THE CAMPBELL COLLABORATION

# Protocol for a Systematic Review: Teach For America (TFA) for Improving Math, Language Arts, and Science Achievement of Primary and Secondary Students in the United States: A Systematic Review

**Herbert Turner, Robert Boruch, Mackson Ncube, Annette Turner**

Submitted to the Coordinating Group of:

| | |
|---|---|
| ☐ | Crime and Justice |
| ☒ | Education |
| | ☐ Disability |
| ☐ | International Development |
| | ☐ Nutrition |
| ☐ | Social Welfare |
| ☐ | Other: |

Plans to co-register:

| | | | |
|---|---|---|---|
| ☒ | No | | |
| ☐ | Yes | ☐ Cochrane | ☐ Other |
| ☐ | Maybe | | |

Date Submitted: 31 December 2014
Date Last Revision Submitted: 21 November 2015
Approval Date: 22 November 2015
Publication Date: 04 January 2016

## BACKGROUND

### *Description of the Condition*

Research shows a shortage of effective teachers in many rural and urban K-12 public schools serving the highest proportions of high-poverty students across the US (Clotfelter, Ladd & Vigdor, 2006; Peske & Haycock, 2006; Monk, 2007). This shortage has persisted for decades (Darling-Hammond, 1984; Ingersoll 2001; Ingersoll & Perda, 2010). In the past decade, alternative route teacher preparation programs aimed at addressing this shortage proliferated across the United States (Kane, Rockoff, & Staiger, 2007). Alternative route teacher preparation programs seek to increase the supply of teachers more rapidly than traditional teacher preparation programs (Hess, 2002; Raymond & Fletcher, 2002; Blazer, 2012). Although their requirements vary widely, most of these programs are shorter, less expensive, and more practically oriented than traditional teacher preparation programs (Blazer, 2102). These programs also vary widely in their selection criteria for teacher candidates, approach to training these candidates, notoriety among education stakeholders, and evidence of their effectiveness (Hess, 2002; Kaine, Rockoff, & Staiger, 2007; Constantine et al. 2009). Teach For America (TFA) is nationally recognized as an alternate route teacher preparation program that has sought to address the shortage of effective teachers specifically in high-poverty rural and urban schools across the US (Teach For America, 2010).

TFA stands out among its peer preparation programs for several reasons. TFA is the largest source of new teachers and is the largest recipient of philanthropic funding for teacher recruitment for K-12 education (Blazer, 2012; Mead, 2015). Since 1990, TFA has recruited, selected, trained, placed, and supported approximately 25,000 new public school teachers (corps members) in the highest-poverty school districts in rural and urban areas. As of 2010, TFA corps members, represented between 10-15% of all new hires in high-poverty schools in the 35 regions served by TFA. In the 2013-14 school year alone, 11,000 TFA corps members reached more than 750,000 students in high-poverty K-12 rural and urban schools.[1] TFA is also the most publicly visible and widely debated alternative route teacher preparation program as noted by the Conner P. Williams (2014) article entitled "Stop Scapegoating Teach for America."[2] Finally, TFA is the most evaluated program of its kind. There have been multiple quasi-experimental and experimental studies conducted on the effectiveness of TFA to improve student outcomes. However, this body of primary studies has not yet been

[1] Data from a randomized controlled trial conducted in 2004 of students in 17 study schools with TFA corps members were on average 66% African American and 25% Latino/Hispanic. Approximately 95% of students in these schools received free or reduced price lunches, and began the year, on average, at the 14th percentile compared to the national norm (Decker, Mayer, & Glazerman, 2004).

[2] http://www.thedailybeast.com/articles/2014/09/24/stop-scapegoating-teach-for-america.html.

systematically reviewed and meta-analyzed (Raegan Miller, personal communication, 16 May 2014).

### *Description of the Intervention*

TFA is a selective alternative route teacher preparation program that recruits college graduates (many from top colleges) and professionals to teach in low-income schools (Clark, Isenberg, Liu, Makowsky & Zukiewicz, 2015). Ensuring that corps members become effective teachers "who lead their students to significant academic achievement" is ensconced in TFA's mission of eliminating educational inequity in US public schools (Teach For America, 2010). To fulfill this mission, TFA developed a data-driven program model that includes (1) a rigorous selection process, (2) intensive pre-service training for selected corps members, (3) two years of ongoing professional development for corps members, and (4) programming that fosters alumni leadership after TFA corps members have completed their two-year commitment (Teach For America, 2010).

Rigorous Selection Process. TFA's selection process includes a writing activity, telephone interview, sample teaching lesson plan, group discussion, and an in-person interview. TFA selects potential corps members who demonstrate competency in areas such as academics, leadership, critical thinking, ability to influence and motivate others, organizational ability, respect for students and families in low-income communities, and perseverance (Teach For America, 2010). Selected corps members, receive a five-week intensive summer training and agree to teach in their assigned school for at least two years. Those that complete the two-year commitment become alumni, and continue to be a part of the TFA community with continued access to resources and support for alumni (see Programming for Alumni section below).

Intensive Pre-Service Training. The five-week intensive pre-service summer training covers (1) instructional and pedagogical philosophies and practices, (2) classroom management skills, (3) attitudes towards teaching, and (4) academic ability. These skills and attitudes are hypothesized to have a positive and meaningful effect on students' academic achievement. This effect is hypothesized to be larger than the effect the same students would have experienced had the TFA corps member not been placed in their classroom.

On-going Professional Development. TFA corps members continue to receive training and support throughout their two-year teaching commitment to help them further develop skills and attitudes introduced during the pre-service training. This ongoing professional development includes observation and coaching from program directors; access to online classroom resources, advice, community support, and self-directed online learning on a private secure website for corps members and alumni.

Programming for Alumni. At the end of their two-year assignment, TFA alumni are encouraged to continue to engage in meaningful ways to advance the mission of TFA and become influential education leaders and advocates for children. TFA alumni have access to

teaching resources as well as support of the TFA community as they continue their professional careers.

### *How the Intervention Might Work*

The hypothesized theory of change for TFA, distilled from the literature including TFA's 2010-15 Business Plan is as follows:

- TFA recruits and selects applicants using a selection model. The model is based on the organization's data on the relationship between TFA student-achievement and TFA corps member characteristics and how these characteristics are associated with implementing a Teaching as Leadership approach in the classroom.

- TFA trains selected individuals starting with a five-week, intensive summer institute training coupled with fieldwork opportunities, such as classroom observations and delivery of instruction prior to the initial teaching year. The training is guided by the Teaching as Leadership framework that codifies for corps members a goal-oriented approach to teaching in the classroom.

- On-going peer mentoring and professional development. This includes continuous feedback on how corps members' teaching is impacting the classroom, including the impact on student achievement.

- After completing their two-year teaching commitment, alumni who continue to teach are eligible to continue to receive access to TFA professional development and leadership resources.

- The TFA Theory of Change is summarized in Figure 1.

Figure 1. TFA Theory of Change


Source: Teach For America Business Plan 2010-15

### *Why it is Important to do the Review*

***No systematic reviews and meta-analysis on TFA**:* The effects of TFA corps members and alumni on student academic outcomes have been investigated by educational researchers and economists alike using correlation, quasi-experimental, and randomized controlled trial designs.[3] There have been vigorous debates about TFA's effectiveness based on these primary studies. When the results are compared by research design, the quasi-experimental studies send mixed signals but experimental findings have consistently found a positive and statistically significant effect in math, but not reading (Raymond, Fletcher, and Luque, 2001; Laczko-Kerr & Berliner, 2002; Seftor & Mayer, 2003; Decker, Glazerman, and Mayer, 2004; Clark et al., 2013). However, until a systematic review of these studies is conducted, we do not know the average effect of TFA across these experiments. Furthermore, this average effect may vary according to academic outcome, grade level, teacher experience, and teacher certification status, and this variation can only be investigated through a meta-analysis. Also, the methodological quality of the quasi-experimental studies (e.g., such as establishment of baseline equivalence between groups in the analysis sample) and experimental studies (e.g., high attrition disrupting random assignment) has not been systematically and rigorously evaluated using C2 systematic review methods. By systematically reviewing the primary studies on TFA, we can apply methods designed to limit the bias in the retrieval, appraisal, and statistical synthesis of the primary study findings (Petticrew and Roberts, 2006; Cooper, 2010).

Using C2 systematic review and meta-analysis methods, we can empirically investigate whether effect sizes reported in primary studies are consistent, and can be generalized across populations and settings. We can also investigate whether findings vary by subsets of primary studies. The use of meta-analysis, after primary studies have been systematically reviewed, will allow us to statistically synthesize study findings, potentially increase the power and precision of effect sizes reported in primary studies, and potentially enhance their generalization (Chalmers and Altman, 1995; Petticrew and Roberts, 2006; Cooper, 2010).

***Narrative reviews are not a substitute for systematic reviews**:* The narrative reviews of quasi-experimental studies and experimental studies that have been conducted on the effects of TFA on K-12 students' academic outcomes are helpful to gain an approximate idea of the amount of agreement or disagreement of treatment effects across studies. They also help us understand what treatment effects look like across different samples. However, the primary limitation of a narrative review is the lack of coding of study characteristics and effect sizes and a statistical synthesis of these effect sizes. Without this, it is difficult, if not impossible for narrative reviews to cognitively and systematically manage and control for the many sources of variation in primary study characteristics and effect sizes (Chalmers and

[3] Value-added designs have also been used, but they are beyond the methodological scope of a C2 systematic review.

Altman, 1995; Petticrew and Roberts, 2006; Cooper, 2010). These variations arise from the different time periods, within study sampling, sample characteristics, group comparisons, outcomes measures, and designs. Reporting of such studies, if not handled systematically, could produce the appearance of conflicting results, or produce consistent results without empirical information across studies to understand why. In contrast, a systematic review transparently and systematically combs through the evidence, controls for study quality, and, when appropriate, statistically synthesizes the results with a view to presenting findings with greater clarity, and less potential bias, than narrative (or literature) reviews (Chalmers and Altman, 1995; Petticrew and Roberts, 2006; Cooper, 2010).

***Reliable and valid systematic evidence to address future scale up of TFA***: With the continued shortage of effective teachers in high-poverty rural and urban schools, it is reasonable to predict that the demand for alternative route teacher preparation programs, like TFA, will increase, not decrease. TFA used the i3 scale up funds to more than double its corps members from 7,300 to 15,000 teachers and increase its presence from 46 to 60 urban and rural regions across the country. By the end of 2015, TFA teachers would reach nearly one million students in some of our country's highest-need communities. There were enough internally valid randomized control trials and matched comparison studies with substantially positive and statistically significant findings to motivate the US Department of Education's Office of Innovation and Improvement (i3) to award TFA in the fall 2010 a 50 million dollar grant to scale up nationally at the elementary, middle, and high school levels. However, the randomized controlled trials and matched comparison studies were presented in a narrative review, by an independent evaluator, to make the case for the i3 scale up funding. The i3 award and narrative review is not a substitute for using C2 systematic review methods to objectively review the quality of randomized controlled trials and matched comparison studies and synthesizing the effect sizes to estimate the average effect of TFA that report the effects of TFA on student academic outcomes

***A TFA systematic review and meta-analysis, at this time, as an important benchmark***: The Investing in Innovation Fund (i3) scale up impact evaluation of TFA is presently being conducted. The effect of TFA on student academic outcomes will be evaluated with an RCT at the elementary grades, and matched comparison QEDs at the middle and high school grades because of the challenges in randomly assigning students beyond elementary grades. The findings from these studies will be released by the US Department of Education in spring 2016. Appropriately synthesized effect sizes in a C2 systematic review prior to TFA scale up could serve as the "maintenance" benchmark for comparative purposes when the i3 scale up results are released, and allow us to evaluate if the effects are maintained or changed during and at the end of TFA scaling up. The comparison of pre-scale up average effect sizes (from a C2 systematic review) to scale up effect sizes (from the i3 evaluation) can make an important contribution to knowledge because one of the primary challenges associated with scaling up interventions is maintaining the effectiveness of the intervention as the intervention goes to scale (Klingner, Boardman, and McMaster, 2013). The systematic review will also create an empirical

database of TFA effect sizes (and corresponding study characteristics) that can be used to summarize the empirical landscape of the highest quality research on TFA, based on C2 systematic review standards, prior to release of the scale up findings. In addition, the systematic review can be used to compare effect sizes from the scale up study (across treatments, samples and settings) to average effect sizes (across studies, treatments, samples, and settings) from the systematic review .

## OBJECTIVES

The purpose of this review is to use systematic procedures that limit bias in the retrieval, critical appraisal, synthesis, and reporting of quasi-experimental and experimental studies that examine the effects of TFA on K-12 student academic outcomes in Math, English Language Arts and Science as reported in the peer reviewed literature and grey literature during the past 20 years. To aid education policymakers and stakeholders (including researchers) in using the review results, we organize the questions according to the policy relevance and methodological issues raised in our review of the literature on the effectiveness of TFA:

1. What are the study characteristics of RCTs and QEDs conducted on TFA that met our inclusion criteria and were reported in this systematic review?

2. What are the sample characteristics of the schools, teachers, and students on RCTs and QEDs that met our inclusion criteria and were reported in this systematic review?

3. What are the main effects of TFA corps members on elementary school students in Math, ELA, or Science outcomes by research design?

4. What are the main effects of TFA corps members on middle school students in Math, ELA, or Science outcomes by research design?

5. What are the main effects of TFA corps members on high school students in Math, ELA, or Science outcomes by research design?

6. Are the main effects estimated by research design similar enough to be combined? If so, what is the combined main effect of TFA at each grade level and corresponding outcome?

7. How does the magnitude and statistical significance of the main effect of TFA change when controlling for the following teacher characteristics separately, in a moderator analyses:

    a. TFA candidate status (e.g., corps member or alumnus)

b. Teacher certification status (e.g., traditionally certified, alternatively certified, and not certified)

c. Teacher average years of teaching experience

8. To what extent does the main effect of TFA differ by fidelity of TFA implementation? (If there is sufficient fidelity of implementation information reported in TFA studies.)

9. Is there sufficient information on teacher turnover in TFA studies to evaluate TFA's main effect on teacher retention? If so, what is the main effect on teacher retention?

10. Is there sufficient information on teacher leadership, content knowledge, years of teaching experience, or overall academic ability to evaluate TFA's main effect on teacher quality? If so, what is the main effect on teacher quality?

11. Is there sufficient cost information in TFA studies, or a subset of studies, to evaluate whether TFA is reported as cost effective? If so, is TFA reported to be cost effective?

The answers to the first and second questions will provide education policymakers and stakeholders with a systematic profile that presents the author(s), date of publication (or reporting), sample characteristics, TFA and comparison groups, design, and outcomes. This profile will help education policymakers understand the potential sources of variability in TFA studies and how the reporting of such studies, if not handled systematically, could produce the appearance of conflicting results.

The answers to the third, fourth, and fifth questions are on the main effect of TFA and are considered confirmatory. The main effect is defined as the effect of TFA corps members on a particular student academic outcome, such as math, relative to non-TFA teachers without controlling for certification status and years of experiences (these controls are implemented when addressing question seven). We focus on TFA corps members because of TFA debates and the popular press focuses more on corps members and less on the alumni. This is partly because a smaller percentage of corps members transition to alumni status and continue to teach for five years (Noell & Gansle, 2009).

The answer to the sixth question addresses the methodological issue of whether to combine results of RCTs and QEDs. This decision should be based on the methodological quality of the studies, and how similar the average effect sizes are. For studies that passed the methodological quality screening and are included in the meta-analysis, we will evaluate whether the effect size differences between the RCTs and QEDs exceed .05 standard deviation units. We focus on the magnitude of the difference. The total number of RCTs and QEDs used to test for this difference may result in low statistical power. For this reason, we define a "substantial" difference, in the weighted average effect for RCTs and QEDs, using the WWC baseline equivalence standards (WWC Handbook Version 3.0). If the difference between the average effect sizes for the RCTs and QEDs exceeds .05 standard deviations, we

will not combine the RCTs and QEDs into a single meta-analysis to produce an overall, weighted average effect size across the two design types.

The answer to the seventh question is based on an exploratory analysis but will provide evidence to inform future research on TFA that speaks directly to the debates between TFA critics and TFA proponents. This will be accomplished by estimating whether the main effects of TFA are moderated by TFA status (corps members and alumni), certification status, or years of teaching experience, through a series of ANOVA analysis for categorical moderators and bivariate meta-regression analyses for the continuous moderator. Similarly, the answer to the eighth question examines whether the main effect of TFA differs by level of fidelity of TFA implementation as reported in the primary study.

Teacher turnover in TFA is an important issue in TFA studies and if teacher turnover is reported as an outcome in both the TFA and comparison groups in primary studies. The answer to the ninth question will address whether TFA has a main effect on this outcome.

The answer to the eleventh question is descriptive, relies on what the research reports, and is designed to provide contextual information for study finding by reporting what authors found regarding the cost effectiveness of TFA.

## METHODOLOGY

### *Criteria for including and excluding studies*

*Types of study designs*

This review will include primary studies with research designs that, when implemented well, are capable of generating data that can be used to make generalizable causal inferences about the effects of TFA on student academic outcomes. Eligible designs that meet these criteria are randomized controlled trials (RCTs), regression discontinuity designs (RDDs), single-case designs (SCDs), and quasi-experimental designs (QEDs). However, we limit the eligible designs for this review to RCTs, where random assignment is used to form intervention and comparison groups, and QEDs, where non-random methods such as matching or other statistical methods are used to form a counterfactual group that is comparable to the intervention group on measured characteristics. RDDs and SCDs will be excluded from this review since statistical methods for incorporating RDD and SCD data into meta-analyses are, to the best of our knowledge, not well established. For example, the Campbell Collaboration Methods Policy Brief is silent on the statistical synthesis of the RDD and SCD. Furthermore, the nature of TFA interventions and the results of our cursory literature search indicate that RDDs and SCDs are rare. Research designs that lack a comparison group, such as single-group "pretest/posttest" designs, will be excluded from the review. It is well established in the methodological literature and in the practice of educational research that designs without a comparison group cannot rule out a competing

explanation for observed differences between intervention and comparison groups on an outcome (Shadish, Cook, and Campbell, 2002).

*Types of participants*

We will include studies with participants who are K-12 students with TFA corps members, TFA alumni, and non-TFA teachers in rural and urban public schools in the United States. At the time of the intervention, the teachers in the treatment condition must be TFA corps members or TFA alumni; the control condition must include non-TFA teachers who have never participated in TFA. Non-TFA teachers may vary in their years of teaching experience and certification status. During the time frame of the study, all students must have a teacher who meets the eligibility criteria for TFA teachers or for non-TFA teachers. We will exclude studies that focus on Teach for America's Early Childhood Initiative, defined as initiatives that start prior to kindergarten, since early childhood studies are outside the policy relevant scope of this review.

*Types of interventions*

The TFA intervention condition will include TFA corps members, who are serving their two year commitment, TFA alumni who have completed the two year program but continue to teach, or both. The non-TFA comparison condition will include teachers who have never participated in TFA. These teachers must not have received preparation or training in programs associated with TFA. Teachers in the non-TFA comparison condition can vary in their certification status: traditional, alternative, emergency and uncertified. To be included in the review, the study must include a TFA condition (as described) and a non-TFA condition (as described). Studies that create an intervention group by bundling the TFA corps members or alumni with teachers trained in other alternative teacher preparation, such as the New York Teaching Fellows Program, will be excluded from the review. The reason is that when TFA is bundled in this way it is impossible to disentangle the effect of TFA from the effect of other alternatively prepared teachers in the intervention group.

*Types of outcome measures for students*

The review will include studies with at least one academic student outcome in math, English language arts, or science domains. Student outcomes in other non-academic (or non-cognitive) domains will be documented in the coding guide but will not be reported in the review. Multiple types of outcome measures will be included, although our experience with reviewing the TFA literature indicates that the primary types of outcome measures we will encounter will be state assessments, end-of-course assessments, and other standardized assessments.

State assessments, end-of-course assessments, and other standardized assessments are eligible for inclusion in the review provided that they were administered as intended. Non-standardized assessments, such as researcher-developed assessments, are eligible for

inclusion however the study must provide evidence that the measure (1) has face validity and reliability, (2) is not over aligned with the intervention, and (3) administered in the same way for both intervention and comparison groups. The first criterion is that the measure has face validity and sufficient reliable. A description that shows that the measure is clearly defined and measures the construct it is supposed to measure can serve as evidence for face validity in this review. Reliability evidence may come in the form of internal consistency, test-retest reliability, or inter-rater reliability. The second criterion is that a study must provide evidence that the measure does not closely resemble aspects of the intervention. For example, the measure should not have items or materials that intervention teachers have access to through their TFA training materials but that comparison teachers do not. The third criterion is that eligible outcome measures must have been used the same way in the treatment and comparison conditions.

*Type of outcome measures for teachers*

The review will include studies that meet the outcome inclusion criteria for students and have at least on teacher outcome on teacher leadership which is a key mediator between TFA teacher training and student achievement in the TFA theory of change. Additional teacher outcomes that will be included in the review are content knowledge, years of teaching experience, or overall academic ability.

Teacher Leadership. There is no single definition, however. One that comes closest to aligning with the TFA framework is teachers who take on leadership roles and additional professional responsibilities such that leadership roles and decision making responsibilities extend beyond the school or district administrative team to the teacher. Reliable and valid measures designed to tap into this construct will be eligible for the review.

Content knowledge. To be eligible for the review, an outcome should measure a teacher having a solid background in a subject or content area as exhibited by a college minor or major in the subject or content area such as math, reading, or science.

Teaching experience. To be eligible for the review, an outcome should measure the total number of years of classroom teaching experience in the field.

Academic ability. To be eligible for the review, an outcome should tap into the construct of of academic skills as measured by SAT scores, ACT scores, grade point average, or selectivity of the college attended.

*Validity criteria for student and teacher outcomes*

When reviewing outcome measures according to the three criteria, we will apply the definitions in the WWC Procedures and Standards Handbook Version 3.0, page 16, section 4. For example, thresholds for the psychometric properties that determine the reliability of an outcome measure will be based on the WWC Evidence Standards that require (a) internal

consistency of 0.50 or higher; (b) temporal stability/test-retest reliability of 0.40 or higher; or (c) inter-rater reliability of 0.50 or higher. Teacher outcomes such as years of teaching experience that do not have psychometric properties must show evidence of being collected consistently across groups in the study.

*Duration of follow-up*

One school year is the minimum dosage that study participants must have had in order for the study to be included in the review. All durations of follow-up above or equal to the minimum dosage will be included in the review. In the meta-analysis we will control for study to study differences in the follow-up period by meta-analyzing studies with the same follow-up periods (studies with one-year follow up outcome will be meta-analyzed on that outcome together, studies with two-year follow-up outcome will be meta-analyzed on that outcome together, and so on). Studies will be excluded if the minimum dosage is less than one school year or if the treatment and comparison groups do not have a comparable dosage.

*Types of settings*

The review will include studies that take place in K-12 public schools, including charter schools, in the United States. Limiting the setting to K-12 public schools helps ensure that the review will generate evidence that informs the TFA policy debate. Privately funded schools, early childhood education programs, higher education programs, adult education programs, and alternative schools, such as correctional programs, will not be included in the review.

### *Search strategy*

The goal of the literature search, consistent with the C2 Information Retrieval Policy Brief, is to identify all eligible studies on the effectiveness of TFA that are formally published (peer review literature) and informally published (grey literature). This involves developing search strategies that are efficient, capture the relevant studies while minimizing the amount of irrelevant material, and minimize bias. With this goal in mind, the final search strategy will be developed in consultation with a C2 Trials Search Co-coordinator and academic librarian at the University of Pennsylvania.

The literature search will be implemented by (1) searching electronic databases, (2) searching the grey literature in which studies are published informally, (3) soliciting previous authors of TFA studies, and (4) manually scanning the Table of Contents of the most current issues of those journals in which TFA effectiveness studies are published. To ensure study relevance, electronic searches will limit retrieved articles to those published (formally or informally) between 1994 and 2015. Our experience with the TFA effectiveness literature leads us to predict that this twenty year window is wide enough to include all of the effectiveness studies that have a comparison group and, therefore, would be eligible for the review.

*Searching Electronic Databases*

For the main electronic database search, we will use the ProQuest search engine available through the University of Pennsylvania's library web portal. This search engine provides access to and individual searching of 52 databases that index studies published formally and informally in a range of disciplines including education, economics, psychology, and sociology. Using this engine, we plan to search the following databases:

- ERIC,
- PsycINFO,
- EconLit,
- Sociological Abstracts,
- PAIS International,
- ProQuest Dissertations and Theses: UK and Ireland ,
- ProQuest Dissertations and Theses Global,
- Worldwide Political Science Abstracts

We also plan to search additional databases not covered by ProQuest. These are JSTOR, Academic Search Premier, and Education Next/Full Text.

When searching electronic databases, we will use search terms in consultation with our PENN library liaison at PENN GSE, using the following four domains:

1. Program names for Teach for America. (Searching only Title and Abstract).
2. Terms for the student outcomes targeted by the program.
3. Names for the student populations targeted by the program.
4. Terms for RCTs, QEDs, systematic reviews, and meta-analyses.

Based on these domains, we will employ search terms that connect the domains with the Boolean "AND" operator. Within domains, we will use the Boolean "OR" operator in order to search multiple keywords. The search terms will be similar to the following:

(TI ("Teach for America") OR AB ("Teach for America" OR "TFA Corps"))

AND

("academic achievement" OR "success*" OR "grade level" OR "grading" OR "academic ability" OR "Attainment" OR "failure" OR "educational indicator*")

AND

("kindergarten" OR "elementary school*" OR "primary school*" OR "high school*" OR "public school")

AND

("random assignment" OR "randomized experiment" OR experiment* OR "experimental design" OR "control group" OR "non-experiment" OR "non-experimental" OR "quasi-experiment" OR "quasi-experimental" OR "comparison group" OR "matched comparison group" OR "matched comparison" OR "matched groups" OR "statistical matching" OR "propensity score matching" OR "systematic review" OR "review" OR "meta-analysis" OR "research synthesis" OR "research review")

Publication Date = 1 Jan 1994 – 2015

Due to controlled vocabulary differences across databases, the main search term will change for each database. We will consult the database thesauruses in order to build the final search terms for each database.

We will search each database separately and tailor our strategy for each. For example, we will look up search words in each database thesaurus to see which descriptors are available, and make use of grade level and publication type filters. Results from our literature searches and other searches described next will be stored and managed using RefWorks online bibliographic software.

*Searching Grey Literature*

In addition to the main database searches, there will be a five-step grey literature search that involves 1) searching grey literature databases, 2) manually searching targeted websites, 3) searching conference presentation databases, 4) searching existing reviews, and 5) searching Google. The database searches will use PolicyFile, PsycExtra, and OpenGrey.eu. The manual grey literature searches will include general websites for organizations that conduct research across many areas of education (Table 1) as well as targeted websites for organizations that have a focus on teacher education or TFA research (Table 2). In order to make sure that relevant conference presentations are included in the meta-analysis, we will search the EditLib and Index of Conference Proceedings databases for conference abstracts using search criteria similar to what will be used for the main database search. We will also search existing reviews in order to refine the search strategy and check references for studies that should be included. Existing reviews will be identified through the main database searches and the grey literature searches as well as through searching the Campbell Library.

Lastly, we will do an advanced Google search where we use criteria similar to the main database search and screen the first twenty pages of results.

***Table 1: General Websites***

| Websites | |
|---|---|
| Abt Associates | Hoover Institute |
| Alliance for Excellent Education | Mathematica Policy Research |
| American Education Research Association | MDRC |
| American Enterprise Institute | National Association of State Boards of Education |
| American Institute of Research | National Governors' Association |
| Best Evidence Encyclopedia | Policy Archive |
| Brookings Institute | Policy Study Associates |
| Carnegie Corporation of New York | RAND |
| Center for Research and Reform in Education | Regional Educational Laboratories |
| Congressional Research Service | SRI |
| Government Accountability Office | Thomas B. Fordham Institute |
| Grants/contracts awarded by IES | Urban Institute |
| Heritage Foundation | |

***Table 2: Targeted Websites***

| Websites | |
|---|---|
| After-School Alliance | Database of Abstracts of Reviews of Effects |
| Campbell Collaboration | Florida Center for Reading Research |
| Carnegie Corporation for the Advancement of Teaching | Harvard Family Research Project |
| Center for Social Organization of Schools | Institute for Higher Education Policy |
| Chapin Hall Center for Children | Institute for Public Policy and Social Research |
| CINAHL | Natl. Association of State Directors of Career Tech. Ed. |
| Cochrane Central Register of Controlled Trials | NBER Working Papers |
| Cochrane Database of Systematic Reviews | |

*Soliciting Previous Authors of TFA Studies*

Based on studies we retrieved from our cursory/background searches, we will develop an email list of all researchers who authored an effectiveness study. We will also develop an email template that briefly describes the C2 systematic review on TFA, provides a bibliography of all effectiveness studies identified from our literature search whether or not these studies are eligible for our review, and requests that study authors refer us to 1) any

effectiveness studies not in the bibliography or 2) any authors not in the bibliography who may be aware of TFA studies that have not been formally published, or 3) both.

*Hand Searches of Journals*

The limited resources and personnel prevent us from conducting a comprehensive hand search of social science journals where TFA may be published. Moreover, our search of bibliographic databases that indexed grey literature, grey literature databases, and our outreach to previous authors of TFA studies should identify many of the studies, if any, not identified in our search of electronic databases. That said, consistent with guidance from the C2 Information Retrieval Policy Brief, we will have master level graduate students on our review team manually scan the table of contents of the most recent two journals that, based on our bibliography of effectiveness studies, are most likely to publish effectiveness studies on TFA. We will conduct this hand search in December 2015 towards the end of the review to compensate for any lag before articles are indexed by the databases.

### *Description of methods used in primary research*

Our cursory search and review of the literature helped us identify study characteristics and methodological issues that we may encounter in the main review. We will review experimental studies with random assignment and quasi-experimental studies that form groups through non-random methods such as matching. Studies we have encountered thus far, and expect to encounter in the future, use state assessments, end-of-course tests, or their standardized assessments as both pretest and posttest (or outcome measures), suggesting that it is likely that a number of studies will have reliable and valid outcomes.

Attrition is a major threat to the internal validity and causal inference that is typically associated with an RCT (WWC Procedures and Standards Handbook 3.0). This is because the combination of overall and differential attrition can disrupt the initial equating properties of randomization resulting in pre-existing differences between intervention and control groups being confounded with post-intervention differences between intervention and control groups, depending of course, on the amount of attrition. An RCT with high attrition will require that the reviewer establish baseline equivalence on a pre-intervention measure of the outcome using the analysis sample with complete case data for both the pretest and posttest. We will use the liberal attrition thresholds in the WWC Procedures and Standards Handbook Version 3.0, and encoded in the WWC Study Coding Guide for Group Design Standards, to determine whether an RCT has high attrition.

A major threat to the internal validity is a lack of baseline equivalence between the intervention and comparison groups' pre-intervention characteristics which could result in pre-intervention differences between the two groups being confounded with post-intervention differences between the two groups. Reviewers will be required to calculate baseline equivalence, using the WWC Study Coding Guide for Group Design Standards, with complete case data in the analysis sample. We will use the WWC standardized effect size

thresholds for determining whether a QED or an RCT with high attrition has groups that are baseline equivalent on a pre-intervention measure of the outcome. Group differences less than or equal to |.05| standard deviations are equivalent, differences greater than |.05| standard deviations but less than or equal to |.25| standard deviations are non-equivalent but can satisfy the baseline equivalence requirement by using a pre-test covariate adjustment in the analysis model, and differences greater than |.25| standard deviations are non-equivalent and cannot satisfy the requirement through a covariate adjustment.

A major challenge we anticipate when reviewing TFA effectiveness studies that are QEDs or RCTs with high attrition is reporting of data needed to calculate baseline equivalence between the analysis sample of TFA and non-TFA groups. The preferred data required for this calculation is un-imputed pre-intervention (or pre-test) values, sample sizes, and standard deviations for the groups compared. The t-statistic and sample sizes for the groups compared can also be used. Sometimes authors report this information; sometimes they don't. When they don't, we plan to initiate an author query to request this information. When authors refuse or cannot provide requested data, we will be forced to exclude the stud y from reporting in the meta-analysis, but the study will be reported in the appendix with a list of excluded studies and the reason for exclusion.

Based on all studies reviewed during the development of this protocol, we expect that missing data will be an issue. How researchers handle missing data can affect impact estimates and corresponding causal inferences. Acceptable methods for handling missing data are outlined in the WWC Procedures and Standards Handbook Version 3.0 section on appropriate missing data methods, and we will use those standards when evaluating the appropriate methods.

### *Criteria for determination of independent findings*

We will use methodology outlined in the C2 Statistical Analysis Policy Brief and the WWC Procedures and Standards Handbook Version 3.0 to deal with dependency in the data. Effects of TFA will be estimated for each comparison (TFA vs. non-TFA group) within the Math, English Language Arts, and Science outcome domains. If a study has more than one outcome in a domain--such as algebra and geometry in the math domain--then these outcomes will be combined, using the Comprehensive Meta-Analysis (CMA) software, into an average effect size resulting in a single outcome in the domain. When a study presents findings for several groups of students separately for the same outcome--findings (for example, for 5$^{th}$ grade TFA students contrasted with 5$^{th}$ grade non-TFA students on the math outcome along with 4$^{th}$ grade TFA students contrasted with 4$^{th}$ grade non-TFA students on the same math outcomes), we will use the CMA software to combine or average the 4$^{th}$ and 5$^{th}$ grade results into a single effect size. In sum, when there are multiple "related" comparisons on a "single" outcome or multiple "related" outcomes on a "single" comparison, we will always include one effect size in the meta-analysis. The effect size will be a weighted average implemented through CMA software.

### *Details of study coding categories*

An initial screening for eligibility will be conducted on all studies retrieved through the searches in order to determine eligibility for full review and coding. The initial screen will focus on the study's title and abstract; however, if the title and abstract do not contain the necessary information to complete screening, then we will obtain the full article. A study may be screened out as ineligible due to its topic, timeframe, sample, geographic location, design, or outcome relevance according to specifications outlined in the draft screening guide in Appendix A.

Studies that pass the initial title and abstract screen will be re-screened with further documentation based on the full text, this screen will take place on the first stage of the study coding guide. The second stage of the study coding guide will document research design issues in order to determine if the study has sufficient internal validity for inclusion in the statistical synthesis. Some of the design criteria that we will evaluate studies against include assignment to study groups, attrition, confounds, baseline equivalence and the reliability and validity of outcomes. The evaluation will differ slightly for randomized controlled trials and quasi-experimental designs as outlined in-depth in the draft coding guide found in Appendix A.

All coding and screening will be conducted by coders who are primarily Penn graduate students that have taken coursework in quantitative research methods; moreover, the coders will attend a training session on research design and evidence standards conducted by the review team. Coders independently drafted a coding guide. Then we went over that coding guide as a group. Next, coders were assigned studies, we gave coders independent feedback on their first draft of their first coding guide. Studies will be double screened and double coded. Coding disagreements will be reconciled to 100% agreement in a conference with both coders on the coding team. If both coders cannot come to agreement, a Principal Investigator (PI) for the review team will resolve the disagreement resulting in a final, master coding guide that will be used as "inputs" for the meta-analysis conducted with CMA 3.0. All lead members of the review team are certified in the use of What Works Clearinghouse Group Design Standards Version 3.0.

### *Statistical procedures and conventions*

*Goal of the Meta-Analysis*

The primary goal of the meta-analysis is to address review questions 3 through 10 by estimating the average effect of TFA on student academic outcomes, quantifying its precision (95% confidence intervals), evaluating whether the effect is real or due to chance using the actual p-value compared to the alpha level. Questions 10 is descriptive, and will report--for studies included in the meta-analysis--what authors reported on TFA's cost effectiveness.

*Data Extraction Using Excel Based Coding Guides*

Two reviewers will extract data from the articles independently and code methods, participant characteristics, intervention characteristics, and outcomes into the coding guide in Appendix A. Issues will be resolved in conference. If still unresolved, then they will be resolved in consultation with PI. If further information (e.g., missing data) is required regarding study data in order to conduct appropriate analyses of outcomes or to establish baseline equivalence, the first author of the study will be contacted. Should requested data be unavailable, the study will still be reported, but not included in the final meta-analysis. Information extracted from the studies eligible for review will be displayed in an appendix table of the review that report study characteristics for studies included and excluded in the statistical synthesis.

*Meta-Analysis*

Individual study effects will be synthesized statistically using Comprehensive Meta-Analysis software version 3.0 (CMA 3.0). The software allows for over 100 different data entry formats for effect size calculations. The choice of which effect size computation to use depends on three key factors: (1) the measures of the outcome variable(s), (2) the designs of studies being reviewed, and (3) the statistical analyses that have been reported. Depending on the study results, we will compute effect sizes from one of two types described next.

*Standardized Mean Difference*. TFA effectiveness studies reporting continuous outcomes will be summarized using standardized mean differences where different measures have been used to measure the same outcome. When the outcome has been measured using the same measure, a weighted mean difference will be used.

*Odds Ratio*. Although unlikely, TFA effectiveness studies may report binary data in which mean outcomes are compared in the experimental and control (or comparison) groups. If so, we will use CMA 3.0 to convert odds ratios, risk difference, or risk ratios to standardized mean differences with 95% confidence intervals for inclusion in the Meta-Analysis.

*Combining Effect Sizes*. Using CMA 3.0, we will convert all effect size indices to Hedge's g which is a standardized mean difference with a small sample size bias correction factor, it is unbiased for both small and large samples. When substantively feasible, effect sizes will be averaged across studies by using an inverse variance weighting of the individual effect sizes to account for differences in sample sizes for individual studies. This weighting will result in the individual effect sizes of larger n studies being given more weight in the combined effect size. We will calculate this effect size using a random effects model, which assumes that the effect size in each study is a sample estimate from a different population, and it estimates the parameter for that population. The estimates differ from study to study owing differences among the study population parameters (between-studies variation) and to sampling of different subjects within the study populations (within-study variation). More importantly, from a policy perspective, results from the random effects model allow for inferences from

the sample of studies to the population of studies from which the set was sampled. In this case, results of the data synthesis can be extrapolated beyond the studies in the set. However, we will also calculate effect sizes using fixed effects to provide a basis for comparison—this is easy to do with the CMA software. For both random effects and fixed effects models, individual study effect sizes and average effect sizes across studies will be reported with confidence intervals and corresponding p-values using a Forrest Plots.

*Homogeneity Analysis*

CMA 3.0 will allow us to quantify the amount of heterogeneity in the individual effect sizes that comprise the average effect size. This involves empirically distinguishing between variation in the individual effect sizes that is due to sampling error and variation in the individual effects that is due to true differences among studies. This can be done using two homogeneity statistics—$I^2$ and Q with the former preferred because it is not dependent on sample size and does not lead to inferential errors due to low statistical power.

*Sensitivity Analysis*

We will test the robustness of the conclusions drawn from the statistical synthesis through a sensitivity analysis on publication source and influential studies (or outliers with respect to effect size). To evaluate the possibility of publication bias, we will use funnel plots to assess the relationships between effect size and study precision. Such a relationship could be due to identifiable biases or to less obvious systematic differences between studies. To increase confidence in the robustness of the results of the data synthesis to overly influential studies (or outliers), we will conduct a "one-study removed" meta-analysis (provided there are at least three studies in the analysis) aimed at examining whether the results are sensitive to the inclusion or exclusion of particular studies. This analysis will be implemented in CMA 3.0 and estimates the overall effect size for all studies included in the meta-analysis with "one study" removed.

*Publication Bias*

Mean effect sizes of studies retrieved from peer reviewed sources will be compared to mean effect sizes of studies retrieved from unpublished sources (e.g., dissertations, government reports, and conference presentations). To evaluate whether the overall estimate of TFA's average effect is affected by publication bias, we will address three questions using methods recommended by Borenstein, Hedges, Higgins, Rothstein (2009, p. 277 - 291).

- Is there evidence of bias?
- Is it possible that the entire effect is an artifact of bias?
- How much of an impact might the bias have?

To address the first question, we will use a funnel plot to determine whether the studies are symmetrically distributed around the mean effect size within the funnel from top to bottom. To address the second question, we will use Rosenthal's Fail-safe N and Orwin's Fail-safe N with the second, an improvement on the first, by allowing us to 1) determine how many missing studies would bring the overall effect to a specified level other than zero and 2) specify the mean effect in the missing studies as some value other than zero. To address the third question, we will use Duval and Tweedle's Trim and Fill. This approach allows us to produce our best estimate of the unbiased effect size, although it is highly dependent on the assumptions of the model for why studies are missing--assumptions that we will state in the systematic review. These methodological approaches will be implemented with statistical validity using CMA 3.0.

*Moderator Analysis for Categorical Study-Level Variables*

The moderator analysis will be used to address research questions 7 - 9. We plan to code all moderators specified in these questions as categorical except for average teacher experience in the study--it will be coded as continuous. For categorical moderators, we are interested in "exploring" through a post-hoc analysis whether study effect sizes vary by the categories of the moderator. This analysis will be conducted for each moderator separately.

We assume that the studies within each moderator category do not share a common effect size. Under this assumption, we will use a random effects statistical model to weight the effect sizes in each moderator category and estimate the average effect size for each category. A separate estimate of Tau-squared (i.e., the true between study effect size variance) should always be estimated under random effects (Borenstein, Hedges, Higgins, and Rothstein, 2009). However, if there are fewer than five studies in a category, which is highly likely for this topic area, this estimate can be imprecise and we will use a pooled estimate. Using a pooled estimate makes more sense because the increased accuracy we get by pooling more studies is likely to exceed any real differences between groups (Borenstein, Hedges, Higgins, and Rothstein, 2009).

To evaluate whether average effect size differs by moderator category, when the moderator has only two categories, we will use the Z-test. When the moderator has more than two categories, we will use a Q-test based on analysis of variance. Although the multiple hypothesis tests across all the moderators could raise the "real" Type I error well above the "nominal" Type I error rate (or alpha level), we will not apply a multiple comparison correction to the alpha level because we are conceptualizing this post-hoc analysis as exploratory research to inform future research on TFA and not to determine whether TFA works or is effective. However, in analyses where such a correction is warranted, we will alert the reader to consider this correction who could use the results to inform their own statistical power analysis for a new RCT or QED on TFA or similar intervention.

*Moderator Analysis for Continuous Study-Level Variables*

For the one moderator (average years teaching experience) coded as continuous, we will use bivariate meta-regression with the effect size for an academic outcome as the dependent variable, and the moderator as a predictor. In this model, multi-collinearity among predictors is not a concern because there is only one. The meta-regression will be estimated using a random effects model (the model used in the main effects analysis) with a Z-test of the relationship between the two variables (e.g., effect size and predictor). To quantify the relationship between the two variables, we will report all of the meta-regression parameter estimates, standard errors, p-values, and variance explained index (e.g., R-squared) analogous to regression in primary studies.

*Reporting Results and Risk Bias*

Results from the main effect, subgroup, and moderator analyses will be reported according to the research questions these analyses are designed to address. Results will be reported using forest plots with study sample sizes, effect sizes, 95% confidence intervals, p-values, tests of homogeneity, and model choice of fixed or random effects. Moderator analyses will be reported with three important methodological caveats.

- First, the analysis is post-hoc and should be interpreted as exploratory to inform future research.

- Second, the analysis is observational and therefore the interpretation is correlational rather than causal. Thus any meaningful and statistically significant effect size differences between moderator categories cannot be interpreted as "caused" by the moderator.

- Third, failing to reject the null hypothesis of no effect size difference between or among moderator categories is not necessarily definitive evidence of no "True" effect size difference between or among categories. We anticipated low statistical power (of the statistical tests) due to the potential small number of studies within a category.

Primary studies will be assessed for quality, risk of bias, and assigned a rating of meets design quality standards with or without reservations based on the criteria (Table B1 in Appendix B)

*Missing Data and Author Queries*

In research studies, authors sometimes do not report results from some of the analyses that were mentioned in the study. If any study mentions but does not report information that we need for our statistical synthesis, then we will send a letter to the study authors in an effort to obtain the relevant information that was mentioned. The authors will have two weeks to respond, but can request more time if needed.

### *Treatment of qualitative research*

To address research question 10, we plan to report what researchers have found on the cost effectiveness of TFA to provide context for study effect sizes.

## REFERENCES


Blazer, Christie. (2012). What the research says about alternative teacher certification programs. Information Capsule. Volume 1104. Research Services, Miami-Dade County Public Schools

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (2009). *Introduction to meta-analysis*. West Sussex, UK: Wiley & Sons.

Chalmers, I., & Altman, D.G. (Eds.). (1995). Systematic reviews. London: BMJ.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from teach for america and the teaching fellows programs*. NCEE 2013-4016. National Center for Education Evaluation and Regional Assistance, P.O. Box 1398, Jessup, MD 20794-1398.

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2015). *Impacts of the Teach for America Investing in Innovation Scale-Up*. Princeton, NJ: Mathematica Policy Research.

Clotfelter, C.T., Ladd, H.F., and Vigdor, J.L., (2006). *The Academic Achievement Gap in Grades 3 to 8,* NBER Working Papers 12207, National Bureau of Economic Research, Inc.

Constantine, (2009). Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). An evaluation of teachers trained through different routes to certification. final report. NCEE 2009-4043. National Center for Education Evaluation and Regional Assistance, P.O. Box 1398, Jessup, MD 20794-1398.

Darling-Hammond, L. (1984). *Beyond the commission reports. The coming crisis in teaching*. Santa Monica: RAND. (No. R-3117-RC). Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED248245

Copper, H. (2010). *Research synthesis and meta-analysis (4th Edition)*. Thousand Oaks: Sage.

Decker, P.T., Mayer, D.P., and Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation*. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from http://www.mathematica-mpr.com/publications/pdfs/teach.pdf

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 37*(3), 499–534.

Ingersoll, R.M. and Perda, D. (2010). Is the Supply of Mathematics and Science Teachers Sufficient? *American Educational Research Journal, 43*(3). pp. 563-594.

Hess, F.M. (2002). Tear down this wall: The case for a radical overhaul of teacher certification. *Educational Horizons, 80*(4), 169-183.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2007). Photo finish: Certification doesn't guarantee a winner. *Education Next,* 7(1), 60-67.

Klingner, J.K., Boardman, A.G., & Mcmaster, K.L. (2013). What Does It Take to Scale Up and Sustain Evidence-Based Practices? *Exceptional Children, 79*(2), 195-211.

Laczko-Kerr, I., & Berliner, D. C. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10*(37), 1-53.

Monk, D.H. (2007). Recruiting and retaining high-quality teachers in rural areas, *Future of Children, 17(1),* 155-174.

Mead, S. (2015 ). *Love 'Em or Hate 'Em, Here's What You Should Learn From Teach For America's Success* [Real Clear Education]. Retrieved from http://www.realcleareducation.com/articles/2015/02/03/teach_for_america_growth_1152.html

Monk, D.H. (2007). Recruiting and Retaining High-Quality Teachers in Rural Area. *Future of Children, 17(1),* 155–174.

Noell, G.H., and Gansle, K.A. (2009). Teach for America teacher's contribution to student achievement in Louisiana in grades 4-9: 2004-2005 to 2006-2007.

Peske, H. G., & Haycock, K. (2006). *Teaching Inequality: How Poor and Minority Students Are Shortchanged on Teacher Quality: A Report and Recommendations by the Education Trust*. Washington, DC: The Education Trust.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell.

Raymond, M., and Fletcher, S. (2002). The Teach for America Evaluation. *Education Next*, 2(1), 62–68.

Raymond, M., Fletcher, S. H., and Luque, J. (2001). Teach For America: An evaluation of teacher differences and student outcomes in Houston, Texas. Retrieved from http://credo.stanford.edu/downloads/tfa.pdf.

Sanders, W.L., & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Seftor, N., & Mayer, D. P. (March 31, 2003). *The effect of alternative certification on student achievement: A literature review*. Unpublished manuscript. Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/effectalt.pdf.

Shadish, W.R., Cook, and Campbell (2002*). Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Teach For America. (2010). *Building the movement to eliminate educational inequity* Retrieved from: http://www.socialimpactexchange.org/files/TFA%20Business%20Plan.pdf

Williams, C.P (2014). Stop scapegoating teacher for America. Education Nation. Retrieved from: http://www.thedailybeast.com/articles/2014/09/24/stop-scapegoating-teach-for-america.html

## APPENDIX A: CODING GUIDE

See Excel-based coding guide in C2 Library with this protocol.

## APPENDIX B: RISK OF BIAS

### *Table B1: Assessing risk of Bias in Primary Studies for TFA Review*

| Criteria | Description |
|---|---|
| **Screening** | In order to be included in the design quality review, a study must have all of the following characteristics: |
| Focus | Study must focus on the effectiveness of the TFA intervention. |
| Time | Study must be published or reported between 1995 and the present. |
| Age | Study must focus on students in grades K-12. |
| Location | TFA must be implemented in the United States. |
| Outcome | Study must report at least one student academic outcome in math, English language arts, or science. |
| Exposure | Students in the TFA group must have at least one school year of exposure to the TFA corps member or alumni before outcome measurement. |
| Setting | Study of TFA must take place in a US public or charter school. |
| Design | The TFA and counterfactual groups must be formed with random assignment or quasi-experimental methods. |
| **Design Quality** | In order to be included in the statistical synthesis, a contrast must satisfy the following criteria: |
| Bundled Treatment Group Confound | To qualify as an intervention teacher, individual must be a TFA corps member or TFA alumni. |
| Bundled Intervention Confound | The TFA intervention must *not* be completely aligned with another intervention such as another alternative route teacher preparation program. |
| N=1 Confound | For each condition, there must be more than one unit at each level (student, teacher, school, district, state, and so on). |
| Outcome Face Validity | A description of the outcome must provide evidence that the measure is well defined, interpretable, and measures what it is purported to measure. |
| Outcome Reliability | Measures must demonstrate an internal consistency reliability of 0.50, inter-rater reliability of 0.50, or a temporal stability reliability of 0.40. |
| Outcome Alignment | Outcomes must not be over aligned with the intervention. |
| Outcome Measure Confound | Outcomes must be measured in the same way for both conditions. |
| Standardized Outcomes | Standardized outcomes are assumed to satisfy the face validity, reliability, alignment, and measurement confound criteria. |
| Attrition | RCTs must demonstrate low attrition. Otherwise a baseline equivalence test is required. Attrition is considered high if the combination between overall and differential attrition exceeds the thresholds defined by the WWC's liberal attrition standard. |
| Cluster Attrition | Cluster RCTs must test for high attrition at the cluster level. In addition, they must test for high attrition at the sub-cluster level using the clusters with outcome data. |

| Criteria | Description |
|---|---|
| Baseline equivalence | All QEDs and RCTs with high attrition must use a pre intervention measure of the outcome to show evidence of baseline equivalence on the analysis sample.<br>• If the absolute value of the effect size is less than or equal to 0.05, the groups are considered equivalent.<br>• If the absolute value of the effect size is great than 0.05 and less than or equal to 0.25 then the analysis model must statistically adjust for the pre-intervention measure.<br>• If the absolute value of the effect size is greater than 0.25 then the outcome is not eligible for the statistical synthesis. |
| Design Quality | All studies that satisfy the above criteria will be included in the statistical synthesis and assigned a study rating as follows<br>• RCTs that with no confounds, reliable outcomes, and low attrition will receive a rating of meets design quality standards without reservations.<br>• RCTs with high attrition and QEDs with no confounds, reliable outcomes, and baseline groups equivalence in analysis samples meets design quality standards without reservations. |
| **Risk of Bias Conduct** | The following describes how this review will comply with the Adaptations on MECIR Version 2.2 Conduct Standards document on risk of bias issues. |
| Assessing Risk of Bias/Study Quality | The Risk of bias will be assessed for all RCTs and QEDs that pass screening using the design criteria outlined in the WWC Procedures and Standards Handbook Version 3.0. This specific criteria is outlined in the rows above under the design quality heading. |
| Assessing Risk of Bias /Study Quality in Duplicate | All studies that pass the initial title and abstract screen will be double coded by trained coders. |
| Supporting Judgments of Risk of Bias/Study Quality | Coders will use information directly from the study to justify all decision. |
| Providing Sources of Information for Risk of Bias/Study Quality Assessments | Coders will collect and document the source of information for each study quality assessment. They will clearly document what information comes directly from the report, what information was obtained from the author query, and what assumptions were made. |
| Differentiating Between Performance Bias and Detection Bias | Selection bias and attrition bias will be assessed through our use of WWC Evidence Standards 3.0 with the former assessed through how groups are formed and the latter assessed for RCTs based on whether there is high attrition.<br>Detection bias and reporting bias will be assessed based on whether the authors report all outcomes for which data were collected or only reported a subset (without justification). |
| If Applicable, Assessing Risk of Bias Due to Lack of Blinding for Different Outcomes | This assessment is rare in TFA studies in particular and in education research studies in general because “concealment” of the intervention condition from the investigator or study participants or both, is rarely feasible or practical. |
| If Applicable, Assessing Completeness of Data for Different Outcomes | Within a study completeness of data may be handled differently for different contrasts. RCTs with low attrition may use the missing data techniques outlined in the WWC Procedures and Standards Handbook Version 3.0. |

| Criteria | Description |
|---|---|
| If Applicable, Summarizing Risk of Bias Assessments When Using the Cochrane Risk of Bias tool | Not applicable. The Cochrane Risk of Bias Tool is not being used. |
| Addressing Risk of Bias/Study Quality in the Synthesis | This review will use sensitivity analysis to check if the results are sensitive to the inclusion or exclusion of particular studies using a "one study removed" analysis. In addition, results for RCTs and QEDs will be reported separately. |
| Incorporating Assessments of Risk of Bias | Within each study, each contrast (intervention vs. comparison group on each outcome) will be assigned a quality rating (see design quality row above) that incorporates the risk of bias. |
| **Risk of Bias Reporting** | The following describes how this review will comply with the Adaptations on MECIR Version 1.1 Reporting Standards document on risk of bias issues. |
| 'Risk of bias' and/or Study Quality Table<br>Summary Assessments of Risk of Bias/Study Quality<br>Study Quality/Risk of Bias in Included Studies | After assessing all design quality issues, the study coding guide will require coders to select a study rating, select a study disposition code (with an explanation) as follows:<br>• RCTs that with no confounds, reliable outcomes, and low attrition will receive a rating of meets design quality standards without reservations.<br>• RCTs with high attrition and QEDs with no confounds, reliable outcomes, and baseline groups equivalence in analysis samples meets design quality standards without reservations |

## REVIEW AUTHORS

**Lead review author:**

| **Name:** | **Herbert M. Turner, III** |
|---|---|
| Title: | President and Principal Scientist / Adjunct Associate Professor |
| Affiliation: | Analytica, Inc. / University of Pennsylvania |
| Address: | 35 Goldfinch Circle |
| City, State, Province or County: | Phoenixville, PA |
| Postal Code: | 19460 |
| Country: | USA |
| Phone: | 610.933.1005 |
| Email: | herb@analytica-inc.com |

**Co-author(s):**

| **Name:** | **Robert F. Boruch** |
|---|---|
| Title: | University Trustee Chair Professor of Education and Statistics |
| Affiliation: | University of Pennsylvania |
| Address: | 3700 Walnut Street |
| City, State, Province or County: | Philadelphia, PA |
| Postal Code: | 19104-6216 |
| Country: | USA |
| Phone: | 215.898.0409 |
| Email: | Email: robertb@gse.upenn.edu |

| Name: | **Annette E. Turner** |
|---|---|
| Title: | Chief Executive Officer |
| Affiliation: | Analytica, Inc. |
| Address: | 35 Goldfinch Circle |
| City, State, Province or County: | Phoenixville, PA |

| | |
|---|---|
| Postal Code: | 19460 |
| Country: | USA |
| Phone: | 610.933.1005 |
| Email: | annette@analytica-inc.com |

## ROLES AND RESPONSIBLIITIES

Content:

- *Review team*: Dr. Herb Turner, Dr. Robert Boruch, Mr. Mackson Ncube, Mrs. Annette Turner
- *Content Advisory*: David Goodman

Systematic review methods:

*Review team*: Dr. Herb Turner, Dr. Robert Boruch, & Mr. Mackson Ncube

- *Methodological Advisor*: Dr. Michael Borenstein

Statistical analysis:

- *Review team*: Dr. Robert F. Boruch & Mr. Mackson Ncube
- *Statistical Advisor*: Dr. Michael Borenstein

Information retrieval:

- *Review team*: Dr. Herb Turner, Mr. Mackson Ncube, and Mrs. Annette Turner
- *Content group*: David Goodman

## SOURCES OF SUPPORT

This review is supported in part by a grant from the Campbell Collaboration Education Coordinating Group awarded to the first author.

## DECLARATIONS OF INTEREST

The authors have no conflicts to declare.

## PRELIMINARY TIMEFRAME

| Stage | Month of Completion |
|---|---|
| Title Registration Submitted | September, 2014 |
| Draft Protocol Submitted | December, 2014 |
| Revised Protocol Submitted | July, 2015 |
| Review Submitted | February, 2016 |

## PLANS FOR UPDATING THE REVIEW

Dr. Turner and Mr. Mackson Ncube will be responsible for updating the review.

## AUTHOR DECLARATION

### Authors' responsibilities

By completing this form, you accept responsibility for preparing, maintaining and updating the review in accordance with Campbell Collaboration policy. The Campbell Collaboration will provide as much support as possible to assist with the preparation of the review.

A draft review must be submitted to the relevant Coordinating Group within two years of protocol publication. If drafts are not submitted before the agreed deadlines, or if we are unable to contact you for an extended period, the relevant Coordinating Group has the right to de-register the title or transfer the title to alternative authors. The Coordinating Group also has the right to de-register or transfer the title if it does not meet the standards of the Coordinating Group and/or the Campbell Collaboration.

You accept responsibility for maintaining the review in light of new evidence, comments and criticisms, and other developments, and updating the review at least once every five years, or, if requested, transferring responsibility for maintaining the review to others as agreed with the Coordinating Group.

### Publication in the Campbell Library

The support of the Coordinating Group in preparing your review is conditional upon your agreement to publish the protocol, finished review, and subsequent updates in the Campbell Library. The Campbell Collaboration places no restrictions on publication of the findings of a Campbell systematic review in a more abbreviated form as a journal article either before or after the publication of the monograph version in *Campbell Systematic Reviews*. Some journals, however, have restrictions that preclude publication of findings that have been, or will be, reported elsewhere and authors considering publication in such a journal should be aware of possible conflict with publication of the monograph version in *Campbell Systematic Reviews*. Publication in a journal after publication or in press status in *Campbell Systematic Reviews* should acknowledge the Campbell version and include a citation to it. Note that systematic reviews published in *Campbell Systematic Reviews* and co-registered with the Cochrane Collaboration may have additional requirements or restrictions for co-publication. Review authors accept responsibility for meeting any co-publication requirements.

**I understand the commitment required to undertake a Campbell review, and agree to publish in the Campbell Library. Signed on behalf of the authors**:

**Form completed by:** Herbert M. Turner, III **Date: 12 December 2014**