Criteria for Classifying Rigorous Designs of MSP Evaluations

This appendix includes the *Criteria for Classifying Rigorous Designs of MSP Evaluations* used to determine the number of projects that successfully conducted rigorous evaluations. The criteria were developed as part of the Data Quality Initiative (DQI) through the Institute for Education Sciences (IES) at the U.S. Department of Education. The results of the review of final year MSP projects according to these criteria were presented in Appendix A.

Criteria for Classifying Rigorous Designs of MSP Evaluations

❒ **Experimental study—**the study measures the intervention's effect by randomly assigning individuals (or other units, such as classrooms or schools) to a group that participated in the intervention, or to a control group that did not; and then compares post-intervention outcomes for the two groups

❒ **Quasi-experimental study—**the study measures the intervention's effect by comparing post-intervention outcomes for treatment participants with outcomes for a comparison group (that was not exposed to the intervention), chosen through methods other than random assignment. For example:

- *Comparison-group study with equating*—a study in which statistical controls and/or matching techniques are used to make the treatment and comparison groups similar in their pre-intervention characteristics

- *Regression-discontinuity study*—a study in which individuals (or other units, such as classrooms or schools) are assigned to treatment or comparison groups on the basis of a "cutoff" score on a pre-intervention non-dichotomous measure

Criteria for Assessing whether Experimental and Quasi-experimental Designs Were Conducted Successfully and Yielded Scientifically Valid Results

A. Data Reduction Rates (i.e. Attrition Rates, Response Rates)[1]

❒ **Met the criterion**. Key post-test outcomes were measured for at least 70 percent of the original sample (treatment and comparison groups combined) and differential attrition (i.e., difference between treatment group attrition and comparison group attrition) between groups was less than 15 percentage points.

❒ **Did not meet the criterion**. Key post-test outcomes was measured for less than 70 percent of the original sample (treatment and comparison groups combined) and/or differential attrition (i.e., difference between treatment group attrition and comparison group attrition) between groups was 15 percentage points or higher.

❒ **Not applicable.** This criterion was not applicable to quasi-experimental designs unless it was

[1] The data reduction and baseline equivalent criteria were adapted from the What Works Clearinghouse standards (see http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf).

required for use in establishing baseline equivalence (see the *Baseline Equivalence of Groups* criterion below).

B. Baseline Equivalence of Groups

❒ **Met the criterion (quasi-experimental studies).** There were no significant pre-intervention differences, as defined below, between treatment and comparison group participants in the analytic sample on the outcomes studied, or on variables related to the study's key outcomes. Two groups are considered to have baseline equivalence when:

- the mean difference in the baseline measures was less than or equal to five percent of the pooled sample standard deviation; *or*
- the mean difference in the baseline measures was more than five percent but less than or equal to twenty-five percent of the pooled sample standard deviation, and the differences were adjust for in analyses (e.g., by controlling for the baseline measure); *or*
- If the data required for establishing baseline equivalence in the analytic sample were missing (and there was evidence that equivalence was tested), then baseline equivalence could have been established in the baseline sample *providing the data reduction rates criterion above was met.*

❒ **Met the criterion (experimental evaluations that did not meet the data reduction rates criterion above).** There were no significant pre-intervention differences, as defined above, between treatment and comparison group participants in the analytic sample on the outcomes studied, or on variables related to the study's key outcomes.

❒ **Did not meet the criterion.** Baseline equivalence between groups in a quasi-experimental design was not established (i.e. one of the following conditions was met):

A. Baseline differences between groups exceeded the allowable limits; *or*
B. The statistical adjustments required to account for baseline differences were not conducted in analyses; *or*
C. Baseline equivalence was not examined or reported in a quasi-experimental evaluation (or an experimental evaluation that did not meet the data reduction rates criterion above) and the necessary information was not provided such that reviewers could calculate it themselves.

❒ **Not applicable.** This criterion was not applicable to experimental designs that met the data reduction rates criterion above.

C. Quality of the Measurement Instruments

❒ **Met the criterion**—the study used existing data collection instruments that had already been deemed valid and reliable to measure key outcomes; a new instrument was created from an existing instrument(s) that has been validated and found to be reliable; or data collection instruments developed specifically for the study were sufficiently pre-tested with subjects who were comparable to the study sample or high reliability was established.

❒ **Did not meet the criterion**—the key data collection instruments used in the evaluation lacked evidence of validity and reliability

❒ **Did not address the criterion**

D. Relevant Statistics Reported

❒ **Met the criterion**—the final report includes treatment and control group post-test means, and tests of statistical significance that directly compare the treatment and comparison groups for key outcomes; or provides sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/standard error).

❒ **Did not meet the criterion**—the final report does not include treatment and control group post-test means, and/or tests of statistical significance for key outcomes; or provide sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/standard error).

❒ **Did not address the criterion**