

Annual Performance Report

<https://apr.ed-msp.net/>

The MSP Program **requires** projects to report on two aspects of evaluation findings:

- Changes in teacher content knowledge based on pre- and post-testing; and
- Proficiency levels on state-level assessments of students of teachers who received professional development.

GPRM Measures

Indicators for the MSP Program focus on evaluation design:

- The percentage of MSP projects that report using an **experimental or quasi-experimental design** for their evaluations.
- The percentage of MSP projects using an experimental or quasi-experimental design for their evaluations whose evaluations are conducted successfully and **yield scientifically valid results**.

Projects that meet the criteria are highlighted in the MSP annual report, and these studies become more visible to federal and state policy makers.

Rigorous Evaluation Criteria

The criteria identify four key elements for assessing whether the MSP evaluations were conducted in a rigorous manner (Westat, IES, and Abt Associate):

- ✓ Attrition
- ✓ Baseline Equivalence of Groups
- ✓ Quality of Measurement Instruments
- ✓ Relevant Statistics Reported

Outcomes

An evaluation may meet the criteria using any of the following outcomes:

- ***Teacher content knowledge:*** cell biology, angular momentum,, square waves, light and sound, quadratic equations, quantum mechanics,
- ***Classroom practices:*** the number of minutes a teacher spends on a topic, how often teachers engage students one-on-one, what the classroom environment looks like, technology integration,
- ***Student achievement:*** Measures of student achievement can include state and standardized tests.

Criterion #1: Attrition

An *experimental* evaluation meets the attrition criterion if the following two conditions are met:

- The overall attrition rate for the treatment and comparison groups is less than or equal to 30 percent,

AND

- The difference in the attrition rates between the two groups is equal to or less than 15 percent.

Attrition rate

The overall change in the sample size:

Group	N (baseline)	N (end of year)
Treatment	100	90
Control	100	66

overall attrition rate is: $[(100 + 100) - (90 + 66)] / (100 + 100) = 22\%$

differential attrition rate is: $[(100 - 66) / 100] - [(100 - 90) / 100] = 24\%$

A common practice among projects is to present varying sample sizes at different time points without an accompanying explanation. Another common mistake is to report the number of people in each group, but fail to report the number of people who were missing data for each outcome.

Criterion #2: Baseline Equivalence of Groups

One of the following two conditions must be met:

- **The difference between treatment and comparison group means on the outcome measure is less than or equal to 5 percent of the pooled standard deviation of the two groups,**

OR

- **The difference between treatment and comparison group means on the outcome measure is greater than 5 percent and less than 25 percent, and the analysis controls for the baseline differences in the analysis.**

Analytic Sample

Analytic Sample			
	N	Mean of Baseline Measure	Standard Deviation of Baseline Measure
Treatment Group	90	309	68
Comparison Group	66	312	59

Using the analytic sample, the difference in means is:

$$\Delta mean = |309 - 312| = 3.0$$

The pooled standard deviation (PSD) is:

$$PSD = \sqrt{\frac{(90 - 1)68^2 + (66 - 1)59^2}{90 + 66 - 2}} = 64.35$$

$$5\% \text{ of PSD} = 3.22, 25\% \text{ of PSD} = 16.09$$

The difference in means, 3.0, is less than 5 percent of the PSD, 3.22, and so this evaluation meets the baseline equivalence criterion and does not need to control for pre-test values.

Criterion #3: Quality of the Measurement Instruments

The Quality of Measurement Instruments criterion can be met in one of three ways:

- **Use existing instruments that have *already* been deemed valid and reliable, or**
- **Create a new instrument from an existing instrument(s) that has been validated and found to be reliable, or**
- **Create a new instrument and pre-test it with subjects comparable to the study sample or establish high reliability.**

Instruments

1- Existing assessments that have been shown to be reliable by developers and state tests are assumed to be valid and reliable for the purposes of MSP evaluations.

2- For existing instruments, grantees can refer to information on validity and reliability reported by other studies. Projects may also use subscales of existing instruments.

3- For new instruments developed from existing instruments, reliability do not need to be demonstrated if the following standards are met:

- At least 10 items are from the validated and reliable instrument(s), *and*
- At least 70 percent of the items on the new instrument are drawn from the validated and reliable instrument(s).

Examples

Teacher Content Knowledge in Mathematics

Learning Mathematics for Teaching (LMT)

Diagnostic Mathematics Assessments for Middle School Teachers

State Teacher Assessment

Knowledge of Algebra for Teaching

PRAXIS II

Teacher Content Knowledge in Science

MOSART: Misconception Oriented Standards-Based Assessment

Diagnostic Teacher Assessments in Mathematics and Science (DTAMS)

State Teacher Assessment

Assessing Teacher Learning about Science Teaching (ATLAST)

Force Concept Inventory

PRAXIS II

Classroom Practices

Reformed Teaching Observation Protocol (RTOP)

Surveys of Enacted Curriculum

Inside the Classroom Observation Protocol

Guskey's Model

Criterion #4: Relevant Statistics Reported

One of the following conditions must be met:

- **Post-test means for treatment and comparison groups and tests of statistical significance for key outcomes are presented. *Tests of statistical significance should directly compare the treatment and comparison groups, or***
- **Sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation and standard error of measurement) is presented.**

Analysis

When assessing differences between treatment and comparison groups for the evaluation, the two groups should be directly compared using an appropriate analytic strategy (ANOVA, ANCOVA, HLM or t-test).

Common Pitfalls

A common practice is to use results from the MSP TCK tool (included as part of the annual performance reporting system) to report on impacts