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 insrument.

Rating	Points and Description
	3= High (5 or more other studies, or 1 meta-analysis, provide evidence in support of the practice).
	2= Medium (2 to 4 other studies provide evidence in support of the practice).
	1= Low (1 other study provides evidence in support of the practice).
	0= None (No other studies provide evidence in support of the practice).

Notes:

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).

Rating	Points and Description
	3= Practice theory is fully described and conceptually sound.
	2= Practice theory is adequately described and appears conceptually sound.
	1= Very little information is provided about practice theory, but it may be conceptually sound.

	0= No information about practice theory or practice theory is invalid.

Notes:

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.

Rating	Points and Description
	3= All defining details of the practice details are evident.
	2= Most defining details of the practice are evident.
	1= Some defining details of the practice are evident.
	0= No defining details of the practice are specified.

Notes: Please specify the targeted population, the targeted behaviors, and the key elements of the practice:

CONCEPTUAL FRAMEWORK SCORING TABLE		
	Prior Research Points	0
+	Theoretical Base Points	0
+	Practice Description Points	0
=	TOTAL	0
/	NUMBER OF ITEMS	3
=	CONCEPTUAL FRAMEWORK SCORE	0.00

Instructions Please carefully assess the practice in terms of design quality, outcome evidence, and practice fidelity.**Part 2 should be completed for each study in the research base. Please record your answers for each article on this form.**(Note: The research base for each practice can include up to three studies.). In scoring items in this section, it should be kept in mind that all items are to be scored with respect to only those aspects of the study that pertain to evaluating the relevant practice (e.g., in a randomized controlled design evaluation of a program, in which quasi-experimental comparisons are made among those in the treatment group with respect to receipt of naturally occurring variaition in receipt of the practice, the appropriate Research Design rating would be Quasi-Experimental rather than Experimental).

PRACTICE NAME:			
STUDY #:		**CITATION**	

REVIEWER'S NAME _____ **DATE OF REVIEW** _____

DESIGN QUALITY

A. RESEARCH DESIGN rates 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 (controls) does not. A quasi-experiment 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 comparison group. Since these designs differ in their 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 des (Note: Not all designs easily fit into this hierarchy. The reviewer should specify the design and note the reason for the score.). 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 general, such comparisons would be considered quasi-experimental, 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.

Rating	Points and Description
	3= Experimental (well-designed randomized field trial). Participants are randomly assigned into a practice group and a non practice group.
	2= Quasi-experimental Level 1 (design uses a credible comparison group of participants who receive the practice and participants who do not).
	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).
	0= Non-experiment level 1 (one group pretest–posttest, one- and two-group posttest only, or case studies).

Specify Design: Experimental

Notes:

1 In some cases, random assignment takes place at a different level than the analysis. For example, schools are randomly assigned to conditions, but the students are the unit of analysis. These cases should not be treated as random assignments.

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 rarely straightforward. Statistical power is a function of several factors: 1) the size of the sample; 2) the magnitude of tl 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 26 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 a program or site level (Note: The same rules of thumb do not apply for 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 estimating the structure of the correlated error. Conversely, although it may not account for the randomness of the data, roughly 15 observations are gener considered the minimum.) The reviewer should use his or her expertise to assess the adequacy of the sample.

Rating	Points and Description
	3= High Power: The sample is sufficient to detect a small effect (.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/cluster 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 disproportional).
	2= Medium Power: The sample is sufficient to detect a medium effect (.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/cluster 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 disproportional).
	1= Low Power: The sample is sufficient to detect a large effect (.80) using appropriate tests. (In general, the N should be between 26 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/cluster sizes are small and/or the number of studies or sites with and without the practice under review are markedly disproportional).
	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 10.)

Specify treatment group sample size:

Specify comparison group sample size:

Specify number of observations (Time Series design):

Notes:

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.

Rating	Points and Description
	3= No statistical adjustments required in the analysis. Random assignment or selection modeling (propensity score matching) with a sufficiently large sample resulted in no group differences.
	2= The analysis employs appropriate statistical adjustments (includes control variables that are presumed to be related to the outcome) to control for group differences.
	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.
	0= The analysis does not employ necessary statistical adjustments to control for group differences.
	NA= Not applicable.

Notes:

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 assessments 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 measure, as well as the reading level, native language, and attention span of respondents.

Rating	Points and Description
	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.
	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.

	1= Below Average. The reliability (the extent to which an item produces the same results when used repeatedly) and/or validity (the extent to which an item measures what it is intended to measure) of the measures are below average.
	0= None. No information is provided on the reliability (the extent to which an item produces the same results when used repeatedly) and/or validity (the extent to which an item measures what it is intended to measure) of the measures.

Notes:

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.

Rating	Points and Description
	3= No threats to internal validity are identified or all threats have been adequately addressed.
	2= Marginal threats to internal validity are identified and remain.
	1= Moderate threats to internal validity are identified and remain.
	0= Serious threats to internal validity are identified and remain.

Notes:

Check all that apply	Threat	Description
□	Attrition or Mortality	This threat occurs when participants drop out of the study between the pretest and the posttest. Attrition is important because it affects whether the groups are equivalent except for practice effects at the time of the post-practice outcon measure. The study should have low overall attrition of study participants and minimal differential attrition between the practice and control groups. While there are exceptions, the general guideline states that a study should obtain outcome data for at least 80 percent of the original study subjects. Furthermore, the attrition rate should be approximately the same for the practice and control groups. Severe differential attrition makes the results suspect, because it may compromise the comparability of the groups.
		Notes:
□	Maturation	This threat is caused by the natural maturation process, where respondents grow experienced or bored.
		Notes:
□	Instrumentation	This threat occurs when there is a change in the measuring instrument.
		Notes:
□	Regression Toward the Mean	This threat occurs whenever there is measurement error and participants are selected based on the extremeness of their measured values. The measured values will tend to be closer to the overall mean on a second administration of t instrument.
		Notes:
□	Selection	This threat occurs when the groups to be compared differ on factors besides the practice. Even if the subjects are randomly assigned, this threat is of particular importance with small sample studies.
		Notes:
□	Contamination	This threat refers to situations where the separation between the groups is less than it should be.
		Notes:
□	History	This threat occurs when an observed effect might be due to an event that takes place between the pretest and the posttest that has nothing to do with the practice.
		Notes:
□	Other	Other threats may include: multiple practice interference, obtrusive testing, secular trends, intervening events, etc.
		Notes:

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.

Rating	Points and Description
	3= More than 1 year.
	2= More than 6 months but less than or equal to 1 year.
	1= Less than or equal to 6 months.
	0= Not specified.

Specify follow-up period in months:

Notes:

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
/	NUMBER OF ITEMS	6
=	Design Quality Score	0.00

SCORING DIRECTIONS. Points are summed and divided by the number of items in the dimension. (Note: Due to the diversity in research designs, some items are not appropriate for all designs. Consequently, the number of items varies by design.)

OUTCOME EVIDENCE

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 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. If no such information is available, a small effect size should be assumed. There 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 two-tailed or p < .05 one-tailed). In these instances, the review should make an informed judgment based on available information (e.g., if a coefficient and standard error are available, the p value may be able to be reasonably inferred). 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 distal either conceptually (e.g., likely to flow from and thus be dependent on more proximal or immediate 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 stream-lined 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 can 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 such outcomes 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 of the practice (significant finding in positive direction or an effect size in positive direction of .15 or greater)
2 = The finding provides moderate evidence of the absence of a negative effect of the practice (non-significant finding in a negative direction with p-value of .25 or greater or, if less than .25, the effect size is no greater than .10 in a negative direction)
1 = The finding provides marginal evidence of the absence of a negative effect of the practice (non-significant finding in a negative direction with a p-value greater than .10 or, if less than .10, the effect size is in a positive direction)
0 = The finding provides no evidence of either an absence of a negative effect of the practice or the presence of a favorable effect (significant or marginally significant finding in negative direction or effect size in a negative direction with p-value less than .20)

Suggested Guidelines for Gauging Effect Size

For Standardized Mean Difference Effect Sizes (Cohen's *d*, also Hedge's *g*, Glass' *g*):

small	=	0.15
medium	=	0.45
large	=	0.90

For Odds Ratios (OR):

small	=	1.31
medium	=	2.26
large	=	5.12

For Correlation Coefficients (*r*):

small	=	0.07
medium	=	0.22
large	=	0.41

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

PRIMARY OUTCOMES CHART

Number of Primary Outcomes	PRIMARY OUTCOMES*	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	DIRECTION (1 = FAVORABLE -1 UNFAVORABLE; IF UNWEIGHTED SCORE = 0, SCORE AS 1)
Primary Outcome 1				x 3	x
Primary Outcome 2				x 3	x
Primary Outcome 3				x 3	x
Primary Outcome 4				x 3	x
Primary Outcome 5				x 3	x
Primary Outcome 6				x 3	x
Primary Outcome 7				x 3	x
Primary Outcome 8				x 3	x
Primary Outcome 9				x 3	x
Primary Outcome 10				x 3	x
	SUM		*	0	

SECONDARY OUTCOMES CHART

	SECONDARY OUTCOMES*	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	DIRECTION (1 = FAVORABLE -1 UNFAVORABLE; IF UNWEIGHTED SCORE = 0, SCORE AS 1)
Secondary Outcome 1				x 1	x
Secondary Outcome 2				x 1	x
Secondary Outcome 3				x 1	x
Secondary Outcome 4				x 1	x
Secondary Outcome 5				x 1	x
Secondary Outcome 6				x 1	x
Secondary Outcome 7				x 1	x

Secondary Outcome 8				x 1	x
Secondary Outcome 9				x 1	x
Secondary Outcome 10				x 1	x
	SUM			0	

*Denotes an outcome for which the alternative scoring guidelines, described above, should be used.

CALCULATION WORKSHEET

	SUM OF WEIGHTED DIRECTIONALSCORE		SUM OF WEIGHT VALUES		
Primary Outcomes	0		1		
Secondary Outcomes	0		0		SUBSTANTIVE PRACTICE EFFECTS SCORE
Total	0	+	1	=	0.00

*If there are no secondary outcomes, the score is the average of the primary outcomes' unweighted score.

B. BEHAVIOR assesses the degree to which a practice produces change(s) 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 all 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 meet 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 eff 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.

Notes:

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	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.

Notes:

OUTCOME EVIDENCE SCORING TABLE

	Behavior Points	0
x	DIRECTIONAL INDICATOR	0
=	SUB TOTAL	0
+	Substantive Practice Effects Points	0.0
=	TOTAL POINTS	0.0
/	NUMBER OF ITEMS	2
=	OUTCOME EVIDENCE SCORE	0.00

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 as 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
	3= 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.
	2= The implementation evidence of the practice is systematically assessed, but the information collected is limited in scope (e.g. addresses only broad aspects of practice's implementation).
	1= Information regarding implementation of the practice is provided, but is non-systematic (ad hoc), incomplete, and/or anecdotal.
	0= No information about implementation of the practice.

Notes:

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

	1= Adherence to practice appears most likely to be satisfactory (i.e., at or above a threshold required to capture most of the practice's effect).
	0= No information about practice implementation. (Note: This option may be selected only if Documentation is scored as 0)
	-1= Adherence to practice appears most likely to be unsatisfactory (i.e., below the threshold required to capture most of the practice's effect).

Notes:

PRACTICE FIDELITY SCORING TABLE		
	Documentation Points	0
=	TOTAL	0
/	NUMBER OF ITEMS	1
=	SUB TOTAL	0
x	ADHERENCE: DIRECTIONAL INDICATOR	0
=	FIDELITY EVIDENCE SCORE	0.00

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 sufficient practice fidelity while a negative value indicates poor practice fidelity. A zero indicates that no information was provided regarding fidelity.

REVIEWER CONFIDENCE/OVERRIDE OPTION

The Override Option is intended to be used sparingly and only if the reviewer lacks confidence in the results of this 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 in 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.

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.

Anomalous Findings: 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.

Rating	Points and Description
	1= Confidence should be placed on the results of this evaluation because the number and type of limitations are minimal.
	0= Very limited or no confidence should be placed in the results of this evaluation because the number and type of limitations are too serious.*

*Note: If "0" is selected, the reviewer must explain below why you do not have confidence in the results and why this was not captured in the scoring instrument.
Notes:

OVERALL SCORE

		Conceptual Framework	Design Quality	Outcome Evidence	Practice Fidelity
Overall Score*		0.00	0.00	0.00	0.00

*The Reviewer Confidence/Override Option score is not included in the final score. If it is determined by both reviewers that no confidence should be placed on the results, 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.

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:

Check	Class	DESCRIPTION
No	**Class 1 (Effective)**	**This study must have exceptional scores (at least 2.0) for the Conceptual Framework, Design Quality, and Practice Fidelity dimensions and a score of at least 1.75 for the Outcome Evidence dimension.** In general, this study demonstrates strong evidence in favor of the practice when evaluated with a design of high quality (quasi-experimental) and implemented with sufficient fidelity.
No	**Class 2 (Promising)**	**This study must have a score of at least 1.5 for the design dimension and at least 1.0 for the outcome evidence dimension.** In general, this study demonstrates promising (perhaps inconsistent) evidence in favor of the practice when evaluated with a design of high quality (quasi-experimental). More extensive research is required.
No	**Class 3 (Ineffective)**	**This study must have 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.
No	**Class 4 (Null Effect)**	**This study must have a score for the outcome evidence dimension that is positive but less than 1.0 and scores of at least 2.0 for the design and fidelity dimensions of practice effectiveness** 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.
Yes	**Class 5 (Insufficient Information)**	**This study must have a score less than 1.5 for the design dimension or a score of less than 1 for outcome evidence dimension in combination with a score of less than 2.0 for the fidelity dimension or the design dimension.** In general, there is insufficient evidence to rate this study.

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% studies in Class 4

• A practice will be listed as "**No Effects**" 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

Development Services Group, Inc., 7315 Wisconsin Avenue, Suite 800E, Bethesda, MD 20814

• A practice will be listed as **"Insufficient Evidence"** if the following is true:
It does not qualify as Effective, Promising, or No Effects

Instructions: Please carefully assess the evidence regarding the practice in terms of external validity. The reviewer should complete this part of the scoring instrument only once for each practice, regardless of the number of studies that are being reviewed. It should be noted that both within-study and across-study evidence should be considered. For example, a particular study may test the effects of a practice separately for community- and school-based mentoring programs, but there may also be some studies that investigate the practice solely within community-based programs and others that do so solely within school-based programs. Each aspect of the evidence base would provide information relevant to the Program Setting/Structure dimension of external validity, although in general within-study comparisons are likely to be more informative relative to cross-study comparisons given that cross-study comparisons of findings for a practice are subject to the influence of other potential differences between the studies involved. Complete this section by using the pertinent information from the studies provided related to the practice The following assessment will not affect the overall classification of the practice, but will be featured in the profile of the practice on the NMRC website.

PRACTICE NAME:	

REVIEWER'S NAME: **DATE OF REVIEW:**

EXTERNAL VALIDITY

A. EXTENT OF TESTING: PRACTICE VARIATIONS assesses the degree to which available studies provide for tests of the practice's effects across different variations of the practice. Dimensions that are relevant to consider will vary by practice but may include key activities and content of the practice as well as the frequency and duration of the activities associated with the practice. Particular emphasis should be given to variations that fall within the scope of the practice as generally defined -- 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.

	Points and Description
	3= The available studies test effects of the practice across the preponderance of the relevant variations of the practice.
	2= The available studies test effects of the practice across some of the relevant variations of the practice.
	1= The available studies test effects of the practice across only limited relevant variations of the practice.
	0= The available studies test effects of the practice with no variation in relevant variations of the practice.

Notes:

B. CONSISTENCY OF EFFECTS: PRACTICE VARIATIONS (Answer only if Question A above is NOT rated 0) refers to the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant variations in 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
	3 = Findings regarding effects of the practice are highly consistent across relevant variations of the practice.
	2= Findings regarding effects of the practice are moderately consistent across relevant variations of the practice.
	1 = Findings regarding effects of the practice show only limited consistency across relevant variations of the practice.
	0 = Findings regarding effects of the practice show very little or no consistency across relevant variations of the practice.

Notes:

C. EXTENT OF TESTING: YOUTH assesses the degree to which available studies provide for tests of the practice's effects across different subgroups of youth. Subgroups relevant to consider will vary by practice, but may include those associated with differences in youth demographics (e.g., age, gender, race/ethnicity), levels and types of adversity exposure (e.g., low socioeconomic status, incarcerated parent), and indications of individual-level 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.

	Points and Description
	3= The available studies test effects of the practice across the preponderance of the relevant subgroups of youth.
	2= The available studies test effects of the practice across some of the relevant subgroups of youth.
	1= The available studies test effects of the practice across only limited relevant subgroups of youth.
	0= The available studies test effects of the practice with no variation in relevant subgroups of youth.

Notes:

D. CONSISTENCY OF EFFECTS: YOUTH (Answer only if Question C above is NOT rated 0) refers to the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant subgroups of youth as determined under C. 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
	3 = Findings regarding effects of the practice are highly consistent across relevant subgroups of youth.
	2= Findings regarding effects of the practice are moderately consistent across relevant subgroups of youth.
	1 = Findings regarding effects of the practice show only limited consistency across relevant subgroups of youth.
	0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of youth.

Notes:

E. EXTENT OF TESTING: MENTORS assesses the degree to which available studies provide for tests of the practice's effects across different subgroups of mentors. Dimensions that are relevant to consider will vary by practice but may include demographic characteristics (e.g., age, gender, race/ethnicity), prior background (e.g., professional training or experience), and paid vs.volunteer status. Particular emphasis should be given to the types 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
	3= The available studies test effects of the practice across the preponderance of the relevant subgroups of mentors.
	2= The available studies test effects of the practice across some of the relevant subgroups of mentors.
	1= The available studies test effects of the practice across only limited relevant subgroups of mentors.
	0= The available studies test effects of the practice with no variation in relevant subgroups of mentors.

Notes:

F. CONSISTENCY OF EFFECTS: MENTORS (Answer only if Question E above is NOT rated 0) refers to the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant subgroups of mentors as determined under E. 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
	3 = Findings regarding effects of the practice are highly consistent across relevant subgroups of mentors.
	2= Findings regarding effects of the practice are moderately consistent across relevant subgroups of mentors.
	1 = Findings regarding effects of the practice show only limited consistency across relevant subgroups of mentors.
	0 = Findings regarding effects of the practice show very little or no consistency across relevant subgroups of mentors.

Notes:

G. EXTENT OF TESTING: PROGRAM SETTINGS/STRUCTURES assesses the degree to which available studies provide for tests of the practice's effects across different program settings and structures. Dimensions that are relevant to consider will vary by practice but may include location of mentoring activities (e.g., school, CBO, community-at-large, on-line), mentoring format (1-to-1, group, team), program goals (e.g., academic achievement, delinquency prevention), and nature and extent of program practices other than the practice under review (e.g., ongoing training or match support in the context of a review of pre-match training). 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.

Rating	Points and Description
	3= The available studies test effects of the practice across the preponderance of the relevant variations in program setting/structure.
	2= The available studies test effects of the practice across some of the relevant variations in program setting/structure.
	1= The available studies test effects of the practice across only limited relevant variations in program setting/structure.
	0= The available studies test effects of the practice with no variation in relevant variations in program setting/structure.

Notes:

H. CONSISTENCY OF EFFECTS: PROGRAM SETTINGS/STRUCTURES (Answer only if Question G above is NOT rated 0) refers to the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant variations in program setting and structure as determined under G. 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
	3 = Findings regarding effects of the practice are highly consistent across relevant variations in program setting/structure.
	2= Findings regarding effects of the practice are moderately consistent across relevant variations in program setting/structure.
	1 = Findings regarding effects of the practice show only limited consistency across relevant variations in program setting/structure.
	0 = Findings regarding effects of the practice show very little or no consistency across relevant variations in program setting/structure.

Notes:

I. EXTENT OF TESTING: OUTCOMES assesses the degree to which available studies provide for tests of the practice's effects across different types of outcomes. The types of outcomes that are relevant to consider will vary by practice but may include those relating to mentoring relationship quality and duration, youth emotional, behavioral, and academic functioning, and program operations or functioning (e.g., volunteer retention, staff turnover). Particular emphasis should be given to outcomes that fall within the scope of potential targeted outcomes 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
	3= The available studies test effects of the practice across the preponderance of the relevant variations in outcomes.
	2= The available studies test effects of the practice across some of the relevant variations in outcomes.
	1= The available studies test effects of the practice across only limited relevant variations in outcomes.
	0= The available studies test effects of the practice with no variation in relevant variations in outcomes.

Notes:

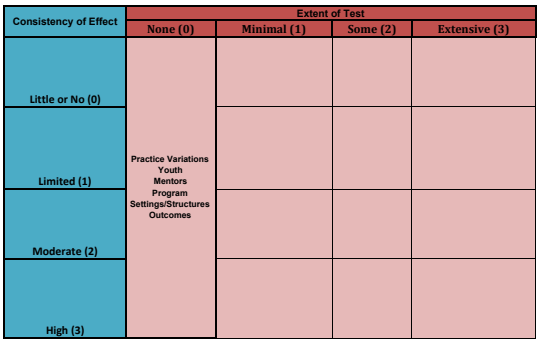
J. CONSISTENCY OF EFFECTS: OUTCOMES (Answer only if Question I above is NOT rated 0) refers to the degree to which, where evidence is available, estimated effects of the practice are consistent across relevant variations in outcomes as determined under I. There may be instances in which differences in 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.

	Points and Description
	3 = Findings regarding effects of the practice are highly consistent across relevant variations in outcomes.
	2= Findings regarding effects of the practice are moderately consistent across relevant variations in outcomes.
	1 = Findings regarding effects of the practice show only limited consistency across relevant variations in outcomes.
	0 = Findings regarding effects of the practice show very little or no consistency across relevant variations in outcomes.

Notes:

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, E,G, and I above) and the extent to which available findings show evidence of consistency across that dimension (Items B, D, F, H, and J above).



Admin Only	
Source	Source
1	3
0	2
-1	1
	0