Academic Intervention Rating Rubric

Group Design

Participants (Group 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; or students with identified disability related to the focus of the intervention).
Empty Bubble	Fails full bubble.

Design (Group Design)	
Does the study design allow us to conclude that the intervention program, rather than extraneous variables, was responsible for the results?	
Full Bubble	Students were randomly assigned. At pretreatment, program and control groups were not statistically significantly different; and 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. There was no attrition bias[1]. 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., for randomizing at the student level, controlling for variance at the classroom level).
Half Bubble	Students were randomly assigned but other conditions for full bubble not met. OR Students were not randomly assigned but a strong quasi-experimental design was used. At pretreatment, program and control groups were not statistically significantly different and 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. Unit of analysis matched assignment strategy.

[1] NCII 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

Design (Group Design)	
Empty Bubble	Fails full and half bubble.

Fidelity of Implementation (Group 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.

Measures (Group Design)		
Were the study measures accurate and important?		
	Targeted[2]	**Broader**[3]
Full Bubble	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).
Half Bubble	Targeted measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most targeted measures were psychometrically reliable (i.e., most	Broader measure(s) appropriately represented outcome(s) relevant to the program's instructional content. Most broader measures were psychometrically reliable (i.e., most

[2] 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.

[3] 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.

Measures (Group Design)		
Were the study measures accurate and important?		
	Targeted[2]	**Broader**[3]
	internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).	internal consistency coefficients > 0.59; interscorer agreement not accepted for measures other than writing).
Empty Bubble	Fails full and half bubble.	Fails full and half bubble.

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*[4], 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:

$$\left(\frac{\text{Posttest mean for program group} - \text{Posttest mean for control group}}{\text{Pooled unadjusted posttest standard deviation}}\right) * \left(1 - \frac{3}{4N - 9}\right)$$

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 pre-and posttest unadjusted mean differences. However, the unadjusted mean differences are 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 the unadjusted mean differences for studies (a) that are unable to provide adjusted means, ***and*** (b) whose pretest differences on outcome measures are not statistically significant and fall within 0.25 standard deviations. Note also that the NCII will not be able to report effect size on any variable for which only posttest data are known because of the need for pretests in calculating adjusted posttest scores.

[4] See pages 22-24 of this document: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

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.

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 (i.e., below 30th percentile on local or national norm; or sample mean below 25th percentile on local or national test; or students with identified disability).
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.

Measures (Single Subject Design)		
Were the study measures accurate and important?		
	Targeted[5]	**Broader**[6]
Full Bubble	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).
Half Bubble	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).
Empty Bubble	Fails full and half bubble.	Fails full and half bubble.

Results (Single Subject Design)	
Does visual analysis of the data demonstrate evidence of a relationship between the independent variable and the primary outcome of interest?	
Full Bubble	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.

[5] 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.

[6] 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.

Results (Single Subject Design)	
Half Bubble	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.
Empty Bubble	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)