CrimeSolutions.gov Program Scoring Instrument Version 2.0

Instructions: Please carefully assess the program in terms of the conceptual framework. **The reviewer should complete Part 1 only once for each program**, regardless of the number of studies to be reviewed. Complete this section by using the pertinent information from the studies and any other program materials you have received. Please record your answers on this form.

PROGRAM 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) supports the conceptual framework of comparable programs. Significantly, the scope of comparable programs will vary by program. For instance, Multisystemic Therapy (MST) has undergone numerous evaluations, thus the scope of comparable programs can be narrowed to consist solely of MST rather than include the other family-based treatment models. In contrast, a program such as the Chicago Alternative Policing Strategy is unlikely to have been the subject of repeated evaluation. In this case, the scope of comparable programs can be widened to include other similar community-based policing programs. Similarly, an intervention such as a Gunshot Detection System (when it was first implemented) had not been not the subject of repeated evaluation and was a novel approach to policing. In this case, the scope of comparable programs cannot be widened, because there are no other similar programs.
Note 1. A meta-analysis will typically include five or more studies on a single practice. Consequently, if a meta-analysis provides evidence to support the program, the research base should receive the highest score.
Note 2. Even though an independent evaluator is NOT required for consideration, specify the association between the program and evaluator (if known).
Note 3. Multiple publications based on the different samples should carry more weight than multiple publications based on the same sample, which for CrimeSolutions.gov and the Model Programs Guide are considered a single study.
Note 4. Please consider, in addition to the evidence presented in the study or supplemental materials, your own knowledge of the empirical evidence on comparable programs.

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

Program-Evaluator Association:

Notes:

B. *Theoretical Base* measures the degree to which the program is based on a well-articulated, conceptually sound program theory–it should explain why the program should effect change. Acceptable program theory may be articulated or implicit. The program should provide an explanation of why and how it is expected to achieve its intended results.

- A well-articulated program theory is clearly defined and sound; previous empirical work related to the theory is described; and there is an explanation about how the theory relates to the specific program components and how this should result in change for the participants. It would be appropriate to give a program with this level of detail a score of 3 for being fully described.
- A program that defines and describes an empirically supported theory, but does not necessarily connect the theory with the specific components of the program or provide a theory of change, would be considered adequately described and it would be appropriate to score the program no higher than 2.
- A program that provides very little information on program theory–perhaps referring to a theory but not describing it, referencing prior empirical support, describing the theory of changes, or tying it into program components–would be given a score no higher than 1.
- A program that does not mention a theory; or a program based on a theory known to be unsound would be scored 0.
- An implicit theory that appeals to common sense could receive any of the range of scores described depending on the degree to which its theory of change is fully described, empirically supported, and connections are made to program components.

Rating	Points and Description
	3= Program theory is fully described and conceptually sound.
[redacted]	2= Program theory is adequately described and appears conceptually sound.
	1= Very little information is provided about program theory, but it may be conceptually sound.
	0= No information about program theory or program theory is invalid.

Notes:

C. *Program Description* rates the degree to which the program details are described. A full and thorough description should serve as a guide for the implementation of the program. It would include the following information: 1) the logic of the program, 2) the details of all key components, 3) the frequency and duration of the program activities, 4) the targeted population, 5) the targeted behavior(s) (i.e., the intent of the program), and 6) the setting. The rating should reflect the degree to which the provided materials afford an adequate program description and/or direct the reader to references containing such a description.

Rating	Points and Description
	3= All program details are specified (5-6 items are described).
	2= Most program details are specified (3-4 items are described).
	1= Some program details are specified (1-2 items are described).
	0= No program details are specified.

Please specify/indicate inclusion of the following:

1. Logic of the program:	
2. Details of all key components:	
3. Frequency/duration of program activities:	
4. Targeted population:	
5. Targeted behaviors:	
6. Setting:	

Additional Notes:

CONCEPTUAL FRAMEWORK SCORING TABLE		
	Prior Research Points	**FALSE**
+	**Theoretical Base Points**	**FALSE**
+	**Program Description Points**	**FALSE**
=	**TOTAL**	**0**
/	**NUMBER OF ITEMS**	**False**
=	**CONCEPTUAL FRAMEWORK SCORE**	**#VALUE!**

Instructions: Please carefully assess the program in terms of design quality, outcome evidence, and program 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 program can include up to three studies.)

PROGRAM NAME:			0
STUDY #:	1	**CITATION**	

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

DESIGN QUALITY

A. *Research Design* rates the ability of the design to infer a causal relationship between program treatment and outcome. 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 program (treatment), while the other (controls) does not. A quasi-experiment research design is similar 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 one or both of the above characteristics. Since these designs differ in their assignment strategy, it is likely they will differ in terms of their strength with respect to internal validity.
Note 1: Not all designs easily fit into this hierarchy. The reviewer should specify the design and note the reason for the score.
Note 2: 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 analysis. These cases should not be treated as random assignments.

Rating	Points and Description
	3= Experimental (well-designed randomized field trial).
	2= Quasi-experimental Level 1 (design uses a credible comparison group with extensive information provided on pre-treatment equivalence of groups; time series comparison group design).
	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-experimental (one group pretest-posttest, one- and two-group posttest only, or case studies).

Specify Design:

Notes:

B. *Sample Size (Power).* The purpose of this item is to estimate the precision of the program effects and assess the adequacy of the sample to detect meaningful program effects. Though several factors can affect the precision of a parameter estimate, sample size is always a factor, because the greater the sample size the closer the sample is to the actual population, and if you take a sample that consists of the entire population, there is no sampling error.
Note 1. The sample size should be based on the unit of analysis: For an individual-level assignment, base your rating on the number of individuals in the analytic sample. For a cluster-level assignment, base your rating on the number of clusters in the analytic sample, regardless of whether 1) the analysis is at the individual level or the cluster level and whether 2) the analytic method adjusts for clustering (e.g., like a multilevel model).
Note 2: The same rules do not apply for time-series designs, as precision is determined by the number of observations. Most textbooks suggest that about 50 observations, with a reasonable distribution among pretest and posttest measurements, is required for a competent analysis, on the 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 generally 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: Analytic sample size greater than 394 per group for person-level assignment, greater than 90 clusters per group for cluster-level assignment, or greater than 75 in a time-series design (80 percent or greater chance of detecting an effect size of 0.20).
	2= Moderate Power: Analytic sample size between 64 to 393 for person-level assignment, between 15 to 90 clusters per group for cluster-level assignment, or between 51 and 75 in a time-series design (20 percent to 80 percent chance of detecting an effect size of 0.20).
	1= Low Power: Analytic sample size between 26 and 63 per group for person-level assignment, between 5 and 14 clusters per group for cluster-level assignment, or between 15 and 50 in a time-series design (20 percent or less chance of detecting an effect size of 0.20).
	0= Insufficient: The sample is not sufficient to detect an effect. (In general, the N is fewer than 25 per group for person-level assignment, fewer than 5 clusters per group for cluster-level assignment, or fewer than 15 in a time-series design.

Specify treatment group sample size:	
Specify comparison group sample size:	
Specify number of observations (Time Series design):	
Notes:	

C. *Statistical Adjustment (if applicable)* 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. (Note 1: Some program studies, such as place and field studies in situational crime prevention, do not lend themselves to the use of statistical controls. In such cases, please choose not applicable.)

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 study. Reliability refers to the stability and consistency of the measures. Validity refers to the accuracy of the measures. 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. In addition to what is reported in the article, please use your broader knowledge of the literature in the field to judge the degree to which the instrumentation reflects your field's standards for measurement.

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 can be attributed to the program. The validity of a study depends on both the research design and the measurement of the program activities and outcomes. Threats to internal validity will affect the accuracy of the results and draw into question the effect of the intervention.
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 program effects, at the time of the postprogram outcome measure. The study should have low overall attrition of study participants and minimal differential attrition between the treatment 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. Further, the attrition rate should be approximately the same for the treatment and control groups. Severe differential attrition makes the results suspect, because it may compromise the comparability of the groups. For more information on important considerations of attrition, such as specific guidance on acceptable rates of overall and differential attrition, please see the What Works Clearinghouse brief on Attrition Standards: What Works Clearinghouse brief on Attrition Standards 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 the instrument. Notes:
☐	Selection	This threat occurs when the groups to be compared differ on factors besides the treatment. 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 treatment. Notes:
☐	Other	Other threats may include: multiple treatment 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 program ends to ascertain its sustained effects. For example, in an evaluation of a 3-month long drug treatment intervention, the researchers assessed the outcomes at baseline (time 0), immediately following the end of the program (time 1), and 6 months after the end of the program (time 2). The follow-up period in this case should be rated a "1," as the time between the end of the program and the assessment period is equal to 6 months. Some programs, however, do not have clearly defined endpoints or are ongoing. In such cases, the period to be rated extends from the baseline to the last assessment period. For example, in a 2-year study of parolees placed on GPS monitoring for life, there is no clearly defined endpoint because the intervention is ongoing. In this case, the follow-up period is rated a "3," as the assessment period is more than 1 year (that is, from baseline to 2 years).

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:

G. *Displacement/Diffusion/Anticipatory Benefits (if applicable)* assesses the degree to which the evaluation examined for the presence of any crime displacement, diffusion of benefits, or anticipatory benefits surrounding the program implementation. (Note: This type of examination typically occurs in the evaluation of community level crime prevention efforts. The examination may involve one or many inspections and any form of displacement or diffusion, whether spatial, temporal, target, tactical, or offense.)

Rating	Points and Description
	3= Central (assesses displacement as integral part of the evaluation and includes appropriate research design containing at least one treatment area, one buffer area, and one control area).
	2= Post-hoc (secondary assessment of displacement or diffusion with demonstration/presentation of indicators).
	1= Cursory (brief mention of displacement or diffusion but no demonstrated examination).
	0= None (displacement or diffusion effects should have assessed but were not).
	NA= Not applicable.

Notes:

DESIGN QUALITY SCORING TABLE		
	Research Design Points	**FALSE**
+	**Sample Size Points**	**FALSE**
+	**Statistical Adjustment Points**	**FALSE**
+	**Instrumentation Points**	**FALSE**
+	**Internal Validity Points**	**FALSE**
+	**Follow-Up Period Points**	**FALSE**
+	**Displacement/Diffusion/ Anticipatory Benefits Points (if applicable)**	**0**
=	**TOTAL**	**0**
/	**NUMBER OF ITEMS (Change if needed; see note to right)**	**False**
=	**Design Quality Score**	**#VALUE!**

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

Please note: If Item G is scored (Displacement/Diffusion/Anticipatory Benefits), please increase the number of item (e.g., change from 6 to 7). Similarly, if Item C (Statistical Adjustment) is scored N/A, please reduce the number of items by 1.

PROGRAM FIDELITY

A. *Documentation* refers to the process of recording information about program fidelity (i.e., the degree to which the core program services or components are implemented as designed via the program description). To effectively establish causality, program designers should operationally define the core components of the program that are necessary and sufficient to achieve the outcomes desired. The implementation of these core components should then be empirically assessed and recorded to determine if the program under study meets a minimum threshold of implementation. Program evaluation studies should then include these measures of implementation fidelity to identify the underlying casual mechanism of the program.

Rating	Points and Description
	3= The collection of program implementation evidence is systematic and measured quantitatively (dosage, time spent in training, adherence to guidelines or a manual, etc.).
	2= The collection of program implementation evidence is systematic and assessed qualitatively (non-numeric data obtained through direct means, such as site observations, staff interviews, focus groups, etc.).
	1= The collection of program implementation evidence is non-systematic (ad hoc), incomplete, and/or assessed anecdotally.
	0= No information about of program implementation.

Notes:

B. *Adherence (directional indicator).* This section is concerned with assessing whether the participants received the proper amount, type, and/or quality of service or treatment. In other words, it is a measure of the degree to which the core program services were implemented as designed. An assessment of adherence is important because many programs that fail to show impacts suffer from a failure to deliver the intervention as specified (implementation failure). In general, there are three types of implementation failure: 1) no, or not enough, treatment; 2) the wrong treatment; or 3) unstandardized treatment. Judgments regarding adherence should be made on the available information and necessitate an inherent degree of subjectivity. The reviewer should review the available evidence to rate the sufficiency of the service delivery.
Note 1. This item is meant to credit programs that demonstrate adherence to the program design. A score of –1 should be applied only in cases where the researchers demonstrate poor adherence. The absence of information on adherence should be rated a "0."
Note 2. The rating concerning adherence should be focused on data about service delivery from the study under review, and not the program in general, as the service delivery at the point of evaluation is of interest.

Rating	Points and Description
	1= Adherence to program appears satisfactory.
	0= No information about program implementation.
	-1= Adherence to program appears poor.

Notes:

PROGRAM FIDELITY SCORING TABLE		
	Documentation Points	FALSE
=	TOTAL	FALSE
/	NUMBER OF ITEMS	FALSE
=	SUB TOTAL	#VALUE!
x	ADHERENCE: DIRECTIONAL	FALSE
=	FIDELITY EVIDENCE SCORE	FALSE

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

OUTCOME EVIDENCE

A. *Substantive Program Effects* estimates the magnitude of the program effect and rates the level of confidence that an effect is the result of the program rather than other factors (such as the selection process or by chance). In short, it is the difference between the outcome level attained with participation in a program and that which the same individuals would have attained had they not participated in the program. The substantive program effects rating is divided into primary and secondary outcomes, with primary outcomes given three times the weight of secondary outcomes. State the intent/core purpose of this program in the box as indicated. Use the following scale to assess the program's achievement of each of the outcomes. Estimate the magnitude of each identified program effect, and rate it by using Cohen's "rules of thumb" (described below). When estimating the magnitude of a program effect, please note the following:

- Assess only the results related to the full sample of study participants; do not assess results related to any subgroups examined in the study.
- Only significant findings (that is, $p \leq 0.05$, two-tailed) receive a score greater than zero.
- When using Wilson's calculator to estimate effect sizes, please be sure to input the appropriate sample size: For individual-level assignment, use the number of individuals in the analytic sample, and for a cluster-level assignment, use the number of clusters in the analytic sample, regardless of whether 1) the analysis is at the individual level or the cluster level or 2) the analytic method adjusts for clustering (for example, like a multilevel model).

SCORING GUIDELINES

Points	Description
3 =	The finding provides very strong evidence of a *positive* program effect (significant finding; large effect).
2 =	The finding provides moderate evidence of a *positive* program effect (significant finding; moderate effect).
1 =	The finding provides marginal evidence of a *positive* program effect (significant finding; small effect).
0 =	The finding provides no evidence of a program effect (comparison groups do not differ; no effect).
-1 =	The finding provides marginal evidence of a *negative* program effect (significant finding; small effect).
-2 =	The finding provides moderate evidence of a *negative* program effect (significant finding; moderate effect).
-3 =	The finding provides very strong evidence of a *negative* program effect (significant finding; large effect).

Only significant findings receive a score greater than zero. In order to determine the magnitude of the effect size, **CrimeSolutions uses Cohen's 'rules-of-thumb"** in order to score an outcome as either having a small, medium, or large effect. **Please use the rules laid out below:**

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

small	**=**	**0.20**
medium	**=**	**0.50**
large	**=**	**0.80**

For Odds Ratios (OR):

small	**=**	**1.50**
medium	**=**	**2.50**
large	**=**	**4.30**

For Correlation Coefficients (*r*):

small	**=**	**0.10**
medium	**=**	**0.25**
large	**=**	**0.40**

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

Intent/Core Purpose of the Program *(In one sentence, state the intent/core purpose of the program.)*

Core Purpose/Intent:

PRIMARY OUTCOMES CHART

	PRIMARY OUTCOMES	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	WEIGHTED SCORE
Primary Outcome 1				x 3	0
Primary Outcome 2				x 3	0
Primary Outcome 3				x 3	0
Primary Outcome 4				x 3	0
Primary Outcome 5				x 3	0
	SUM		*	0	0

SECONDARY OUTCOMES CHART

	SECONDARY OUTCOMES	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	WEIGHTED SCORE
Secondary Outcome 1				x 1	0
Secondary Outcome 2				x 1	0
Secondary Outcome 3				x 1	0
Secondary Outcome 4				x 1	0
Secondary Outcome 5				x 1	0
	SUM			0	0

CALCULATION WORKSHEET

	SUM OF WEIGHTED SCORE		SUM OF WEIGHT VALUES		
Primary Outcomes	0		0		
Secondary Outcomes	0		0		**SUBSTANTIVE PROGRAM EFFECTS SCORE**
Total	0	÷	0	=	#DIV/0!

*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 program demonstrates change (or changes) in behavior. Programs that demonstrate behavioral change (reductions in criminal behavior, substance abuse, and the like) are considered more effective than programs that demonstrate changes only in knowledge, beliefs, or attitudes (knowledge of characteristics of healthy relationships, attitudes about the acceptability of delinquent behavior, etc.). This is because behavior is not always consistent with a person's attitudes and beliefs—behavior that reflects a given attitude may be suppressed because of a competing attitude, or in deference to the views of others. Notably, behavior change does not need to be limited to individual behavior but may also include organizational change or changes in community-level behavior, such as an increase in convictions or a drop in crime rates. A drop in arrests in a particular group or community may also be considered behavioral change. This item relies on identifying both what constitutes a behavioral measure versus a measure of attitude/belief/knowledge, and where the preponderance of evidence lies. Please consult the Training Manual for elaboration on each choice below:

Rating	Points and Description
	3= The preponderance of the findings provides strong evidence of behavioral or systemic change (consistent and mostly significant findings; large effects), and may also provide evidence of change in knowledge, beliefs, and/or attitudes. A score of "3" is appropriate when the preponderance of behavioral scores have been scored at least a "1" and more likely a "2" or "3" in Outcome Evidence. A lack of significant attitude/belief/knowledge scores should not be used to justify lowering this score, whereas the presence of significant attitude/belief/knowledge scores could be used to justify raising a score that was between a "2" and a "3" to a "3."
	2= The preponderance of the findings provides moderate evidence of behavioral change or systemic (inconsistent but some significant findings; small to moderate effects) and may also provide evidence of change in knowledge, beliefs, and/or attitudes. A score of at least "2" is appropriate when the preponderance of behavioral scores have been scored at least a "1" in Outcome Evidence. A lack of significant attitude/belief/knowledge scores should not be used to justify lowering this score.
	1= There is marginal evidence or no evidence of significant behavioral results, but there are significant results for a preponderance of attitude/belief/knowledge scores.
	0= The findings provide no evidence (the groups do not differ; no attitudinal or behavioral effect) or evidence of negative behavioral, systemic, or attitudinal/knowledge change. A score of "0" is appropriate when the preponderance of behavioral scores have been scored "0" or less than "0" in Outcome Evidence.

Notes:

OUTCOME EVIDENCE SCORING TABLE		
	Substantive Program Effects Points	**#DIV/0!**
+	**Behavior Points**	**FALSE**
=	**TOTAL**	**#DIV/0!**
/	**NUMBER OF ITEMS**	**FALSE**
=	**OUTCOME EVIDENCE SCORE**	**#DIV/0!**

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 program effects while a negative value indicates negative program effects. A zero indicates no effect.

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 (Inconclusive Evidence) 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 program and be valid measures relating to the program's purpose. The reviewer should take into account if the specified outcomes match the intent of the program.

Anomalous Findings: Anomalous findings may contradict the intent of the program 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	Program Fidelity
Overall Score*		#VALUE!	#VALUE!	#DIV/0!	FALSE

*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 (Inconclusive Evidence) 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 program 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
#VALUE!	**Class 1 (Effective)**	**This study must have exceptional scores (at least 2.0) on all four dimensions of program effectiveness.** In general, this study demonstrates strong evidence in favor of the program when evaluated with a design of high quality (randomized controlled trial or quasi-experimental) and implemented with sufficient fidelity.
#VALUE!	**Class 2 (Promising)**	**This study must have above-average scores (at least 1.5) on the design-quality and outcome-evidence dimensions.** In general, this study demonstrates promising (perhaps inconsistent) evidence in favor of the program when evaluated with a design of high quality (quasi-experimental). More extensive research is required.
#VALUE!	**Class 3 (Ineffective)**	**This study must have a poor score (less than 0) on the outcome-evidence dimension, but have exceptional scores (at least 2.0 in design and fidelity) on other dimensions of program effectiveness.** In general, when implemented with sufficient fidelity and using an evaluation design of high quality (quasi-experimental), this study demonstrates negative program effects.
#VALUE!	**Class 4 (Null Effect)**	**This study must have a neutral score (from 0 to 1.4) on the outcome-evidence dimension, but have exceptional scores (at least 2.0 in design and fidelity) on other dimensions of program effectiveness.** In general, this study demonstrates no evidence in favor of the program when evaluated with a design of high quality (quasi-experimental) and implemented with sufficient fidelity.
#VALUE!	**Class 5 (Inconclusive Evidence)**	**This study must have a neutral score (less than 1.5) on the design-quality dimension, or alternatively, have scores for the dimensions that do not meet the stated thresholds for Classes 1–4.** This study could not be rated due to inconclusive evidence. (Note: Programs with inconclusive evidence will appear on the list of "Programs reviewed but not assigned a rating" on the CrimeSolutions.gov Web site.)

Integration of Evidence

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

- A program will be listed as "**Effective" if it has at least one *Class 1* study.**
- A program will be listed as "**Promising**" **if it has at least one Class 2 study.**
- A program will be listed as "**No Effects" if it has at least one *Class 3 or Class 4* study.**

2 The conceptual framework and program dimensions are effect modifiers. These modifiers will not be used to exclude a program from inclusion in the Crime Solution Resource Center, but will be applied as a gauge to increase confidence regarding the underlying causal mechanism of the program.

Please type your first and last name in the box above

Note: by electronically signing in the box, you are verifying that the information you have entered on Part 1 and Part 2 of this instrument is correct, and that you agree with the final scores and study rating.