U.S. Department of Education Mathematics and Science Partnership Program

Guide for Summarizing MSP Evaluation Designs and Results

October 2008

Guide for Summarizing MSP Evaluation Designs and Results

One of the goals of the Mathematics and Science Partnership (MSP) program is to contribute to the knowledge base on effective professional development in mathematics and science. To this end, the MSP legislation (Title II, Part B of the No Child Left Behind Act) requires every MSP project to design and implement an evaluation and accountability plan that allows for a rigorous assessment of its effectiveness, and which includes information on the project's impact on teachers and students. In order to ensure that projects are providing high-quality information on program outcomes, the rubric, *Criteria for Classifying Designs of MSP Evaluations* (see Appendix A), was developed as part of the Data Quality Initiative through the Institute for Education Sciences (IES) at the U.S. Department of Education. The criteria that comprise the rubric specify conditions that projects that use experimental designs and quasi-experimental designs must meet in order to be deemed rigorous evaluations.

In 2008, the rubric was applied to the final evaluation reports of completed MSP projects for the first time. In doing so, it became apparent that most projects evaluate more than one component of their project (e.g., teacher content knowledge in mathematics and/or science, teacher attitudes and beliefs, student content knowledge in mathematics and/or science), that different evaluation techniques are often applied to the different components, and that some components meet all the criteria for being classified as a rigorous evaluation while other components do not. It also became apparent that while most projects collect most of the information needed to assess their evaluation design(s), few report the information in a manner that allows it to be easily evaluated with the rubric.

This *Guide* was developed to provide Project Directors and Evaluators with guidance on how best to summarize their evaluation data to facilitate the review and assessment of their evaluation design(s). We recommend that you present the results for each of the criteria discussed below in an Executive Summary at the beginning of your *final* evaluation report.

Screening Process

MSP evaluations undergo a two-stage screening process. They are first screened for the type of evaluation design and then for the strength of the implementation of the individual elements of the design. Below we present the criteria that are used in each stage of the screening process followed by recommended summary tables or narrative reporting guidelines, where relevant, for presenting information about your evaluation.

Evaluation Design

To be classified as having a strong design, only one component of the evaluation has to be either 1) an experimental study that compares the outcomes of a randomly assigned treatment and control group or 2) a quasi-experimental study that compares the outcomes of a treatment and comparison group that meets one of two design criteria:

- *comparison group study with equating*—statistical controls or matching techniques were used to make the treatment and comparison groups similar on their pre-intervention characteristics; or
- *regression-discontinuity study*—individuals (or other units such as classrooms or schools) were assigned to treatment or comparison groups on the basis of a "cutoff" score on a pre-intervention non-dichotomous measure.

Summary Information

List each outcome that your are evaluating and the participant group to whom it applies, and check the type of evaluation design applied to that group. *The table below provides an example of a project that evaluated five outcomes using three different designs.*

Table 1. Evaluation Design Type

Participant Group and Outcome	Experimental Design	Quasi-Experimental Design		Other Design
		with equating (matching)	regression discontinuity	
Elementary teachers science knowledge	x			
Elementary teachers mathematics knowledge	x			
Elementary students science achievement		x		
Elementary students mathematics achievement		x		
Elementary teacher classroom practice science				x
Elementary teacher classroom practice mathematics				x

Experimental Designs

For each participant group and outcome that was evaluated using an experimental design, please describe how units (i.e., participants, classroom schools, or districts) were randomly assigned to groups.

1. Participant Group/Outcome: _____: (*describe random assignment*)
2. Participant Group/Outcome: _____: (*describe random assignment*)
3. Participant Group/Outcome: _____: (*describe random assignment)*

Elements of the Design

To be classified a strong design each participant group/outcome that was evaluated using a quasi-experimental design must meet all of the following six criteria. Participant group/outcomes that were evaluated using an experimental design must meet every criterion except the first, baseline equivalence of groups, as randomly assigned groups are assumed to be equivalent at baseline.

A. Baseline Equivalence of Groups (quasi-experimental designs only)

Criterion:

- No significant pre-intervention differences between treatment and comparison group on variables related to the study's key outcomes; or
- Adequate steps were taken to address the lack of baseline equivalence in the statistical analysis.

Summary Information:

For each participant group and outcome provide the treatment and comparison groups' pre-test score (mean or percent) and the p-value of the statistical test used to assess equivalence.

Table 2. Baseline Equivalence of Groups			
Participant Group/Outcome and Matching Variables	**Treatment Group Pre-test Score**	**Comparison Group Pre-test Score**	**p-value**
Participant Group and Outcome: *Middle School Students/Middle School Mathematics*			
Student achievement	*mean or percent*	*mean or percent*	
Student demographic characteristics	*mean or percent*	*mean or percent*	
Participant Group and Outcome: *Middle Schools Students/Middle School Science*			
Student achievement	*mean or percent*	*mean or percent*	
Student demographic characteristics	*mean or percent*	*mean or percent*	
Participant Group and Outcome: *Middle School Teachers/Middle School Science*			
Teacher characteristics	*mean or percent*	*mean or percent*	
Participant Group and Outcome: *Middle School Teachers/Middle School Science*			
Teacher characteristics	*mean or percent*	*mean or percent*	

B. Sample Size

Criterion:

- Sample size was adequate based on a power analysis with recommended:
    - Significance level = 0.05
    - Power = 0.8
    - Minimum detectable effect informed by actual data; or

Absent a power analysis, a study will qualify as meeting the criterion in the following scenarios assuming the level of the intervention is the same as the unit of assignment or grouping (see Working Definitions for Projects in Appendix A for the assumptions that each scenario is based on).

Teacher Outcomes

- Case #1: For interventions at the school or district level, an evaluation would need a sample of at least 12 schools or districts.

- Case #2: For interventions at the teacher or classroom level, an evaluation would need a sample of at least 60 teachers

Student outcomes

- Case #1: For interventions at the school or district level, an evaluation would need a sample of at least 12 schools or districts.
- Case #2: For interventions at the teacher or classroom level**,** an evaluation would need a sample of at least 18 classrooms/teachers.
- Case #3: For interventions at the individual student level, an evaluation would need a sample of at least 130 students.

If the design is unbalanced (i.e., there are more treatment units than control/comparison or vice versa), the smaller of the two groups must at least meet the minimum sample size divided by 2. For example, for teacher outcomes Case #1, it is acceptable if there are 6 control/comparison schools and more than 6 treatment schools or vice versa.

Summary Information

For each participant group and outcome provide the *final* sample size at the level of random assignment or matching for the treatment and comparison/control group. Provide power calculation assumptions, if applicable.

Table 3. Sample Size			
Participant Group and Outcome	**Treatment Group (Final sample size)**	**Comparison/Control Group (Final sample size)**	**Power Analysis Findings (if applicable)**
Elementary teachers mathematics knowledge	*N*	*N*	alpha = power = MDE =
Elementary students science achievement	*N*	*N*	alpha = power = MDE =
Elementary students mathematics achievement	*N*	*N*	alpha = power = MDE =
Elementary teacher classroom practice science	*N*	*N*	alpha = power = MDE =
Recommended significant levels: alpha = 0.05, power = 0.8; minimal detectable effect (MDE) = informed by actual data.			

C. Quality of Measurement Instruments

Criterion:

- The study used existing data collection instruments that had already been deemed valid and reliable to measure key outcomes; or

- The study used data collection instruments developed specifically for the study that were sufficiently pre-tested with subjects who were comparable to the study sample.

Using selected items from a validated and reliable instrument or instruments is acceptable if the resulting instrument:

- Includes at least 10 items, *and*

- At least 70 percent of the items are from the validated and reliable instrument(s).

Summary Information

For each participant group and outcome, provide the name of the instrument that was used to measure the outcome and provide evidence of the instrument's validity and reliability. The evidence for borrowed or adapted instruments may be a website or other reference where the evidence is provided, or a narrative description of the evidence. For locally developed instruments that pre-tested the instruments, provide evidence of reliability and validity from those tests. For locally developed instruments that use items from one or more pre-existing valid and reliable instruments, provide the total number of items and the number of items borrowed from each instrument. *The table below provides examples of how to present data on different types of instruments.*

Table 4. Data Collection Instruments

Participant Group and Outcome	Name of Instrument	Evidence for Validity and Reliability
Teacher content knowledge – mathematics	*DTAMS*	*Cite website or other reference where evidence can be found.*
Teacher content knowledge – mathematics	*Locally developed instrument*	*Total items = 20* *NAEP items = 15* *LMT items = 5*
Teacher content knowledge – physics	*Locally developed instrument*	*Narrative description of evidence (e.g., Cronbach alpha, face validity).*
Teacher content knowledge - biology	*Locally developed instrument*	*Not tested for validity or reliability.*

D. Quality of the Data Collection Methods

Criterion:

- The methods, procedures, and timeframes used to collect the key outcome data from treatment and comparison groups were the same.

Summary Information

For each participant group and outcome, describe the method/procedure for collecting data from the treatment group, and indicate whether the same method/procedure was used to collect data from the comparison group. If the same method was not used, describe the method/procedure.

1. Participant Group and Outcome: _____

a. Method/procedure for collecting data from treatment group *(describe):*

b. Was the same method/procedure used to collect data from the comparison group? ____ Yes ___ No
If no, please describe how the method/procedure was different:

c. Time frame for data collection. Indicate the month and year that each test was administered to each group.

Table 5. Time Frame for Data Collect			
Participant Group and Outcome	**Month and Year**		
	Pre-test	**Post-test**	**Repeated Post-test**
Treatment group			
Comparison group			

2. Participant Group and Outcome: _____

a. Method/procedure for collecting data from treatment group *(describe):*

b. Was the same method/procedure used to collect data from the comparison group? ____ Yes ___ No
If no, please describe how the method/procedure was different:

c. Time frame for data collection. Indicate the month and year that each test was administered to each group.

Table 5. Time Frame for Data Collect			
Participant Group and Outcome	**Month and Year**		
	Pre-test	**Post-test**	**Repeated Post-test**
Treatment group			
Comparison group			

3. Participant Group and Outcome: _____

a. Method/procedure for collecting data from treatment group *(describe):*

b. Was the same method/procedure used to collect data from the comparison group? ____ Yes ___ No
If no, please describe how the method/procedure was different:

c. Time frame for data collection. Indicate the month and year that each test was administered to each group.

Table 5. Time Frame for Data Collection			
Participant Group and Outcome	**Month and Year**		
	Pre-test	**Post-test**	**Repeated Post-test**
Treatment group			
Comparison group			

E. Data Reduction Rates

There are two aspects to the data reduction criterion: attrition rates and response rates. An evaluation must meet the criterion for both attrition and response rates in order for it to meet the data reduction rates criterion. One exception is for cross-sectional studies that collect one-time data when only response rates apply. For longitudinal/pre-post studies that collect data from the same individuals over time, one needs to look at both the response rates and attrition rates criteria.

Criterion:

- The study measured the key outcome variable(s) in the post-tests for at least 70% of the original study sample (treatment and comparison groups combined)
- **Or** there is evidence that the high rates of data reduction were unrelated to the intervention; **AND**
- The proportion of the original study sample that was retained in the follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was similar for both the treatment and comparison groups (i.e., less than or equal to a 15% difference),
- **Or** the proportion of the original study sample that was retained in the follow-up data collection was different for the treatment and comparison groups, and sufficient steps were taken to address this differential attrition were not taken in the statistical analysis.

Summary Information

For each participant group and outcome, provide the following information for the treatment and comparison group: original sample size, pre-test sample size and the pre-test response rate (the percent of the pre-test sample that took the pre-test), post-test sample size and post-test response rate (the percent of the post-test sample that took the post-test), and the attrition rate, where the rate is calculated as the number of individuals who took *both* the pre- *and* post-test divided the number of individuals who took the post test.

Table 6. Data Reduction Rates

	Original Sample Size	**Pre-test Sample Size & Response Rate**	**Post-test Sample Size & Response Rate**	**Attrition Rate (for designs with pre-test)**
Participant Group and Outcome: *Elementary teachers science*				
Treatment group	*N*	*N, % responding*	*N, % responding*	*%*
Comparison group	*N*	*N, % responding*	*N, % responding*	*%*
Participant Group and Outcome: *Elementary teachers mathematics*				
Treatment group	*N*	*N, % responding*	*N, % responding*	*%*
Comparison group	*N*	*N, %*	*N,%*	*%*
Participant Group and Outcome: *Elementary students science*				
Treatment group	*N*	*N, % responding*	*N, % responding*	*%*
Comparison group	*N*	*N, % responding*	*N, % responding*	*%*
Participant Group and Outcome: *Elementary students mathematics*				
Treatment group	*N*	*N, % responding*	*N, % responding*	*%*
Comparison group	*N*	*N, % responding*	*N, % responding*	*%*

E. Relevant Statistics

Criterion:

- The final report includes treatment and comparison group post-test means and tests of significance for key outcomes; or

- Provides sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/standard error).

Summary Information

For each participant group and outcome, provide the following information for the treatment and comparison group: post-test sample size, mean or percent, and test of significance; or post-test sample size, mean or percent, and standard deviation (SD) or standard error (SE).

Table 7. Relevant statistics					
	Post-test N	**Mean or Percent**	**SD or SE**	**t, F, or chi square**	**p-value**
Participant Group and Outcome: *Elementary teachers science*					
Treatment group					
Comparison group					
Participant Group and Outcome: *Elementary teachers mathematics*					
Treatment group					
Comparison group					
Participant Group and Outcome: *Elementary students science*					
Treatment group					
Comparison group					
Participant Group and Outcome: *Elementary students mathematics*					
Treatment group					
Comparison group					

Appendix A

Criteria for Classifying Designs of MSP Evaluations[1]

- **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
- **Other**
  - The study uses a design other than a randomized controlled trial, comparison-group study with equating, or regression-discontinuity study, including *pre-post* studies, which measure the intervention's effect based on the pre-test to post-test differences of a single group, and comparison-group studies without equating, or non-experimental studies that compare outcomes of groups that vary with respect to implementation fidelity or program dosage.

[1] To be used for addressing following MSP GPRA measure: *The percentage of MSP projects that use an experimental or quasi-experimental design for their evaluations that are conducted successfully and that yield scientifically valid results.*

Criteria for Assessing whether *Experimental Designs* Were Conducted Successfully and Yielded Scientifically Valid Results

A. Sample size[2]

- **Met the criterion**—sample size was adequate (i.e. based on power analysis with recommended significance level=0.05, power=0.8, and a minimum detectable effect informed by the literature or otherwise justified).
- **Did not meet the criterion** —the sample size was too small
- **Did not address the criterion**

B. 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; or data collection instruments developed specifically for the study were sufficiently pre-tested with subjects who were comparable to the study sample
- **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**

C. Quality of the Data Collection Methods

- **Met the criterion**—the methods, procedures, and timeframes used to collect the key outcome data from treatment and control groups were the same
- **Did not meet the criterion**—instruments/assessments were administered differently in manner and/or at different times to treatment and control group participants

D. Data Reduction Rates (i.e. Attrition Rates, Response Rates)

- **Met the criterion**—(1) the study measured the key outcome variable(s) in the post-tests for at least 70% of the original study sample (treatment and control groups combined) or there is evidence that the high rates of data reduction were unrelated to the intervention, AND (2) the proportion of the original study sample that was retained in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was similar for both the treatment and control groups (i.e. less or equal to a 15-percent difference), or the proportion of the original study sample that was retained in the follow-up data collection was different for the treatment and control groups, but sufficient steps were taken to address this differential attrition in the statistical analysis.

---

[2] The critical sample size here is related to the unit of assignment. For example, if the assignment is made at the school level, the relevant sample size is the number of schools involved.

- **Did not meet the criterion**—(1) the study failed to measure the key outcome variable(s) in the post-tests for 30% or more of the original study sample (treatment and control groups combined), and there is no evidence that the high rates of data reduction were unrelated to the intervention; OR (2) the proportion of study participants who participated in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was significantly different for the treatment and control groups (i.e. more than a 15-percent difference) and sufficient steps to address differential attrition were not taken in the statistical analysis
- **Did not address the criterion**

E. Relevant Statistics Reported

- **Met the criterion**—the final report includes treatment and control group post-test means, and tests of statistical significance 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**

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

A. Baseline Equivalence of Groups

- **Met the criterion**—there were no significant pre-intervention differences between treatment and comparison group participants on variables related to the study's key outcomes; or adequate steps were taken to address the lack of baseline equivalence in the statistical analysis
- **Did not meet the criterion**—there were statistically significant pre-intervention differences between treatment and comparison group participants on variables related to the study's key outcomes; and no steps were taken to address lack of baseline equivalence in the statistical analysis
- **Did not address the criterion**

B. Sample size[3]

- **Met the criterion**—sample size was adequate (i.e. based on power analysis with recommended significance level=0.05, power=0.8, minimum detectable effect size informed by the literature or otherwise justified)
- **Did not meet the criterion** —the sample size was too small
- **Did not address the criterion**

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; or data collection instruments developed specifically for the study were sufficiently pre-tested with subjects who were comparable to the study sample
- **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. Quality of the Data Collection Methods

- **Met the criterion**—the methods, procedures, and timeframes used to collect the key outcome data from treatment and comparison groups were the same.

[3] The critical sample size here is related to the unit of grouping. For example, if the grouping is made at the school level, the relevant sample size is the number of schools involved.

- **Did not meet the criterion**—instruments/assessments were administered differently in manner and/or at different times to treatment and comparison group participants.

E. Data Reduction Rates (i.e. Attrition Rates, Response Rates)

- **Met the criterion**—(1) the study measured the key outcome variable(s) in the post-tests for at least 70% of the original study sample (treatment and comparison groups combined) or there is evidence that the high rates of data reduction were unrelated to the intervention, AND (2) the proportion of the original study sample that was retained in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was similar for both the treatment and comparison groups (i.e. less or equal to a 15-percent difference), or the proportion of the original study sample that was retained in the follow-up data collection was different for the treatment and comparison groups, and sufficient steps were taken to address this differential attrition were not taken in the statistical analysis.
- **Did not meet the criterion**—(1) the study failed to measure the key outcome variable(s) in the post-tests for 30% or more of the original study sample (treatment and comparison groups combined), and there is no evidence that the high rates of data reduction were unrelated to the intervention; OR (2) the proportion of study participants who participated in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was significantly different for the treatment and comparison groups (i.e. more than a 15-percent) and sufficient steps were not taken to address differential attrition in the statistical analysis.
- **Did not address the criterion**

F. Relevant Statistics Reported

- **Met the criterion**—the final report includes treatment and comparison group post-test means, and tests of statistical significance 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 did not include treatment and comparison group post-test means, 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**

MSP Rubric Working Definitions for Projects

The section contains working definitions to help interpret criteria in the *Criteria for Classifying Designs for MSP Evaluations* rubric.

Eligibility of evaluation report

Only final evaluation reports that contain post-test results on key outcomes will be evaluated. The review focuses exclusively on components regarding program impact, and does not cover assessment of implementation fidelity or performance against benchmarks.

Definition of an evaluation

An evaluation design may contain multiple outcomes. For the purpose of implementing this rubric, the major outcomes of interest are 1) teacher content knowledge, 2) teacher instructional practices, and 3) student achievement. The reviewer will apply each rubric criterion as it relates to the three outcomes separately.

Data collected on the three outcomes of interest might come from teachers/students in various grades and use different designs. If the implementation of the study design for an outcome meets all the criteria for at least one grade, the design for that outcome is considered as meeting the criteria. For example, if a study of 4th grade math achievement met the criteria but a study of 5th grade math did not, the student achievement evaluation from the project will be considered meeting the criteria based on the merit of its 4th grade math achievement study.

Baseline equivalence of groups

Variables related to key outcomes may vary. For example, if the key student outcome is achievement, the most relevant variable will be an achievement outcome from the same or similar test conducted prior to the intervention. Other related variables, although not equally effective, can be related to student socio-economic status. If the key outcome is teacher effectiveness, the most relevant variables will be measures of teacher effectiveness from the same or similar pre-test. Other related variables may include measures of teacher quality such as level of education and/or years of teaching experience.

Sample size

The sample size refers to the final sample size; that is the sample for which data have been collected.

Absent a power analysis, a study will qualify as “Met the criterion” in the following scenarios assuming the level of intervention is the same as the unit of assignment/grouping:

Teacher outcomes

- Case #1: For interventions at the school or district level**,** an evaluation would need a sample of at least 12 schools/districts based on following assumptions: 1) a balanced sampling design that randomizes/matches at the school/district level; 2) 0.05 level of significance in a two-tailed test; 3) a minimum detectable effect size of 0.50; 4) the power of the test is 0.8; 5) each school/district has at least 15 teachers; 6) intraclass correlation of 0.05; and 7) a school/district level covariate (i.e. aggregated pre-test score) explains 70 percent of the variation.

- Case #2: For interventions at the teacher or classroom level, an evaluation would need a sample of at least 60 teachers based on following assumptions: 1) a balanced sampling design that randomizes/matches at the teacher/classroom level; 2) 0.05 level of significance in a two-tailed test; 3) a minimum detectable effect size of 0.50; 4) the power of the test is 0.8; and 5) a teacher/classroom level covariate (i.e. pre-test score) explains 70 percent of the variation.

Student outcomes

- Case #1: For interventions at the school or district level, an evaluation would need a sample of at least 12 schools or districts based on following assumptions: 1) a balanced sampling design that randomizes/matches at school/district level; 2) 0.05 level of significance in a two-tailed test; 3) a minimum detectable effect size of 0.35; 4) the power of the test is 0.8; 5) each school or district has at least 75 students; 6) intraclass correlation of 0.05; and 7) a school/district level covariate (i.e. aggregated pre-test score) explains 70 percent of the variation.
- Case #2: For interventions at the teacher or classroom level**,** an evaluation would need a sample of at least 18 classrooms/teachers based on following assumptions: 1) a balanced sampling design that randomizes/matches at the classroom/teacher level; 2) 0.05 level of significance in a two-tailed test; 3) a minimum detectable effect size of 0.35; 4) the power of the test is 0.8; 5) each class has at least 25 students; 6) intraclass correlation of 0.05; and 7) a class/teacher level covariate (i.e. aggregated pre-test score) explains 70 percent of the variation.
- Case #3: For interventions at the individual student level, an evaluation would need a sample of at least 130 students based on following assumptions: 1) a balanced sampling design that randomizes/matches at the student level; 2) 0.05 level of significance in a two-tailed test; 3) a minimum detectable effect size of 0.35; 4) the power of the test is 0.8; and 5) a student level covariate (i.e. pre-test score) explains 70 percent of the variation.

If the design is unbalanced (i.e., there are more treatment units than control/comparison or vice versa), the smaller of the two groups must at least meet the minimum sample size divided by 2. For example, for teacher outcomes case #1, it is acceptable if there are 6 control/comparison schools and more than 6 treatment schools or vice versa.

Quality of measurement instruments

If the evaluators used an existing state accountability assessment or other widely used assessments (i.e. Iowa test, TerraNova) in totality one can assume that their psychometric properties are adequate. Using selected items from a validated and reliable instrument or instruments is acceptable if the resulting instrument includes at least 10 items and at least 70 percent of the items are from the validated and reliable instrument(s).

In addition, all instruments should at least have face validity.

Data reduction rates

There are two aspects to the data reduction criterion: attrition rates and response rates. An evaluation must meet the criterion for both attrition and response rates in order for it to meet the data reduction rates criterion. One exception is for cross-sectional studies that collect one-time data. For cross-sectional studies only response rates apply. For longitudinal/pre-post studies that collect data from the same subject over time, one needs to look at both the response rates and attrition rates criteria.

If not provided in the report, the rates can be loosely calculated a) attrition rates b) response rates:

a. Posttest N/ Pretest N
b. Posttest N/ Original N

The first component of the criterion refers to overall data reduction and the second is related to differential reduction (i.e., between treatment and control/comparison groups).

If the 70-percent data retention rate is not met, an evaluation may meet the criterion if the evaluators provide valid explanations (e.g., the schools are located in high mobility areas) or have addressed potential differences between sample members who have post-test data and those who do not in the analysis.

References

Raudenbush, S.W., Spybrook, J., Liu, X, and Cogndon, R. (2005). Optimal design for longitudinal and multilevel research. Versio