

# Academic Intervention Rating Rubric

## Group Design

Participants (Group Design)	
<b>Are the students in the study at risk?</b>	
<b>Full Bubble</b>	Evidence is convincing that participants were at risk with respect to the focus of the intervention (i.e., below 30th percentile on local or national norm; or sample mean below 25th percentile on local or national test; non-response to Tiers 1 and 2 <sup>1</sup> ; or students with identified disability related to the focus of the intervention).
<b>Empty Bubble</b>	Fails full bubble.
Design (Group Design)	
<b>Does the study design allow us to conclude that the intervention program, rather than extraneous variables, was responsible for the results?</b>	
<b>Full Bubble</b>	Random assignment was used. At pretreatment, program and control groups had a mean standardized difference that fell within 0.25 SD on measures used as covariates or on pretest measures also used as outcomes, and on demographic measures (or if the mean difference is above .25 SDs, the difference was controlled for in the analyses and there was no differential attrition in the sample). There was no attrition bias <sup>2</sup> . Unit of analysis matched random assignment (controlling for variance associated with potential dependency at higher levels of the unit of randomization is permitted, e.g., controlling for variance at the classroom level).
<b>Half Bubble</b>	Random assignment was used but other conditions for full bubble not met. OR Random assignment was not used but a strong quasi-experimental design was used. At pretreatment, program and control groups had a mean standardized difference that fell within 0.25 SD on measures central to the study (i.e., pretest measures also used as outcomes) and demographic measures, and outcomes were analyzed to adjust for pretreatment differences. There was no attrition bias <sup>2</sup> . Unit of analysis matched assignment strategy.
<b>Empty Bubble</b>	Fails full and half bubble.

<sup>1</sup> Non-response to Tiers 1 and 2 is applicable for interventions studied in settings in which a behavioral tiered intervention system is in place, and the student has failed to meet the school's or district's criteria for "response" to both Tier 1 (schoolwide/universal program) and Tier 2 (Tier 2 or secondary behavioral intervention) supports. Detailed information about these non-response criteria should be included in the study description.

<sup>2</sup> NCI follows guidance from the What Works Clearinghouse (WWC) in determining attrition bias. The WWC model for determining bias based on a combination of differential and overall attrition rates can be found on pages 11-13 of this document: [https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc\\_procedures\\_v3\\_0\\_standards\\_handbook.pdf](https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf)



### Fidelity of Implementation (Group Design)

#### Was it clear that the intervention program was implemented as it is designed to be used?

<b>Full Bubble</b>	Measurement of fidelity of implementation was conducted adequately* and observed with adequate intercoder agreement (e.g., between 0.8 and 1.0) or permanent product, and levels of fidelity indicate that the intervention program was implemented as intended (e.g., at least 75% for a single measure, or a reasonable average across multiple measures).
<b>Half Bubble</b>	Levels of fidelity indicate that the intervention program was implemented as intended (e.g., at least at 75% for a single measure, or a reasonable average across multiple measures), but measurement of fidelity of implementation either was not conducted adequately or was not observed with adequate intercoder agreement. OR Measurement of fidelity of implementation was conducted adequately and observed with adequate intercoder agreement (e.g., between 0.8 and 1.0) or permanent product, but levels of fidelity are moderate (e.g., an average below 60% across multiple measures, 60%-75% for a single measure).
<b>Empty Bubble</b>	Fails full and half bubble.

\*In determining whether measurement of fidelity of implementation was conducted adequately, the TRC will consider the following:

- clear and comprehensive rationale for the indicators making up the implementation measures, that reflects what the intervention developers believe are the active intervention ingredients;
- the number of times implementation is measured;
- the extent to which implementation fidelity observers are independent of the intervention development team.



Measures (Group Design)		
Were the study measures accurate and important?		
	Targeted <sup>3</sup> Outcome Measures	Broader <sup>4</sup> Outcome Measures
<b>Full Bubble</b>	Targeted measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Each targeted measure was psychometrically reliable* (i.e., all internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).	Broader measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Each broader measure was psychometrically reliable (i.e., all internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).
<b>Half Bubble</b>	Targeted measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most targeted measures were psychometrically reliable (i.e., most internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).	Broader measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most broader measures were psychometrically reliable (i.e., most internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).
<b>Empty Bubble</b>	Fails full and half bubble.	Fails full and half bubble.
<b>Dash</b>	No targeted measures used in the study.	No broader measures used in the study.

\* For standardized measures, empirical evidence for the quality of the measure does not need to be based on the current sample, but instead psychometric evidence from validation samples (e.g., sample information found in technical manuals) can be reported.

<sup>3</sup> Targeted measures assess aspects of competence the program was directly targeted to improve. Typically, this does not mean the very items taught but rather novel items structured similarly to the content addressed in the program. For example, if a program taught word attack, a targeted measure would be decoding of pseudowords. If a program taught comprehension of cause-effect passages, a targeted measure would be answering questions about cause-effect passages structured similarly to those used during intervention, but not including the very passages used for intervention.

<sup>4</sup> Broader measures assess aspects of competence that are related to the skills targeted by the program but not directly taught in the program. For example, if a program taught word-level reading skill, a broader measure would be answering questions about passages the student reads. If a program taught calculation skill, a broader measure would be solving word problems that require the same kinds of calculation skill taught in the program.



## Effect Size

### *(Group Design)*

The effect size is a measure of the magnitude of the relationship between two variables. Specifically, on this chart, the effect size represents the magnitude of the relationship between participating in a particular intervention and an academic outcome of interest. The larger the effect size, the greater the impact that participating in the intervention had on the outcome. Furthermore, a positive effect size indicates that participating in the intervention led to improvement in performance on the academic outcome measure, while a negative effect size indicates that participating in the intervention led to a decline in performance on the academic outcome measure. According to guidelines from the What Works Clearinghouse, an effect size of .25 or greater is considered to be “substantively important.” Additionally, we note on this tools chart those effect sizes which are statistically significant. Effect sizes that are statistically significant can be considered more trustworthy than effect sizes of the same magnitude that are not statistically significant.

There are many different methods for calculating effect size. In order to ensure comparability of effect size across studies on this chart, the NCII follows guidance from the What Works Clearinghouse and uses a standard formula to calculate effect size across all studies and outcome measures—Hedges  $g$ , corrected for small-sample bias:

$$\frac{(\text{Posttest mean for program group} - \text{Posttest mean for control group})}{\text{Pooled unadjusted posttest standard deviation}} \times \sqrt{\frac{1}{34N-9}}$$

Developers of programs on the chart were asked to submit the necessary data to compute the effect sizes. Where available, the NCII requests **adjusted** posttest means, which refers to posttests that have been adjusted to correct for any pretest differences between the program and control groups. In the event that developers are unable to access or report adjusted means, the NCII will calculate and report effect size based on unadjusted posttests. However, **unadjusted** posttests are typically reported only in instances in which we can **assume pretest group equivalency**. Therefore, the default effect size reported will be Hedges  $g$  based on adjusted posttest means. NCII will only report effect size based on unadjusted posttests for studies that (a) are unable to provide adjusted means, and (b) have pretest differences on the measure that fall within .25SD and are not statistically significant.<sup>5\*</sup>

<sup>5</sup> An exception to this rule will only be made if vendors establish a link between an instrument administered only at posttest and a comparable instrument administered at pretest. If Center staff verify that the pretest and posttest measures assess the same construct and that there were negligible between-group differences in this domain (pretest ES fell within 0.25 SDs and



The chart includes, for each study, the number and type of outcomes measures, and, for each type of outcome measure, a mean effect size. Additionally, for some studies, effect sizes are reported for one or more disaggregated sub-samples. By clicking on any of the individual effect size cells, users can see a full list of effect sizes for each measure used in the study. Studies that include a “—” in the effect size cell either do not have the necessary data or do not meet the assumptions required for calculating and reporting effect size using the associated formula. The reason for the missing data is provided when users click on the cell.

---

was statistically insignificant), NCII will attempt to calculate a difference-in-differences adjusted ES. For further details, see Appendix F (pages F.4-F.5) of the current [WWC Procedures Handbook](#). If you believe that one of your outcome measures is a suitable “proxy” for another measure, please indicate this in your submission or contact us ([ToolsChartHelp@air.org](mailto:ToolsChartHelp@air.org)) and explain your thinking.



## Single Subject Design

### Participants (Single Subject Design)

#### Are the students in the study at risk?

**Full Bubble** Evidence is convincing that participants were at risk with respect to the focus of the intervention (i.e., below 30th percentile on local or national norm; or sample mean below 25th percentile on local or national test; non-response to Tiers 1 and 2<sup>6</sup>; or students with identified disability related to the focus of the intervention).

**Empty Bubble** Fails full bubble.

### Design (Single Subject Design)

#### Does the study design allow us to evaluate experimental control?

**Full Bubble** The study includes three data points or sufficient number to document a stable performance within that phase. There is the opportunity for at least three demonstrations of experimental control.\*

**Half Bubble** The study includes one or two data points within a phase. There is the opportunity for two demonstrations of experimental control. Or, the study is a non-concurrent multiple baseline design.

**Empty Bubble** Fails full and half bubble.

\* For alternating treatment designs, five repetitions of the alternating sequence are required for a full bubble, and four are required for a half bubble.

### Fidelity of Implementation (Single Subject Design)

#### Was it clear that the intervention program was implemented as it is designed to be used?

**Full Bubble** Measurement of fidelity of implementation was conducted adequately and observed with adequate intercoder agreement, and levels of fidelity indicate that the intervention program was implemented as intended (at 75% or above).

**Half Bubble** Levels of fidelity indicate that the intervention program was implemented as intended (at 75% or above), but measurement of fidelity of implementation either was not conducted adequately or was not observed with adequate intercoder agreement.

**Empty Bubble** Fails full and half bubble.

\*In determining whether measurement of fidelity of implementation was conducted adequately, the TRC will consider the following:

<sup>6</sup> Non-response to Tiers 1 and 2 is applicable for interventions studied in settings in which an academic tiered intervention system is in place, and the student has failed to meet the school's or district's criteria for "response" to both Tier 1 (schoolwide/universal program) and Tier 2 (Tier 2 or secondary academic intervention) supports. Detailed information about these non-response criteria should be included in the study description.



- clear and comprehensive rationale for the indicators making up the implementation measures, that reflects what the intervention developers believe are the active intervention ingredients;
- the number of times implementation is measured; and
- the extent to which implementation fidelity observers are independent of the intervention development team.

Measures (Single Subject Design)		
Were the study measures accurate and important?		
	Targeted <sup>7</sup> Outcome Measures	Broader <sup>8</sup> Outcome Measures
<b>Full Bubble</b>	Targeted measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Each targeted measure was psychometrically reliable* (i.e., all internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).	Broader measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Each broader measure was psychometrically reliable (i.e., all internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).
<b>Half Bubble</b>	Targeted measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most targeted measures were psychometrically reliable (i.e., most internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).	Broader measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most broader measures were psychometrically reliable (i.e., most internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).
<b>Empty Bubble</b>	Fails full and half bubble.	Fails full and half bubble.
<b>Dash</b>	No targeted measures used in the study.	No broader measures used in the study.

\* For standardized measures, empirical evidence for the quality of the measure does not need to be based on the current sample, but instead psychometric evidence from validation samples (e.g., sample information found in technical manuals) can be reported.

## Results (Single Subject Design)

<sup>7</sup> Targeted measures assess aspects of competence the program was directly targeted to improve. Typically, this does not mean the very items taught but rather novel items structured similarly to the content addressed in the program. For example, if a program taught word attack, a targeted measure would be decoding of pseudowords. If a program taught comprehension of cause-effect passages, a targeted measure would be answering questions about cause-effect passages structured similarly to those used during intervention, but not including the very passages used for intervention.

<sup>8</sup> Broader measures assess aspects of competence that are related to the skills targeted by the program but not directly taught in the program. For example, if a program taught word-level reading skill, a broader measure would be answering questions about passages the student reads. If a program taught calculation skill, a broader measure would be solving word problems that require the same kinds of calculation skill taught in the program.



---

**Does visual analysis of the data demonstrate evidence of a relationship between the independent variable and the primary outcome of interest?**

<b>Full Bubble</b>	Visual or other analysis demonstrates clear, consistent, and meaningful change in pattern of data as a result of intervention (level, trend, variability, immediacy). The number of data points is sufficient to demonstrate a stable level of performance for the dependent variable; there are at least three demonstrations of a treatment effect*, and no documented non-demonstrations.
<b>Half Bubble</b>	Visual or other analysis demonstrates minimal or inconsistent change in pattern of data. There were two demonstrations of a treatment effect and no documented non-effects, or the ratio of effects to non-effects was less than or equal to 3:1.
<b>Empty Bubble</b>	Visual analysis demonstrates no change in pattern of the data. Fails full and half bubble.

\*In determining demonstration of a treatment effect, the TRC will consider the following:

1. Do the baseline data document a pattern in need of change?
2. Do the baseline data demonstrate a predictable baseline pattern?
  - a. Is the variability sufficiently consistent?
  - b. Is the trend either stable or moving away from the therapeutic direction?
3. Do the data within each phase non-baseline document a predictable data pattern?
  - a. Is the variability sufficiently consistent?
  - b. Is the trend either sufficiently low or moving in the hypothesized direction (i.e., away from anticipated treatment effects during baseline conditions and towards treatment effects in intervention conditions)?
4. Does between phase data document the presence of basic effects?
  - a. Is the level discriminably different between the first and last three data points in adjacent phases?
  - b. Is the trend discriminably different between the first and last three data points in adjacent phases?
  - c. Is there an overall level change between baseline and treatment phases?
  - d. Is there an overall change in trend between baseline and treatment phases?
  - e. Is there an overall change in variability between baseline and treatment phases?
  - f. Is there sufficiently low overlap between baseline and treatment phases to document an experimental effect?
  - g. Do the data patterns in similar phases (e.g., intervention-to-intervention) demonstrate similar patterns? (Only applicable to reversal designs or embedded probe designs)





This resource was produced under U.S. Department of Education, Office of Special Education Programs, Award No. H326Q210001. Celia Rosenquist serves as the project officer. The views expressed herein do not necessarily represent the positions or policies of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this document is intended or should be inferred.

