Directions for Practice Scoring Instrument

This scoring instrument automatically calculates the final scores for each dimension on the instrument (conceptual framework, design quality, outcome evidence, practice fidelity, and external validity). In addition, the final rating for each study (Class 1-Class 5) will also be automatically calculated after scoring it in Part 2 of the instrument.

1. The instrument is in a Read-Only format so, in order to fill it out, the first thing to do is save a copy of the instrument on your hard drive.

2. The tabs at the bottom of the spreadsheet clearly label each part of the instrument. If you receive only one study to score, then only fill out the spreadsheet labeled Part 2_Study 1. If you have more than one study to score, fill out Part 2_Study 2, and so forth for each study.

3. For each item in the instrument, select one score from the dropdown menu located under Rating (click in the rating box to see the dropdown box). For both Part I and Part 2, the scores will automatically be filled in the scoring tables at the end of each dimension. The scoring tables calculate the final score for each dimension. The final scores are then automatically filled in on the overall score table at the end of Part 2.

4. In the Outcome Evidence section, under item A-Substantive Practice Effects, only fill in a score under Unweighted Score (either 3, 2, 1, or 0) and Direction (-1 or 1). The Weighted Score and the final Substantive Practice Effects Score will be automatically calculated.

   Please note: Score only 5 primary outcomes and 5 secondary outcomes. Do not add any additional rows to score additional outcomes, as this will affect the final score calculation that is automatically calculated in the scoring table. Contact your Senior Researcher for clarification if you believe there are more than 5 primary and 5 secondary outcomes that should be scored.

5. At the end of Part 3, there is a signature box. Be sure to type in your first and last name in the signature box before sending the instrument back to the NMRC Post-Doctoral Research Associate. (This serves as a proxy for your electronic signature.)
CRIME SOLUTIONS RESOURCE CENTER: PROGRAM RATING INSTRUMENT--PART 1

Instructions: Please carefully assess the practice in terms of the conceptual framework. The reviewer should complete Part 1 only once for each practice, regardless of the number of studies to be reviewed. Complete this section by using pertinent information from the studies provided as well as your own knowledge of the literature as it pertains to each topic (e.g., Prior Research). Please record your answers on this form.

PRACTICE NAME:

REVIEWER’S NAME ___________________________ DATE OF REVIEW ___________________________

CONCEPTUAL FRAMEWORK

A. Prior Research assesses the degree to which previous empirical evidence (formal evaluations and meta-analyses) provides supports for practices that are comparable to the practice being reviewed. It is important to note that the scope of comparable practices will vary by practice. Take, for instance, the practice of providing ongoing (post-match) training to mentors. Because there is a reasonable amount of research on related types of mentoring program practices, such as pre-match mentor training and staff supervision of matches, the scope of comparable practices could be limited to these types of practices. On the other hand, consider a practice that is relatively novel in its approach such as youth initiated mentoring. In this case, the scope of comparable practices could be widened to include other similar practices as implemented in a broader range of interventions than only mentoring programs. Finally, please note that research on the effectiveness of the practice being reviewed should not be considered in scoring this item since this research will be the focus of Part II of the scoring instrument.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Points and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>High (5 or more other studies, or 1 meta-analysis, provide evidence in support of the practice).</td>
</tr>
<tr>
<td>2</td>
<td>Medium (2 to 4 other studies provide evidence in support of the practice).</td>
</tr>
<tr>
<td>1</td>
<td>Low (1 other study provides evidence in support of the practice).</td>
</tr>
<tr>
<td>0</td>
<td>None (No other studies provide evidence in support of the practice).</td>
</tr>
</tbody>
</table>

B. Theoretical Base measures the degree to which the practice is based on a well-articulated, conceptually sound theory that is logically connected to the characteristics of the intended users and recipients, program settings and structures, and outcomes. Some practices are designed with little regard to conceptual development other than an implicit appeal to common sense. Instead, there should be an explanation provided of why and how the practice is expected to achieve its intended results and this explanation should be supported by prior conceptual development and empirical research. The emphasis in assigning this rating should be on the theoretical basis for the practice at a general level rather than for a particular form or instance of the practice (e.g., training mentors before they begin their matches rather than training them using particular methods or with an emphasis on particular types of content or skills).

<table>
<thead>
<tr>
<th>Rating</th>
<th>Points and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Practice theory is fully described and conceptually sound.</td>
</tr>
<tr>
<td>2</td>
<td>Practice theory is adequately described and appears conceptually sound.</td>
</tr>
<tr>
<td>1</td>
<td>Very little information is provided about practice theory, but it may be conceptually sound.</td>
</tr>
</tbody>
</table>
C. Practice Description rates the degree to which the defining details of the practice are evident. Defining details can be inferred from their consistent presence across different applications of the practice. Typically, these would include the following information: 1) key activities and content of the practice, 2) frequency and duration of the activities associated with the practice, 3) the targeted population, 4) the targeted outcome(s) (i.e., the intent of the practice), and 5) the setting.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Points and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3=</td>
<td>All defining details of the practice details are evident.</td>
</tr>
<tr>
<td>2=</td>
<td>Most defining details of the practice are evident.</td>
</tr>
<tr>
<td>1=</td>
<td>Some defining details of the practice are evident.</td>
</tr>
<tr>
<td>0=</td>
<td>No defining details of the practice are specified.</td>
</tr>
</tbody>
</table>

Notes: Please specify the targeted population, the targeted behaviors, and the key elements of the practice:
### Conceptual Framework Scoring Table

<table>
<thead>
<tr>
<th></th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Research Points</td>
<td>0</td>
</tr>
<tr>
<td>+ Theoretical Base Points</td>
<td>0</td>
</tr>
<tr>
<td>+ Practice Description Points</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>0</td>
</tr>
</tbody>
</table>

/ Number of Items 3

= Conceptual Framework Score 0.00
A. RESEARCH DESIGN: note the ability of the design to infer a causal relationship between the practice and outcomes. There are three general types of designs: experimental, quasi-experimental, and non-experimental. The designs differ in the method of assignment. A randomized field experiment randomly sorts participants into two or more groups. One group receives the practice (treatment), while the other (control) does not. A quasi-experimental research design is similar to the experimental design with the exception that the subjects are assigned to the treatment and comparison groups through a process that is not random. Finally, a non-experiment lacks a comparable control group. Since these designs differ in assignment strategy, it is likely they will differ in terms of their strength with respect to internal validity. Designs in which the two groups differ in their receipt of multiple practices (i.e., the practice of interest and other practices) should be rated as non-experimental because of the inability to infer a causal effect of the practice of interest in such designs (Note: Not all designs fall into this hierarchy. The reviewer should define the design and note the reason for the scores). In some evaluations, comparisons will be made between those who received a practice and those who did not based on unplanned factors relating to program implementation or participant compliance (e.g., in programs where all mentors are intended to receive pre-match training, only some do and the primary comparison then becomes mentors who do and do not receive the training). In such cases, comparisons would be considered quasi-experiments, but not as strong in this regard as quasi-experimental comparisons that are planned (e.g., mentors in one set of programs are trained and those in another set are not by design), and thus a score of 1 is likely most appropriate.

B. SAMPLE SIZE (POWER): assesses the adequacy of the sample to detect meaningful effects of the practice. However, the optimal size of a sample is not easily straightforward. Statistical power is a function of several factors: 1) the size of the sample, 2) the magnitude of the expected effect, 3) the type of statistical test used, and 4) the alpha level set to control Type I error (conventionally set at .05). In general, for a traditional two-group experiment with a statistical power of .80, the N should be roughly 394 per group to detect a small effect (d=.20), 64 to detect a medium effect (d=.50), and 20 to detect a large effect (d=.80). It should be noted, however, that these figures are intended only as guidelines to help direct the review. Furthermore, as detailed in the guidelines for assigning different scores below, separate rules of thumb apply for meta-analyses and when analyses are conducted at a program or site level (Note: The same sample size guidelines are relevant for both traditional experiments and time series designs). Most textbooks suggest that about 50 observations, with a reasonable distribution among pre- and posttest measurements, is required for a competent analysis, on grounds that this figure is usually sufficient for random error. Conversely, although it may not account for the randomness of the data, roughly 15 observations are generally considered the minimum. The reviewer should use No, or not the expert, to assess the adequacy of the sample.

C. STATISTICAL ADJUSTMENT: assesses the use of statistical controls to account for the initial measured differences between the groups. Any outcome-relevant variable on which the groups may differ should be identified and included in the statistical adjustment.

D. INSTRUMENTATION: rates the quality (reliability and validity) of the measures used in the evaluation of the practice – that is, both the measure of the practice itself (in particular, the assessment of whether the practice was received/implemented or not; more nuanced assessment of quality of implementation should instead be considered under Practice Fidelity) and outcomes examined in the practice (only score measures of control variables and outcomes listed under Outcome Evidence). Reliability refers to the stability and consistency of the measures. Validity refers to the accuracy of the measure. The selection of appropriate instrumentation should also consider the developmental and cultural appropriateness of the measures, as well as the reading level, native language, and attention span of the participants.

---

### Table: Design Quality

<table>
<thead>
<tr>
<th>Rating</th>
<th>Public and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= Experimental (well-designed randomized field trial). Participants are randomly assigned into a practice group and a non-practice group.</td>
<td></td>
</tr>
<tr>
<td>2= Quasi-experimental Level 1 (design uses a credit card comparison group of participants who receive the practice and participants who do not).</td>
<td></td>
</tr>
<tr>
<td>1= Quasi-experimental Level 2 (design has a comparison group but lacks comparability on important preexisting variables or lacks information on pre-treatment equivalence of groups; time series single group design).</td>
<td></td>
</tr>
<tr>
<td>0= Non-experimental level 1 (one group pretest-posttest, one- and two-group posttest only, or case studies).</td>
<td></td>
</tr>
</tbody>
</table>

### Table: Sample Size (Power)

<table>
<thead>
<tr>
<th>Rating</th>
<th>Public and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= High Power: The sample is sufficient to detect a small effect (d=.20) using appropriate tests. (In general, the N should be greater than 394 per group in a traditional experiment and greater than 75 in a time series design; in the case of a meta-analysis or when analyses are conducted at the program or site level, the number of studies or sites should generally be greater than 40 with more required if sample sizes are small [e.g., less than 100] and/or the numbers of studies or sites with and without the practice under review are markedly disproportionate).</td>
<td></td>
</tr>
<tr>
<td>2= Medium Power: The sample is sufficient to detect a medium effect (d=.50) using appropriate tests. (In general, the N should be between 64 and 393 per group in a traditional experiment and between 51 and 75 in a time series design. In case of a meta-analysis or program/site-level evaluation, the number of studies or sites should generally be between 20 and 40 with more required if sample sizes are small [e.g., less than 100] and/or the numbers of studies or sites with and without the practice under review are markedly disproportionate).</td>
<td></td>
</tr>
<tr>
<td>1= Low Power: The sample is sufficient to detect a large effect (d=.80) using appropriate tests. (In general, the N should be between 20 and 63 per group in a traditional experiment and between 15 and 50 in a time series design; in the case of a meta-analysis or program/site-level evaluation, the number of studies or sites should generally be between 10 and 20 with more required if sample sizes are small or the numbers of studies or sites with and without the practice under review are markedly disproportionate).</td>
<td></td>
</tr>
<tr>
<td>0= Insufficient: The sample is not sufficient to detect an effect. (In general, the N is less than 25 per group in a traditional experiment and less than 15 in a time series design; in the case of a meta-analysis or program/site-level evaluation, the number of studies or sites is less than 15).</td>
<td></td>
</tr>
</tbody>
</table>

### Table: Statistical Adjustment

<table>
<thead>
<tr>
<th>Rating</th>
<th>Public and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= No statistical adjustments required in the analysis. Random assignment or selection modeling (properly score matching) with a sufficiently large sample resulted in no group differences.</td>
<td></td>
</tr>
<tr>
<td>2= The analysis employs appropriate statistical adjustments (includes control variables that are presumed to be related to the outcome) to control for group differences.</td>
<td></td>
</tr>
<tr>
<td>1= The analysis employs statistical adjustments (includes control variables that are presumed to be related to the outcome) but some important variables are not addressed.</td>
<td></td>
</tr>
<tr>
<td>0= Not applicable.</td>
<td></td>
</tr>
</tbody>
</table>

### Table: Instrumentation

<table>
<thead>
<tr>
<th>Rating</th>
<th>Public and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= Excellent: The reliability (the extent to which an item produces the same results when used repeatedly) and validity (the extent to which an item measures what it is intended to measure) of the measures are excellent.</td>
<td></td>
</tr>
<tr>
<td>2= Adequate: The reliability (the extent to which an item produces the same results when used repeatedly) and validity (the extent to which an item measures what it is intended to measure) of the measures are adequate.</td>
<td></td>
</tr>
</tbody>
</table>
E. INTERNAL VALIDITY assesses the degree to which the observed changes infer a causal relationship with the practice. The internal validity of a study depends on both the research design and the measurement of the practice. Threats to internal validity will affect the accuracy of the results and draw into question the effect of the practice.

Please check the specific threats to validity in the table on the next page and include notes.

<table>
<thead>
<tr>
<th>Threat</th>
<th>Points</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= No threats to internal validity are identified or all threats have been adequately addressed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2= Marginal threats to internal validity are identified and remain.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1= Moderate threats to internal validity are identified and remain.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0= Serious threats to internal validity are identified and remain.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F. FOLLOW-UP PERIOD assesses the length of time that the study period continues after the implementation or delivery of the relevant practice has ended so as to ascertain the sustained effects of the practice. It should be kept in mind that the end of the practice may often differ from the end of the program (e.g., mentor pre-match training). If the practice was still being implemented or delivered at the end of the study period, then a score of 1 should be assigned.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Points and Description</th>
<th>Specify follow-up period in months:</th>
</tr>
</thead>
<tbody>
<tr>
<td>3= More than 1 year.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2= More than 6 months but less than or equal to 1 year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1= Less than or equal to 6 months.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0= Not specified.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DESIGN QUALITY SCORING TABLE

| Research Design Points | 0 |
| Sample Size Points     | 0 |
| Statistical Adjustment Points | 0 |
| Instrumentation Points | 0 |
| Internal Validity Points | 0 |
| Follow-Up Period Points | 0 |
| TOTAL                   | 0 |

\[
\text{TOTAL} = \frac{\text{NUMBER OF ITEMS}}{6} \\
= \text{Design Quality Score} \\
= 0.88
\]
A. Substantive Practice Effects: rates the presence, strength, and direction of effects of the practice. The primary outcomes generally will be those that relate directly to one of the major areas of Crime Solutions (reducing crime/delinquency, improving the justice system, responding to victims, etc.) as well as those that reflect changes in youth behavior. Secondary outcomes will relate less directly to one of the major areas of Crime Solutions and will include both changes in youth attitudes as well as characteristics of mentoring relationships. Scores for primary outcomes are given three times the weight of secondary outcomes. Use the following scale to assess the practice’s achievement of each of the outcomes:

3: The finding provides very strong evidence of an effect of the practice (significant finding – i.e., p < .05, two-tailed; large effect)
2: The finding provides moderate evidence of an effect of the practice (significant finding, moderate effect)
1: The finding provides marginal evidence of an effect of the practice (significant finding or effect size equivalent to Cohen’s d of .15 or greater even if finding is not statistically significant; small effect*)
0: The finding provides no evidence of an effect of the practice (non-significant finding and effect size of less than .15 – i.e., no effect)

If the effect size magnitude is not reported for a statistically significant finding, an effort should be made to make an informed judgment based on other available information. For example, if the p value for a finding reaches p < .05 but not p < .01, and the sample size is large, the effect size is likely to be small. For example, if there is no evidence of an effect, a finding should be scored as none. These also may be cases where statistical significance is reported but not with respect to whether the effect estimate is p < .05, two-tailed (e.g., p < .10 for one-tailed or p < .05 one-tailed). In these cases, the reviewer should make an informed judgment based on other available information (e.g., if a coefficient and standard error are available, the p-value may be likely to be achieved with a moderate effects). In all instances, it is important to keep in mind that the guidelines provided for scoring of this item are also subject to reviewer judgment and discretion – for example, if a finding is for an outcome that is relatively distant from the conceptually (e.g., flow to flow from and thus be dependent on more proximal or current impacts of the program on other outcomes) or temporally (e.g., long-term follow-up period) and thus likely harder to impact, a finding that approaches but does not reach statistical significance or threshold for a small effect may be judged most appropriate to still score as a small effect (or a small effect on the same type of outcome might be judged appropriate to rate as a moderate effect).

In some instances, the primary intended effects of the practice may be other than youth attitudes or behavior or characteristics of mentoring relationships. For example, a practice could focus on decreasing wait-list times for youth referred to mentoring programs through more streamlined screening, intake, and/or matching processes. Another example would be a practice that seeks to increase the number or diversity of volunteers who apply to a program and ultimately be matched with a youth. In these types of instances, demonstration of an absence of an effect on youth or mentoring relationship outcomes may be sufficient rather than requiring each such outcome to be improved. To the extent that such circumstances apply for a given practice, relevant outcomes will be identified on the scoring form for reviewers and will be scored as follows (and favorability will automatically be assigned a score of 1):

3: The finding provides evidence of the presence of a positive effect on the practice (significant finding in positive direction or effect size in positive direction of .15 or greater)
2: The finding provides moderate evidence of a positive effect of the practice (non-significant finding in a positive direction with p-value less than .05; medium effect*)
1: The finding provides marginal evidence of an effect of the practice (significant finding in positive direction or effect size in a positive direction with p-value less than .05 two-tailed or p < .05 one-tailed). In these cases, the reviewer should make an informed judgment based on other available information (e.g., if a coefficient and standard error are available, the p-value may be likely to be achieved with a moderate effects). In all instances, it is important to keep in mind that the guidelines provided for scoring of this item are also subject to reviewer judgment and discretion – for example, if a finding is for an outcome that is relatively distant from the conceptually (e.g., flow to flow from and thus be dependent on more proximal or current impacts of the program on other outcomes) or temporally (e.g., long-term follow-up period) and thus likely harder to impact, a finding that approaches but does not reach statistical significance or threshold for a small effect may be judged most appropriate to still score as a small effect (or a small effect on the same type of outcome might be judged appropriate to rate as a moderate effect).

If a study does not report one of the types of effect sizes above, an effect size can generally be calculated using information provided in the study. This can be done using an Effect Size Calculator. Please use the link below to calculate effect sizes:

Effect Size Calculator

<table>
<thead>
<tr>
<th>PRIMARY OUTCOMES CHART</th>
<th>NUMBER OF PRIMARY OUTCOMES</th>
<th>FINDINGS</th>
<th>UNWEIGHTED SCORE</th>
<th>WEIGHT VALUE</th>
<th>DIRECTION (1 = FAVORABLE - 1 UNFAVORABLE; IF UNWEIGHTED SCORE = 0, SCORE AS 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Outcome 1</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 2</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 3</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 4</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 5</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 6</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 7</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 8</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 9</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Outcome 10</td>
<td>x 3 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td>- 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SECONDARY OUTCOMES CHART</th>
<th>SECONDARY OUTCOMES</th>
<th>FINDINGS</th>
<th>UNWEIGHTED SCORE</th>
<th>WEIGHT VALUE</th>
<th>DIRECTION (1 = FAVORABLE - 1 UNFAVORABLE; IF UNWEIGHTED SCORE = 0, SCORE AS 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary Outcome 1</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 2</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 3</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 4</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 5</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 6</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Outcome 7</td>
<td>x 1 x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CRIME SOLUTIONS RESOURCE CENTER: PROGRAM RATING INSTRUMENT--PART 2

Development Services Group, Inc., 7315 Wisconsin Avenue, Suite 800E, Bethesda, MD 20814
CRIME SOLUTIONS RESOURCE CENTER: PROGRAM RATING INSTRUMENT--PART 2

B. BEHAVIOR assesses the degree to which a practice produces changes in behavior. Such change can be in the youth’s behavior outside of the mentoring relationship (e.g., reductions in criminal behavior, substance use, etc.) as well as in the interactions that occur between the mentor and youth (including the length of time that the mentor and youth remain engaged in their relationship). Both the consistency and magnitude of effects on behavioral outcomes should be considered in scoring this item. For example, consistent effects of small magnitude across or nearly all behavioral outcomes could constitute robust evidence of behavioral change, but so could effects of larger magnitude on selected behavioral outcomes. (Notes: 1. Behavior change need not be limited to individual behavior, but may also include organizational change or changes in community-level behavior, such as an increase in convictions, a reduction in the fear of crime, or a drop in crime rates. A drop in arrests in a particular group or community may also be considered behavioral change. 2. Behavior change could include effects that exceed the threshold for evidence of a marginal effect, but are not statistically significant. In A. above, 3. For behavioral outcomes subject to alternative scoring guidelines, consider the strength of evidence for the absence of an effect on the outcomes, as was done in A., above.)

Rating Points and Description
2= The findings provide robust evidence of behavioral change (must include evidence on more than one measure).
1= The findings provide limited evidence of behavioral change.
0= The findings provide no evidence of behavioral change.
-1= The findings provide evidence of negative behavioral change.

Rating Description
1= The preponderance of evidence indicates positive behavioral effects.
0= The preponderance of evidence indicates no behavioral effect or behavioral effects were not assessed.
-1= The preponderance of evidence indicates negative behavioral effects.

C. BEHAVIOR DIRECTIONAL INDICATOR indicates the direction of the behavioral effects based on the preponderance of the evidence. (Note: This element is a multiplier.)

Rating Points and Description
2= The preponderance of evidence indicates positive behavioral effects.
1= The preponderance of evidence indicates no behavioral effect or behavioral effects were not assessed.
0= The preponderance of evidence indicates negative behavioral effects.

Rating Description
1= The preponderance of evidence indicates positive behavioral effects.
0= The preponderance of evidence indicates no behavioral effect or behavioral effects were not assessed.
-1= The preponderance of evidence indicates negative behavioral effects.

OUTCOME EVIDENCE SCORING TABLE

<table>
<thead>
<tr>
<th>Behavior Points</th>
<th>DIRECTIONAL INDICATOR</th>
<th>SUB TOTAL</th>
<th>Substantive Practice Effects Points</th>
<th>TOTAL POINTS</th>
<th>NUMBER OF ITEMS</th>
<th>OUTCOME EVIDENCE SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Scoring Directions: Points are summed, divided by the number of items in the dimension, and then multiplied by the directional indicator. A positive value indicates positive effects of the practice while a negative value indicates negative effects. A zero indicates a neutral effect.

PRACTICE FIDELITY

A. DOCUMENTATION refers to the process of recording information about practice fidelity (i.e., the degree to which the practice is implemented as designed). To effectively establish causality, practice designers should operationally define the core components of the practice that are necessary and sufficient to achieve the outcomes desired. In mentoring programs, core components of a practice should be considered including both any components that are to be implemented by mentors, even if they are volunteers, and any components of the practice’s implementation for which program staff are responsible. Consideration should be given to both documentation of implementation of the practice, as broadly defined (e.g., did mentor training sessions happen) as well as documentation of the implementation of the practice’s specific elements and content (e.g., the degree to which the content and activities of a mentor training session were delivered).

Rating Points and Description
2= The implementation of the practice is systematically assessed and the information collected addresses both broad (e.g., frequency of training sessions) and more detailed (e.g., adherence to manualized guidelines) aspects of the practice’s implementation.
1= The implementation evidence of the practice is systematically assessed, but the information collected is limited in scope (e.g., address only broad aspects of practice implementation).
0= Information regarding implementation of the practice is provided, but is non-systematic (e.g. hoc), incomplete, and/or anecdotal.
-1= No information about implementation of the practice.

Rating Description
2= The implementation of the practice is systematically assessed and the information collected addresses both broad (e.g., frequency of training sessions) and more detailed (e.g., adherence to manualized guidelines) aspects of the practice’s implementation.
1= The implementation evidence of the practice is systematically assessed, but the information collected is limited in scope (e.g., address only broad aspects of practice implementation).
0= Information regarding implementation of the practice is provided, but is non-systematic (e.g. hoc), incomplete, and/or anecdotal.
-1= No information about implementation of the practice.

B. ADHERENCE (directional indicator) refers to the degree to which the core components of the practice are implemented as designed. (Note: This element is a multiplier.)

Rating Points and Description
2= The implementation of the practice is systematically assessed and the information collected addresses both broad (e.g., frequency of training sessions) and more detailed (e.g., adherence to manualized guidelines) aspects of the practice’s implementation.
1= The implementation evidence of the practice is systematically assessed, but the information collected is limited in scope (e.g., address only broad aspects of practice implementation).
0= Information regarding implementation of the practice is provided, but is non-systematic (e.g. hoc), incomplete, and/or anecdotal.
-1= No information about implementation of the practice.
The Reviewer Confidence/Override Option score is not included in the final score. If it is determined by both reviewers that no confidence can be placed on the results of the review process, the study will be coded as a Class 5 (Insufficient Findings) and will be eliminated from the review process. If one reviewer invokes the Override Option and the other does not, the dispute resolution process will be used to classify the study.

Examples of these further considerations include:

**Outcomes:** Study outcomes should match the intent of the practice and be valid measures relating to the practice's purpose. The reviewer should take into account if the specified outcomes match the intent of the practice.

**Annotations:** Anomalous findings may contradict the intent of the practice and suggest the possibility of confounding causal variables. The reviewer should judge if anomalous findings draw into question the confidence in the results of the evaluation.

**Statistical Analysis:** The type of statistical analysis utilized can sometimes influence the outcomes. The reviewer should take into account whether the statistical analysis was appropriate given the research design.

**Other:** The reviewer should consider whether the study possesses any other limitations not expressly or inadequately addressed in the instrument that reduces the confidence in the results of the evaluation.

### PRACTICE FIDELITY SCORING TABLE

<table>
<thead>
<tr>
<th>Documentation Points</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>总分 (TOTAL)</td>
<td>0</td>
</tr>
<tr>
<td>数目 (NUMBER OF ITEMS)</td>
<td>1</td>
</tr>
<tr>
<td>子总分 (SUB TOTAL)</td>
<td>0</td>
</tr>
<tr>
<td>遵从性方向指示器 (ADHERENCE DIRECTIONAL INDICATOR)</td>
<td>0</td>
</tr>
<tr>
<td>实践性证据分数 (PRACTICE EVIDENCE SCORE)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### REVIEWER CONFIDENCE/OVERRIDE OPTION

The Override Option is intended to be used sparingly and only if the reviewer lacks confidence in the results of the scoring instrument as it pertains to the study. The Override provides an opportunity to exercise judgment and discretion based on the reviewer’s expertise for items that may not have been explicitly captured in the elements of the instrument. If the reviewer feels that no confidence can be placed on the results, detailed reasons must be provided. If this option is invoked by both reviewers, the study will be coded as a Class 5 (Insufficient Findings) and will be eliminated from the review process. If one reviewer invokes the Override Option and the other does not, the dispute resolution process will be used to classify the study.

### OVERALL SCORE

<table>
<thead>
<tr>
<th>维度</th>
<th>Class 1 (Effective)</th>
<th>Class 2 (Promising)</th>
<th>Class 3 (Ineffective)</th>
<th>Class 4 (Null Effect)</th>
<th>Class 5 (Insufficient Information)</th>
</tr>
</thead>
<tbody>
<tr>
<td>本科学领域 (Conceptual Framework)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>设计质量 (Design Quality)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>结果证据 (Outcome Evidence)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>实践性 (Practice Fidelity)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### CLASSIFICATION SYSTEM

The score in each of the four dimensions is calculated separately and used to assess each study. The maximum overall score in each dimension is 3 points. The outcome evidence and practice fidelity dimensions include directional indicators to signify the directional nature of the dimension. These dimensions are then used to classify each study into one of the following five classes:

- **No**
- **Class 1 (Effective)**
- **Class 2 (Promising)**
- **Class 3 (Ineffective)**
- **Class 4 (Null Effect)**
- **Class 5 (Insufficient Information)**

### Integration of Evidence

An aggregation of this research base is used to rate the effectiveness of each practice, as follows:

- A practice will be listed as "Effective" if the following is true:
  - It has at least 2 studies in Class 1 or 1 study in Class 1 and at least 2 studies in Class 2
  - It has no studies in Class 3
  - It has 0 or less than 33% of studies in Class 4

- A practice will be listed as "Promising" if the following is true:
  - It does not qualify as Effective
  - It has at least 1 study in Class 1 or at least 2 studies in Class 2
  - It has 0 studies in Class 3
  - It has 0 or less than 33% of studies in Class 4

- A practice will be listed as "No Effect" if the following is true:
  - It does not qualify as Effective or Promising
  - It has at least 2 studies in either Class 3 or Class 4

- A practice will be listed as "Null Effect" if the following is true:
  - It has a score less than 0 in the outcome evidence dimension and scores of at least 2.0 for the design and fidelity dimensions of practice effectiveness; in general, when implemented with sufficient fidelity and using an evaluation design of high quality (quasi-experimental), this study demonstrates negative practice effects.

- A practice will be listed as "Insufficient Information" if the following is true:
  - The study must have a score less than 1.0 for the design dimension and a score of at least 2.0 for the outcome evidence dimension; in general, this study demonstrates no evidence in favor of the practice when evaluated with a design of high quality (quasi-experimental) and implemented with sufficient fidelity. More extensive research is required.
A practice will be listed as "Insufficient Evidence" if the following is true:
It does not qualify as Effective, Promising, or No Effects.
A. EXTENT OF TESTING: PRACTICE VARIATIONS

assesses the degree to which available studies provide tests of the practice's effects across different variations of the practice. Dimensions that are assessed can include whether the practice is applied to an individual of a certain age, sex, or race, whether the practice is delivered through community-based or school-based programs, and whether the practice is delivered by paid or volunteer mentors. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Rating Points and Description

0 = Findings regarding effects of the practice show very little or no consistency across relevant variations of the practice.

1 = Findings regarding effects of the practice are moderately consistent across relevant variations of the practice.

2 = Findings regarding effects of the practice are highly consistent across relevant variations of the practice.

0 = The available studies test effects of the practice across no relevant variations of the practice.

1 = The available studies test effects of the practice across some of the relevant variations of the practice.

2 = The available studies test effects of the practice across all relevant variations of the practice.

B. CONSISTENCY OF EFFECTS: PRACTICE VARIATIONS (Answer only if Question A above is NOT rated 0)

assesses the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant variations of the practice as determined under A. There may be instances in which differences in effects are limited to variations in the magnitude of what would still consistently indicate substantial (medium effect size or larger) effects of the practice. Such variation should not result in a lower score on this item.

Rating Points and Description

0 = Findings regarding effects of the practice are highly consistent across relevant variations of the practice.

1 = Findings regarding effects of the practice are moderately consistent across relevant variations of the practice.

2 = Findings regarding effects of the practice are only limited consistency across relevant variations of the practice.

0 = The available studies test effects of the practice across no relevant variations of the practice.

1 = The available studies test effects of the practice across some of the relevant variations of the practice.

2 = The available studies test effects of the practice across all relevant variations of the practice.

0 = The available studies test effects of the practice across the preponderance of the relevant variations of the position.

1 = The available studies test effects of the practice across some of the relevant variations of the position.

2 = The available studies test effects of the practice across all relevant variations of the position.

C. EXTENT OF TESTING: YOUTH

assesses the degree to which available studies provide tests of the practice's effects across different subgroups of youth. Subgroups relevant to consider will vary by practice, but may include race, gender, socioeconomic status, family background, and indication of individualization risk or vulnerability (e.g., problem behavior involvement, disability). Particular emphasis should be given to subgroups of youth that fall within the target population for the practice -- see Practice Description in Part I of the instrument. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Rating Points and Description

0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of youth.

1 = Findings regarding effects of the practice are moderately consistent across relevant subgroups of youth.

2 = Findings regarding effects of the practice are highly consistent across relevant subgroups of youth.

0 = The available studies test effects of the practice with no variation in relevant subgroups of youth.

1 = The available studies test effects of the practice across only limited relevant subgroups of youth.

2 = The available studies test effects of the practice across some of the relevant subgroups of youth.

3 = The available studies test effects of the practice across the preponderance of the relevant subgroups of youth.

D. CONSISTENCY OF EFFECTS: YOUTH (Answer only if Question C above is NOT rated 0)

assesses the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant subgroups of youth. Subgroups relevant to consider will vary by practice, but may include race, gender, socioeconomic status, family background, and indication of individualization risk or vulnerability (e.g., problem behavior involvement, disability). Particular emphasis should be given to subgroups of youth that fall within the target population for the practice -- see Practice Description in Part I of the instrument. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Rating Points and Description

0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of youth.

1 = Findings regarding effects of the practice show only limited consistency across relevant subgroups of youth.

2 = Findings regarding effects of the practice are moderately consistent across relevant subgroups of youth.

3 = Findings regarding effects of the practice are highly consistent across relevant subgroups of youth.

0 = The available studies test effects of the practice with no variation in relevant subgroups of youth.

1 = The available studies test effects of the practice across only limited relevant subgroups of youth.

2 = The available studies test effects of the practice across some of the relevant subgroups of youth.

3 = The available studies test effects of the practice across the preponderance of the relevant subgroups of youth.

E. EXTENT OF TESTING: MENTORS

assesses the degree to which available studies provide tests of the practice's effects across different subgroups of mentors. Determinants that are relevant to consider will vary by practice, but may include race, gender, socioeconomic status, family background, and indication of individualization risk or vulnerability (e.g., problem behavior involvement, disability). Particular emphasis should be given to subgroups of mentors to whom the practice is likely to be applied and/or those likely to be serving in programs that utilize the practice -- see Practice Description in Part I of the instrument. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Rating Points and Description

0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of mentors.

1 = Findings regarding effects of the practice show only limited consistency across relevant subgroups of mentors.

2 = Findings regarding effects of the practice are moderately consistent across relevant subgroups of mentors.

3 = Findings regarding effects of the practice are highly consistent across relevant subgroups of mentors.

0 = The available studies test effects of the practice with no variation in relevant subgroups of mentors.

1 = The available studies test effects of the practice across only limited relevant subgroups of mentors.

2 = The available studies test effects of the practice across some of the relevant subgroups of mentors.

3 = The available studies test effects of the practice across the preponderance of the relevant subgroups of mentors.

F. CONSISTENCY OF EFFECTS: MENTORS (Answer only if Question E above is NOT rated 0)

assesses the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant subgroups of mentors. Subgroups relevant to consider will vary by practice, but may include race, gender, socioeconomic status, family background, and indication of individualization risk or vulnerability (e.g., problem behavior involvement, disability). Particular emphasis should be given to subgroups of mentors to whom the practice is likely to be applied and/or those likely to be serving in programs that utilize the practice -- see Practice Description in Part I of the instrument. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Rating Points and Description

0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of mentors.

1 = Findings regarding effects of the practice show only limited consistency across relevant subgroups of mentors.

2 = Findings regarding effects of the practice are moderately consistent across relevant subgroups of mentors.

3 = Findings regarding effects of the practice are highly consistent across relevant subgroups of mentors.

0 = The available studies test effects of the practice with no variation in relevant subgroups of mentors.

1 = The available studies test effects of the practice across only limited relevant subgroups of mentors.

2 = The available studies test effects of the practice across some of the relevant subgroups of mentors.

3 = The available studies test effects of the practice across the preponderance of the relevant subgroups of mentors.
H. EXTENT OF TESTING: PROGRAM SETTINGS/STRUCTURES

This section assesses the degree to which available studies provide for tests of the practice's effects across different program settings and structures. Differences in the types of program settings and structures within which the practice is likely to be applied can result in differences in outcomes (e.g., youth engagement, academic achievement, delinquency prevention). Particular emphasis should be given to variations that the types of program settings and structures within which the practice is likely to be applied — see Practice Description in Part I of the instrument. The rating can be informed by information that is included in study reports and/or their supplementary materials.

Consistency of Effect

<table>
<thead>
<tr>
<th>Extent of Test</th>
<th>Rating Points and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some</td>
<td>Findings regarding effects of the practice show only limited consistency across relevant variations in program setting/structure.</td>
</tr>
<tr>
<td>Extensive</td>
<td>Findings regarding effects of the practice show highly consistent across relevant variations in program setting/structure.</td>
</tr>
<tr>
<td>None</td>
<td>Findings regarding effects of the practice do not show consistent across relevant variations in program setting/structure.</td>
</tr>
</tbody>
</table>

Classification System

The visual below provides a representation of the extent to which, for each of the above dimensions, external validity has been tested for the practice (Items A, C, G, I, and J above) and the extent to which available findings show evidence of consistency across different program settings and structures (Items B, D, E, F, H, and J above).