**SREE 2020 Proposal Abstract**

**Title: Considerations for Evidence Standards in Education Research**

**Authors and Affiliations:**
**Elisabeth Davis, AIR**
**Joseph Taylor, AIR**

## Background

Despite a history that can be tracked to 1867, the US Department of Education is a relative newcomer to applying evidence standards to education research. In fact, it wasn't until criticisms about the failure of the education research community to accumulate knowledge became overwhelming, including those aired in the US House of Representatives, that federal action was taken (NRC, 1992; Fuhrman, 2001). These actions included the establishment of the Institute of Education Sciences via the Education Sciences Reform Act (United States Congress, 2002) and shortly afterward its large investment in the What Works Clearinghouse (Institute of Education Sciences, 2019).

The WWC reviews intervention research to support evidence-based decision making. This is done by applying standards of research quality to studies and giving those studies ratings that indicate the extent to which the findings are trustworthy. In addition, nonprofit entities have also invested in vetting research toward more informed decision-making in education and other policy areas. There is much overlap in the interventions of interest to the WWC and other research vetting entities. As such, comparing the different conclusions that might be drawn about the same program can be insightful to on-going refinement of these standards and to the implications of those differences for the identification and implementation of evidence-based practices under the Every Student Succeeds Act (ESSA).

## Methods

We model the comparative approach using a case study of the *Career Academies* program, reviewed under WWC Group Design Standards, Blueprints for Healthy Youth Development (Univeristy of Colorado Boulder, 2019), and Social Programs that Work (Arnold Ventures', 2019) standards. The goal of this comparison is to examine these and other standards frameworks, extending the case study of *Career Academies* toward the goal of illuminating differences in framework-specific assessments of interventions that could confound decision making for practitioners.

## Considerations for Design and Evidence Standards

There are several components considered by evidence-based clearinghouses when reviewing the evidence of educational programs. Some focus on the program itself, however, most of these components focus specifically on the research study conducted to test the program. Table 1 (see appendix) outlines these components as they relate to the study, program implementation, and program impact.

We feature three online clearinghouses that review educational programs. We first describe the criteria considered for each in determining their standards, then crosswalk one program, *Career Academies*, that has been reviewed by all three clearinghouses to highlight how these different entities review and categorize the same program in different ways.

### The What Works Clearinghouse

The WWC Group Design Standards focus on the quality of individual *study* design methods and have three possible study ratings. The *effectiveness* of interventions is done using a separate rating scheme for those studies that meet the WWC standards. The WWC study ratings are provided in Table 2 and the effectiveness ratings in Figure 1.

### Blueprints for Healthy Youth Development

Blueprints for Healthy Youth Development is a nonprofit registry of programs across the social policy spectrum whose effectiveness have been rigorously tested, mostly through randomized control trials. Blueprints focuses on programs across the social policy spectrum designed to reduce antisocial behavior and promote healthy development and adult maturity. Blueprints ratings, which are determined at the *program* level, are Promising, Model, and Model+. These are described in Table 3 (appendix).

### Social Programs that Work

The mission of Social Programs that Work (SPW) is to help policy makers and the general public identify rigorous evidence in all areas of social policy and aid in data-driven decision making. The clearinghouse exclusively reviews experimental studies and examines the evidence related to several education policy areas. SPW's ratings, designated at the *program* level are Top Tier, Near Top Tier, and Suggestive Tier. These ratings are described in table 4 in the appendix.

### Comparing Ratings Across Clearinghouses

Career Academies are a dropout prevention strategy for at risk students that uses a school-within-a-school model, with each academy focused on a specific workforce theme such as health care or communications. This program has been reviewed by several clearinghouses, including WWC, Blueprints, and SPW. Table 5 in the appendix provides a crosswalk of these clearinghouses and the components considered in their respective evidence standards. In the sections that follow, we provide a comparison of how the *Career Academies* program has been rated across entities.

**WWC review**. The WWC reviewed nine studies, one of which *Meets WWC Standards Without Reservations* (WWC's highest rating). The Career Academies Intervention Report (What Works Clearinghouse, 2015) rates the intervention as having potentially positive effects (WWC's second highest effectiveness rating) for the outcome domains of staying in school and completing school, and as having no discernable effects for the outcome domain of progressing in school.

**Blueprints for Healthy Youth Development review**. Blueprints rates Career Academies as *Promising* (Blueprint's lowest rating). The outcomes of focus for the Blueprints review of Career

Academies were employment and attendance, which indicates that these are the outcome for which Career Academies are promising for improvement.

**Social Programs that Work review**. SPW rates Career Academies as *Top Tier* (SPW's highest rating). The one RCT reviewed was a large, multisite study with sustained effects, and the outcome of interest was average annual earnings eight years after high school graduation.

**Implications**

The review of programs across clearinghouses can yield different ratings bases on divergent standards, and varying emphases on the merits of research design, fidelity of implementation, and the impacts of the program. Clearinghouses address these components of evidence review, but no one entity addresses all with equal scrutiny. For decision makers in education, a decision informed by a well designed and implemented experimental study alone provides no assurance that the program was implemented as intended, could be implemented in their local context, nor does it assume favorable impacts of that program. Further, a program that is effective in one outcome domain may not be equally as effective in others. The results of this comparison have implications for the continued refinement of evidence standards across the educational research field.

## References


Arnold Ventures' Evidence-Based Policy Team. (2019). Social Programs that Work. Retrieved from: https://evidencebasedprograms.org/

Fuhrman, S. (2001). *The policy influence of education R&D centers*. Testimony to The U.S. House Committee on Education and the Workforce. 107th Cong., 1st Sess.

Institute of Education Sciences. (2019). *What Works Clearinghouse*. Retrieved from: https://ies.ed.gov/ncee/wwc/

National Research Council. (1992). Research and education reform: Roles for the Office of Educational Research and Improvement. Committee on the Federal Role in Education Research. R.C. Atkinson and G.B. Jackson, Eds. *Commission on Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

United States Congress. (2002*). Education Sciences Reform Act*. Retrieved from: https://www.congress.gov/107/plaws/publ279/PLAW-107publ279.pdf

University of Colorado Boulder Institute of Behavior Science. (2019). Blueprints for Healthy Youth Development. Retrieved from: https://www.blueprintsprograms.org/

What Works Clearinghouse (2015). *Career Academies Intervention Report*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_careeracademies_092215.pdf

## Tables and Figures

**Table 1. Components to consider when determining the evidence of educational programs**

| Component | Description | Focus |
|---|---|---|
| Methodological rigor | Design and execution of the study, measurement validity and reliability, confidence that the program lead to the outcome | Study |
| Generalizability | Sample size, location and setting, confidence that the program will yield the same results in similar settings | Study |
| Replication | Whether the program has been studied more than once | Study |
| Independence | Whether study team is free of bias/conflict of interest | Study |
| Implementation fidelity | Whether the program was implemented as intended by developers | Program |
| Program replicability | Whether the program can be implemented as intended elsewhere | Program |
| Program impacts | Whether the study yielded favorable effects | Outcomes |
| Outcome differentiation | Effects of the program on specific outcomes, for whom, and under what circumstances | Outcomes |
| | | |

**Table 2. WWC Study Ratings**

| Study Rating | Description |
|---|---|
| **Meets Standards without Reservations**. | This is reserved for randomized control trials (RCTs) with uncompromised random assignment processes, low attrition, and no other disqualifying study artifacts (see below). |
| **Meets Standards with Reservations.** | This rating is given to a RCT with compromised random assignment, or high attrition, but demonstrates baseline equivalence of groups and has no other disqualifying study artifacts<br>OR<br>A quasi-experimental design (QED) that demonstrates baseline equivalence of groups and has no other disqualifying study artifacts. |
| **Does not Meet Standards.** | This rating is given to a RCT that has compromised random assignment or high attrition and cannot demonstrate baseline equivalence or has even one of the other disqualifying artifacts<br>OR<br>A QED that cannot demonstrate baseline equivalence of groups or has even one of the other disqualifying study artifacts |
| **Disqualifying Study Artifacts**<br>These include: confounded treatment effects, outcome measures that lack face validity, reliability, are overaligned to one of the treatment conditions, or were administered differently in one of the treatment conditions. | |

**Figure 1. WWC Effectiveness Levels**

**Table IV.3. Criteria Used to Determine the WWC Rating of Effectiveness for an Intervention**

| | |
|---|---|
| Positive effects: Strong evidence of a positive effect with no overriding contrary evidence. | • Two or more studies show statistically significant positive effects, at least one of which meets WWC group design standards without reservations, AND<br>• No studies show statistically significant or substantively important negative effects. |
| Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence. | • At least one study shows statistically significant or substantively important positive effects, AND<br>• Fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects, AND<br>• No studies show statistically significant or substantively important negative effects. |
| No discernible effects: No affirmative evidence of effects. | • None of the studies show statistically significant or substantively important effects, either positive or negative. |
| Mixed effects: Evidence of inconsistent effects. | EITHER both of the following:<br>• At least one study shows statistically significant or substantively important positive effects, AND<br>• At least one study shows statistically significant or substantively important negative effects, BUT no more such studies than the number showing statistically significant or substantively important positive effects.<br>OR both of the following:<br>• At least one study shows statistically significant or substantively important effects, AND<br>• More studies show an indeterminate effect than show statistically significant or substantively important effects. |
| Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence. | EITHER both of the following:<br>• One study shows statistically significant or substantively important negative effects, AND<br>• No studies show statistically significant or substantively important positive effects.<br>OR both of the following:<br>• Two or more studies show statistically significant or substantively important negative effects, at least one study shows statistically significant or substantively important positive effects, AND<br>• More studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects. |
| Negative effects: Strong evidence of a negative effect with no overriding contrary evidence. | • Two or more studies show statistically significant negative effects, at least one of which meets WWC group design standards without reservations, AND<br>• No studies show statistically significant or substantively important positive effects. |

*Source: WWC Procedures and Standards Handbook, 4.0*

**Table 3 Blueprints For Healthy Youth Development ratings**

| Rating | Description |
|---|---|
| **Promising Programs** | At least one RCT or two QEDs must meet evaluation quality standards (such as treatment group assignment, instrument alignment and psychometrics, attrition and baseline equivalence, independence between data collection and implementation of the intervention, and implementation fidelity), intervention impact standards (statistically significant favorable effects with no "iatrogenic" effects), intervention specificity standards (acceptable documentation of all aspects of the intervention as it is intended to be implemented, for who, and under what circumstances), and dissemination readiness standards (implementation guide, cost information, resources needed to implement the program with fidelity in other settings). |
| **Model Programs** | All standards of promising programs met, plus more that one RCT or one RCT and one QED have examined the effectiveness of the program and sustained significant favorable effects for at least one year on at least one outcome. |
| **Model+ Programs** | All standards and criteria of Model Programs, plus replication of results by an independent research team, with no financial ties to the program. |

**Table 4. Social Programs that Work ratings**

| Rating | Description |
|---|---|
| **Suggesting Tier** | This tier is reserved for programs that have been evaluated by one or more well designed, well implemented experimental studies with favorable effects, but are thus far limited by lack of sustained effects, statistically significant findings, or being conducted in more that one setting. |
| **Near Top Tier** | Programs designated as Near Top Tier meet all the criteria for Top Tier evidence but need an additional step to qualify for Top Tier (in most case, replication of effects). |
| **Top Tier** | For a program to be designated as top tier, it must be evaluated by more than one well designed, well implemented experimental study (or one large, multisite study) in a replicable setting with large, favorable, sustained effects. The studies also must present no countervailing negative effects, and have been conducted in more that one setting. |

**Table 5. Matrix of evidence rating components across WWC, Blueprints, and SPW standards**

| Component | WWC | Blueprints | SPW |
|---|---|---|---|
| Methodological rigor | X | X[a] | X[a] |
| Generalizability | | | X |
| Replication | | X | |
| Independence | | X | |
| Implementation fidelity | | X | |
| Program replicability | | X | X |
| Program impacts | X[b] | X | X |
| Outcome differentiation | X | | |

a. Does not – or rarely – reviews QEDs

b. WWC Intervention Reports report program effectiveness for studies that meet *WWC Standards With or Without Reservations*, however, program impacts are not a component of the WWC design standards.