NOTICE: This version of the Program Scoring Instrument was replaced by CrimeSolutions.gov Program Scoring Instrument Version 2.0 in September 2017.

CRIMESOLUTIONS.GOV: PROGRAM EVIDENCE RATING INSTRUMENT—PART 2

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:			
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 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[1]. 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: Not all designs easily fit into this hierarchy. The reviewer should specify the design and note the reason for the score.)*

CHECK	POINTS	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:

[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 program effects. 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 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 26 to detect a large effect (d=.80). It should be noted however that these figures are guidelines to help direct the review. (*Note 1: For three groups, the N per group drops to roughly 322 for a small effect, 52 for a medium effect, and 21 for a large effect. Group size continues to drop as the number of groups increases. Note 2: 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 generally considered the minimum.)* The reviewer should use his or her expertise to assess the adequacy of the sample.

CHECK	POINTS	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.)
	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.)
	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.)
	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.)

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.)*

CHECK	POINTS	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 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.

CHECK	POINTS	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.

CHECK	POINTS	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 post-program 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. Furthermore, 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. 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. In cases where programs do not have clearly defined endpoints, the follow-up period may be delimited between the first and last assessment period.

CHECK	POINTS	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.)*

CHECK	POINTS	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).
	N/A	Not applicable.

Notes:

DESIGN QUALITY SCORING TABLE	
	Research Design Points
+	**Sample Size Points**
+	**Statistical Adjustment Points**
+	**Instrumentation Points**
+	**Internal Validity Points**
+	**Follow-Up Period Points**
+	**Displacement/Diffusion/Anticipatory Benefits Points (if applicable)**
=	**TOTAL**
/	**NUMBER OF ITEMS** (FILL IN)
=	**DESIGN QUALITY SCORE**

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.)*

OUTCOME EVIDENCE

A. SUBSTANTIVE PROGRAM EFFECTS 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). **State the intent/core purpose of this program in the box on the next page.** The core purpose and primary outcomes should relate to one of the major areas of Crime Solutions (reducing crime/delinquency, improving the justice system, responding to victims, etc.). **Select and score primary and secondary outcomes (up to five each).** Secondary outcomes should relate to the ancillary purposes of the program. Scores for primary outcomes are given three times the weight of secondary outcomes. Use the following scale to assess the program's achievement of each of the outcomes. **Be sure to focus on the core purpose and primary outcomes of the program as these are most relevant to Crime Solutions.** See example below.

POINTS	DESCRIPTION
3 =	The finding provides very strong evidence of a program effect (significant finding; large effect).
2 =	The finding provides moderate evidence of a program effect (significant finding, moderate effect).
1 =	The finding provides marginal evidence of a program effect (significant finding, small effect).
0 =	The finding provides no evidence of a program effect (comparison groups do not differ, no effect).

EXAMPLE

PROGRAM INTENT/CORE PURPOSE: The main intents of the Hawaii Opportunity Probation with Enforcement (HOPE) program are to reduce drug use, recidivism, and incarceration among probationers who are considered at high-risk of failing probation or returning to prison.

	PRIMARY OUTCOMES	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	WEIGHTED SCORE
Primary Outcome 1	Reduce positive urine	Comp. group had significantly higher (46%) positive urinalyses versus 13% for HOPE.	3	x 3	9
Primary Outcome 2	Reduce re-arrest rates	47% of comparison group were arrested compared with 21% of HOPE participants.	3	x 3	9
Primary Outcome 3	Reduce # days incarcerated	Participants spent an average of 48% fewer days incarcerated (138 days vs. 267 days).	2	x 3	6
	Sum			9	24

	SECONDARY OUTCOMES	FINDINGS	UNWEIGHTED SCORE	WEIGHT VALUE	WEIGHTED SCORE
Secondary Outcome 1	Reduce probation revocations	HOPE participants had 7% revocation rate compared with 15% for comparison group (statistically significant).	1	x 1	1
Secondary Outcome 2	Reduce no-shows for prob. appointments	HOPE participants were significantly (61%) less likely to skip or miss appointments than the comparison group (9% vs. 23%).	3	x 1	3
	Sum			2	4

CALCULATION WORKSHEET

	SUM OF WEIGHTED SCORE		SUM OF WEIGHT VALUES		
Primary Outcomes	24		9		
Secondary Outcomes	4		2		**SUBSTANTIVE PROGRAM EFFECTS SCORE**
TOTAL	**28**	**÷**	**11**	**=**	**2.5**

INTENT/CORE PURPOSE OF THE PROGRAM *(In one sentence, state the intent/core purpose of the program.)*

PRIMARY OUTCOMES CHART

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

SECONDARY OUTCOMES CHART

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

CALCULATION WORKSHEET

	SUM OF WEIGHTED SCORE	SUM OF WEIGHT VALUES	
Primary Outcomes			
Secondary Outcomes			**SUBSTANTIVE PROGRAM EFFECTS SCORE**
TOTAL		÷	=

*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(s) in behavior. Programs that demonstrate behavioral change (reductions in criminal behavior, substance abuse, etc.) are considered more effective than programs that demonstrate changes only in knowledge or attitudes, because behavior does not always conform to a person's feelings and beliefs. Behavior that reflects a given attitude may be suppressed because of a competing attitude, or in deference to the views of others. *(Note 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. Note 2: Behavior change could include substantive program effects mentioned in A above.)*

CHECK	POINTS	DESCRIPTION
	3 =	The preponderance of the findings provides strong evidence of behavioral or systemic change (consistent, mostly significant findings; large effects).
	2 =	The preponderance of the findings provides moderate evidence of behavioral change or systemic (inconsistent but some significant findings, small to moderate effects).
	1 =	The preponderance of the findings provides evidence of attitudinal/knowledge change but only marginal evidence of behavioral or systemic changes (significant attitudinal findings with varying effects, but small behavioral effects).
	0 =	The findings provide no evidence of behavioral, systemic or attitudinal/knowledge change (comparison groups do not differ, no attitudinal or behavioral effect).

Notes:

C. *OUTCOME* (*directional indicator*) indicates the direction of the effects based on the preponderance of the evidence. *(Note: This element is a multiplier.)*

CHECK	POINTS	DESCRIPTION
	1 =	The preponderance of evidence indicates positive effects.
	0 =	The preponderance of evidence indicates no effect.
	–1 =	The preponderance of evidence indicates negative effects.

Notes:

OUTCOME EVIDENCE SCORING TABLE		
	Substantive Program Effects Points	
+	**Behavior Points**	
=	**TOTAL**	
/	**NUMBER OF ITEMS**	**2**
=	**SUB TOTAL**	
X	**DIRECTIONAL INDICATOR**	
=	**OUTCOME EVIDENCE SCORE**	

***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 a neutral effect.

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.

CHECK	POINTS	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*) refers to the degree to which the core program services or components are implemented/delivered as designed (via the program description). Adequate adherence to program design is as important as the type of program. An effective program model can be rendered less effective if implemented poorly, without fidelity. *(Note: This element is a multiplier.)*

CHECK	POINTS	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	
=	TOTAL	
/	NUMBER OF ITEMS	1
=	SUB TOTAL	
x	ADHERENCE: DIRECTIONAL INDICATOR	
=	FIDELITY EVIDENCE SCORE	

***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.

REVIEWER CONFIDENCE/OVERRIDE OPTION

The Reviewer Confidence/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 Information) 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.

CHECK	POINTS	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.

OVERALL SCORE

	CONCEPTUAL FRAMEWORK	DESIGN QUALITY	OUTCOME EVIDENCE	PROGRAM FIDELITY
Overall Score*				

*Reviewer Confidence/Override Option: As a final step on the scoring instrument, Study Reviewers provide an assessment as to their overall confidence in the study design. If both Study Reviewers agree that there is a fundamental flaw in the study design (not captured in the Design Quality dimension) that raises serious concerns about the study's results, the study is removed from the evidence base and not factored into the program's evidence rating. If one reviewer invokes the Override Option and the other does not, the dispute resolution process will be used to classify the study.

STUDY CLASSIFICATION SYSTEM

The score in each of the four dimensions is calculated separately and used to assess each study.[2] 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
	Class 1 (Strong Evidence of Positive Effect)	**This study must have exceptional scores (at least 2.0) in 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 (quasi-experimental) and implemented with sufficient fidelity.
	Class 2 (Some Evidence of Positive Effect)	**This study must have above average score scores (at least 1.5) in the design 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.
	Class 3 (Strong Evidence of Negative Effect)	**This study must have a poor score (less than 0) in the outcome evidence dimension yet exceptional scores (at least 2.0 in design and fidelity) in 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.
	Class 4 (Strong Evidence of Null Effect)	**This study must have a neutral score (from 0 to 1.4) in the outcome evidence dimension yet exceptional scores (at least 2.0 in design and fidelity) in 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.
	Class 5 (Insufficient Information)	**This study must have neutral scores (less than 1.5) in design quality dimensions.** In general, there is insufficient evidence to rate this study. *(Note: Programs with only insufficient evidence will not receive an evidence rating.)*

[2] The conceptual framework and program fidelity dimensions are effect modifiers. These modifiers will not be used to exclude a program from inclusion in CrimeSolutions.gov, but will be applied as a gauge to increase confidence regarding the underlying causal mechanism of the program.

A PROGRAM'S EVIDENCE RATING

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

Program's Evidence Rating	Study Classification				
	Class 1 Strong Evidence of Positive Effect	**Class 2 Some Evidence of Positive Effect**	**Class 3 Strong Evidence of Negative Effect**	**Class 4 Strong Evidence of Null Effect**	**Class 5 Insufficient Information**
Effective Programs have strong evidence to indicate they achieve their intended outcomes when implemented with fidelity.	Must have **at least 1** study in Class 1.	May have up to 2 studies in Class 2.	Must have **0** studies in Class 3	May have **up to 1** study in Class 4	Studies do not determine Evidence Rating
Promising Programs have some evidence to indicate they achieve their intended outcomes.	Must have 0 studies in Class 1	Must have **at least 1** study in Class 2	Must have **0** studies in Class 3	May have **up to 1** study in Class 4	Studies do not determine Evidence Rating
No Effects Programs have strong evidence indicating that they had no or harmful effects when implemented with fidelity.	Must have **0** studies in Class 1	Must have **0** studies in Class 2	Must have **at least** 1 study in either Class 3 or Class 4		Studies do not determine Evidence Rating