methodologist will determine if the outcome is eligible for review.

Validity standards will differ for student outcomes and teacher outcomes.

- ***Validity for student outcomes.*** Measures of student outcomes must have *face validity*. That is, reviewers should be able to judge for themselves whether a student-level outcome measure assesses the construct in question.

- ***Validity for teacher outcomes.*** For teacher outcomes to meet the validity standard for this topic area, a statistical relationship must be evident between the teacher outcome and student achievement. Such a relationship can be indicated with a correlation coefficient, regression models, *t*-tests, ANOVAs, or multilevel statistical models. The relationship must be documented in the report being examined, in a publicly available report cited by the authors, or through an author query. The onus for establishing the validity of a teacher

outcome measure rests with the authors of the report. The lead methodologist will have discretion regarding the validity of the measure.

The interval for measuring post-intervention effects. For the Teacher Training, Evaluation, and Compensation topic area, measures obtained at the end of an intervention, as well as any time thereafter, are admissible. The Teacher Training, Evaluation, and Compensation review team prioritizes immediate post-intervention findings for developing intervention ratings and improvement indices because these findings are most prevalent in Teacher Training, Evaluation, and Compensation studies. Measures occurring several months or years after the intervention may provide strong evidence for an intervention's effectiveness. Therefore, the Teacher Training, Evaluation, and Compensation review team will include follow-up findings, when available and appropriate, in supplemental appendices to the intervention report.

Statistical and Analytic Issues

Attrition in RCTs

The WWC considers both the overall sample attrition rate and the differential in sample attrition between the intervention and comparison groups, as both contribute to the potential bias of the estimated effect of an intervention. The WWC has established conservative and liberal standards for acceptable levels of attrition. The conservative standards are applied in cases where the lead methodologist has reason to believe that much of the attrition can be attributed to the intervention reviewed—for example, high school students choosing whether or not to participate in a dropout prevention program. The liberal standards are applied in cases where the lead methodologist has reason to believe that little of the attrition is endogenous to the intervention reviewed. Attrition rates are based on the number of sample cases used in the analysis sample with measured, as opposed to imputed, values of the outcome measures.

The Teacher Training, Evaluation, and Compensation topic area uses the liberal standard. This reflects the assumption that most attrition of teacher training, evaluation, and compensation research results from exogenous factors, such as teachers' absence on the day of observations, reductions in force among teachers (especially new teachers), or parent mobility and students' absence on the days that assessments are conducted.

Table 1 presents the maximum difference in the attrition rate for the intervention and comparison groups that is acceptable for a given level of overall sample attrition. The empirical basis for these thresholds is described in Appendix A of the *WWC Procedures and Standards Handbook*, version 3.0.

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

RCTs with combinations of overall and differential attrition rates that exceed the applicable threshold, based on the liberal standard, must demonstrate baseline equivalence of the analysis sample, or, if non-equivalence falls within the allowable range, statistically control for that nonequivalence, in order to receive the MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS rating. See the Baseline Equivalence section below for more details.

Table 1. WWC Attrition Standards for Randomized Controlled Trials

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

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

Baseline Equivalence

For the Teacher Training, Evaluation, and Compensation topic area, RCTs with high attrition or QED studies must show equivalence between intervention and comparison groups on the outcome measure(s) before the intervention. The onus for demonstrating equivalence in these studies rests with the authors. Sufficient reporting of pre-intervention data should be included in

the study report (or obtained from the study authors) to allow the review team to draw conclusions about the equivalence of the intervention and comparison groups.

Groups are considered equivalent if the reported differences in pre-intervention data are less than or equal to 5% of the pooled standard deviation in the sample, regardless of statistical significance. However, if differences are greater than 5% and less than or equal to 25% of the pooled standard deviation in the sample, the analysis must have controlled analytically for the pre-intervention outcome measure(s) on which the groups differ. If pre-intervention differences are greater than 0.25 for *any* of the outcomes in the same domain, the study does not meet standards. In addition, if there is evidence that the populations were drawn from very different settings (such as rural versus urban, or high-SES versus low-SES), the lead methodologist may decide that the environments are too dissimilar to provide an adequate comparison.

Some RCTs with high attrition or QED studies may not have data on the same outcome measure at baseline. Depending on features of the outcome and unit of assignment, equivalence using alternative baseline measures may be acceptable. There are ten situations for which baseline and acceptable alternative baseline measures are identified. Table 2 provides a summary of the situations.

Table 2. Acceptable Baseline Measures for Different Situations

Situation	Outcome	Level of Analysis	Level of Assignment	Preferred Baseline Measure	Acceptable Alternative Baseline Measures
1	Student academic achievement (*in any subject area domain*)	Students within Teachers	Teachers	Students' achievement in subject at baseline	Students' general academic achievement at baseline
2		Students within Teachers	Schools	Students' achievement at baseline	School-level academic achievement at baseline AND 1 measure of school-level disadvantage AND racial/ethnic composition of schools
3	Student progression (*e.g., promotion in grade; graduation*)	Students within Teachers	Teachers	N/A	Race/ethnicity AND 1 measure of disadvantage AND 1 measure of student academic performance
4		Students within Teachers	Schools	N/A	School-level progression AND [1 measure of disadvantage OR 1 measure of school academic achievement]
5	Teacher (*or school*) value-added measures	Students within Teachers	Teachers	Students' achievement at baseline	Students' general academic achievement at baseline
6		Students within Schools	Schools	Average student achievement at baseline	Students' general academic achievement at baseline

Situation	Outcome	Level of Analysis	Level of Assignment	Preferred Baseline Measure	Acceptable Alternative Baseline Measures
7	Teacher instruction (*e.g., CLASS; FFT*)	Teachers	Teachers	Same teacher outcome at baseline	NONE
8		Teachers	Schools	Same teacher outcome at baseline	NONE
9	Teacher attendance	Teachers	Teachers	Attendance at baseline	NONE
10		Teachers	Schools	Attendance at baseline	NONE
11	Teacher turnover (*e.g., teachers move to other school, teachers leave profession*)	Teachers	Teachers	N/A	Level of teacher experience AND school's academic achievement AND racial/ethnic composition of schools
12		Teachers	Schools	N/A	School-level turnover AND levels of teacher experience AND school academic achievement AND school racial/ethnic composition

Notes: See text for examples of acceptable measures of disadvantage. N/A = not applicable.

Situation 1: Acceptable baseline measures for student academic outcomes in specific subject areas. For subject-specific student academic achievement, when the particular subject-specific measure is not assessed at baseline, group equivalence can be determined using a *general measure* of academic achievement that is highly related to the particular outcome measure (proxy measure). For example, if a study is a QED and the outcome of interest is student achievement in science, a *general measure* of achievement (e.g., combined math and reading scores) can be used to assess baseline equivalence if a baseline measure in the specific subject (in this example, science) is not available.

Situation 2: Acceptable baseline measures for academic outcomes with school-level assignment. When a study compares student achievement in two groups of schools that differ by the presence of a teacher-focused intervention, it is preferable for a measure of student achievement in that domain to be assessed at baseline. In cases where no such pre-intervention measure exists, then equivalence of schools can be determined only by demonstrating equivalence on all three of these alternative measures:

- **School-level achievement**

- **At least one measure of school-level disadvantage, including:** free and reduced-price lunch status, poverty status, family income, being from a single-parent family, parent's education, immigrant or English learner (EL) status, special education or disability status, teen parent status

- **Racial/ethnic composition of the school:** percentage of students within the school who represent a racial/ethnic minority group. The study review guide will calculate the differences in this binary measure between schools.

If the groups of schools differ by more than 0.25 standard deviations on any of these pre-intervention measures, then the study is rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

Situation 3: Acceptable baseline measures for student progression. Because students' progression status (e.g., promotion in grade level; graduation) cannot be determined prior to a teacher-focused intervention, a combination of alternatives may be sufficient for determining whether two groups of teachers are equivalent. These alternatives are:

- **Grade level being taught**

- **Student race/ethnicity**

- **At least one measure of disadvantage:** students' free and reduced-price lunch status, EL status, and special education or disability status

- **At least one measure of academic performance:** standardized test scores, whether behind grade level (could be measured by age among students in the same grade), prevalence of school behavior or discipline issues, rate of school attendance, GPA

If each of these three meet the standards for baseline equivalence, then the study may be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If groups of teachers' students differ by more than 0.25 standard deviations on any of these pre-intervention measures, then the study is rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

Situation 4: Acceptable baseline measures for student progression with school-level assignment. To compare the student progression rate among two sets of schools, the two sets of schools must be equivalent at baseline on these alternative measures:

- **School-level progression from year before**

- **At least one of the following:** A measure of student disadvantage (students' free and reduced-price lunch status, EL status, and special education or disability status) and/or a measure of school academic performance

If each of these meet the standards for baseline equivalence, then the study may be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If groups of schools differ by more than 0.25 standard deviations on either of these pre-intervention measures, then the study is rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

Situations 5 and 6: Acceptable baseline measures for teacher-level or school-level value-added scores. In situations where RCTs with high attrtion or QED studies compare value-added scores for different groups of teachers or for different groups of schools, baseline equivalence of groups can be determined using any of the following: (1) average student achievement in the academic subject of interest; (2) group average achievement on a *general measure* of academic achievement, provided that the general measure is highly related to the academic subject of interest; or (3) average student growth in the subject of interest. For example, if a study is a QED, and the outcome of interest is teachers' value-added scores in mathematics, then the equivalence of groups can be determined using students' average mathematics achievement at baseline, students' average general achievement at baseline (presuming that the authors show a strong relationship between scores on the general measure and students' achievement in mathematics), or students' average growth rates in mathematics prior to implementation of the intervention. If the difference between study groups is less than or equal to 0.25 standard deviations, then the study may be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If the two groups differ by more than 0.25 standard deviations at baseline, the study will be rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

Situations 7–10: Acceptable baseline measures for teacher instruction and attendance. For RCTs with high attrition or QED studies, no alternative measures are acceptable for determining baseline equivalence. That is, there are no known teacher or school characteristics that have been demonstrated to be highly related with teacher observational outcomes or attendance. If RCTs with high attrition or QED studies demonstrate equivalence of groups on the outcome measure at baseline, the study may be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If the two groups differ by more than 0.25 standard deviations at baseline, they will be rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

Situations 11 and 12: Acceptable baseline measures for teacher turnover. There are no baseline measures of teachers' decisions to either move to another school or leave the teaching profession altogether. Instead, group equivalence must be shown on the following study characteristics:

- **Levels of teacher experience.** Groups being compared must have an equivalent percentage of teachers who have less than 3 years' experience teaching in the classroom (i.e., novice teachers) AND an equivalent percentage of teachers within the oldest age group (i.e., nearing retirement age).

- **Academic performance.** The groups being compared must have students with equivalent levels on all of the following that are presented: standardized test scores, whether behind grade level (could be measured by age among students in the same grade), prevalence of school behavior or discipline issues, rate of school attendance, and GPA. Should one or more of these academic-related characteristics not be listed, then equivalence is based on those characteristics that are present.

- **Racial/ethnic composition of the school.**

If each of these meet the standards for baseline equivalence, then the study may be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If groups of schools differ by more than 0.25 standard deviations on either of these pre-intervention measures, then the study is rated DOES NOT MEET WWC GROUP DESIGN STANDARDS.

RCTs with high attrition and QED studies that demonstrate equivalence of groups on each of those characteristics (teacher experience, academic performance, and racial/ethnic composition of the school) will be rated MEETS WWC GROUP DESIGN STANDARDS WITH RESERVATIONS. If the groups differ by more than 0.25 standard deviations at baseline on any of the characteristics, the study will be rated Does NOT MEET WWC GROUP DESIGN STANDARDS.

Given the potential for selection bias in QEDs and RCTs with high attrition, the possibility that the intervention and comparison groups were drawn from different populations is also a concern. Fundamental differences in the settings from which the intervention and comparison groups in a QED study were drawn, or baseline differences in the characteristics of the intervention and comparison groups in QEDs and RCTs with high attrition, may indicate that the students or teachers in the two groups represent very different populations, even if they are equivalent on pretest measures. When there is evidence that the populations being compared are drawn from very different settings, the study will be referred to the review team leadership, who will determine whether the settings are too dissimilar to provide an adequate comparison. If the leadership decides that they are too dissimilar, the study is rated DOES NOT MEET WWC GROUP DESIGN STANDARDS. These characteristics include, but are not limited to, the percentage of students from:

- **Families of different socio-economic status' or racial/ethnic groups;**
- **Special education classifications;** and
- **Different locations (e.g., urban, rural).**

The provision of all such information, however, is not a requirement of the review.

Other Statistical and Analytical Issues

RCT studies with low attrition do not need to use statistical controls in the analysis, although statistical adjustment for well-implemented RCTs is permissible and can help generate more precise effect size estimates. For RCTs with low attrition that do not include statistical controls for the pretest in the analysis of effects of the intervention, the effect size estimates will be adjusted for differences in pre-intervention characteristics at baseline (if available) using a difference-in-differences method if the authors did not adjust for pretest (see Appendix F of the *Handbook*). Authors may make additional statistical adjustments that are not required by the WWC evidence standards.

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

and adjusted post-intervention means, the WWC review will report the adjusted means and unadjusted standard deviations. If effect sizes or the information required to calculate them are not reported, then the missing information will be requested from the author(s).

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

When the unit of analysis in the study is not the same as the unit of assignment, the effect sizes computed by the WWC will incorporate a statistical adjustment for clustering. The default intraclass correlation used for academic achievement domains is 0.20 and 0.10 for the teacher outcomes domain. For an explanation about the clustering correction, see Appendix G of the *Handbook*.

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

The Teacher Training, Evaluation, and Compensation review is unlikely to review regression discontinuity (RD) or single-case design (SCD) studies. If, however, one is identified, the review will follow the pilot standards as described in the *WWC Procedures and Standards Handbook*. Decisions about parameters unique to RD or SCD studies will be determined by the review team in collaboration with content experts at that point; the protocol will then be updated.

LITERATURE SEARCH METHODOLOGY

The WWC literature search is comprehensive and systematic. Detailed protocols guide the entire literature search process. At the beginning of the process, relevant journals, organizations, and experts are identified. The WWC searches core sources and additional topic-specific sources identified by the lead content expert. The process is fully and publicly documented.

In 2013, the WWC staff members conducted a search of the literature related to the interventions of interest. For this search, WWC staff used a comprehensive list of search terms and strategies designed to identify literature pertaining to a broad range of programs and practices in the Teacher Training, Evaluation, and Compensation topic area. WWC staff used an extensive list of keywords to search electronic databases, conducted targeted searches of specific programs identified through the keyword search, and searched the websites of organizations that conduct research on teacher training, evaluation, and compensation programs.

1. Keyword Search Parameters and Databases

The primary objective of the keyword search was to identify interventions with potentially eligible studies and assess the number of studies on each intervention, so that interventions could be prioritized for review. Targeted outcomes and study design terms were included to focus the search on identifying literature that will support an intervention report.

Search Terms

The following table presents the search terms by category.

Category	Search Terms
Study Design	control(s), control group(s), comparison group(s), counterfactual(s), matched group(s), treatment*, random*, assignment, baseline, experiment*, evaluat*, impact*, effect*, efficacy, causal, growth, outcome*, posttest, post-test, pretest, pre-test, randomized controlled trial(s), random control trial, RCT, field trial, quasi-experiment*, quasiexperiment*, quasi experiment*, QED, regression discontinuity, RDD, changing criterion design, intrasubject replication design, multiple baseline design, multi-element design, multi element design, single case design, single subject design, SCD, ABAB design, alternating treatment, simultaneous treatment, meta-analysis, meta analysis, reversal design, withdrawal design.
Keywords	The initial literature search focused on four programs: *The Teacher Advancement Program, Teach For America, My Teaching Partner*, and *New Teacher Center Induction Model.* A full keyword search will be conducted at a later time.
Intervention Name	Examples: *The Teacher Advancement Program, Teach For America, Troops to Teachers.*

Electronic Databases

This review searched the following electronic databases:

- ***Academic Search Premier***
- ***Campbell Collaboration***
- ***EconLit***
- ***Education Research Complete***
- ***E-Journals***
- ***ERIC***
- ***Google***
- ***ProQuest Dissertation & Theses***
- ***PsycINFO***
- ***SAGE Journals Online***
- ***Scopus***
- ***SocIndex with Full-Text***
- ***World Cat***

Websites

This review searched the following websites:

- ***Abt Associates***
- ***Alliance for Excellent Education***
- ***American Education Research Association***
- ***American Enterprise Institute***
- ***American Institutes for Research***
- ***Best Evidence Encyclopedia***
- ***Bill & Melinda Gates Foundation***
- ***Brookings Institution***
- ***Carnegie Corporation for the Advancement of Teaching***
- ***Carnegie Corporation of New York***
- ***Center for Research and Reform in Education***
- ***Center for Teaching Quality***
- ***Center on Great Teachers and Leaders***
- ***Chapin Hall Center for Children at the University of Chicago***
- ***Congressional Research Service***
- ***Consortium for Policy Research in Education***
- ***Government Accountability Office***
- ***Grants/contracts awarded by IES***
- ***Harvard Family Research Project***

- ***Heritage Foundation***
- ***Hoover Institution***
- ***Mathematica Policy Research***
- ***MDRC***
- ***National Association of State Boards of Education***
- ***National Center for Longitudinal Analysis of Longitudinal Data in Education Research (CALDER)***
- ***National Center on Performance Incentives***
- ***National Council on Teacher Quality***
- ***National Governor's Association***
- ***NBER Working Papers***
- ***Policy Archive***
- ***Policy Studies Associates***
- ***RAND Corporation***
- ***Regional Education Laboratories (All 10)***
- ***SRI***
- ***Thomas B. Fordham Institute***
- ***University of Chicago Consortium on Chicago School Research***
- ***Urban Institute***
- ***Westat***
- ***WestEd***
- ***Intervention-specific websites (e.g., Teacher Advancement Program, Teach For America, Troops to Teachers)***