CrimeSolutions.gov
Practices Scoring Instrument

Contents

Part I. Screening

Step 1. Identifying an Eligible Meta-Analysis

What qualifies as an eligible meta-analysis? If 'NO' to any of the items below, the research review does not qualify as a meta-analysis eligible for inclusion in CrimeSolutions.gov. If 'YES' to all items, proceed to Step 3.

Checklist (see detailed instructions for each item following the checklist)

YES	NO	
☐	☐	1. Intervention. Does the meta-analysis include at least two studies of the intervention of interest?
☐	☐	2. Aggregation. Does the meta-analysis aggregate the results from at least two studies? NOTE: If the meta-analysis presents results individually for at least two studies, but elects not to aggregate results across the studies, seek advice from an expert meta-analyst as to whether it is possible and/or justifiable to combine findings from at least two of the included studies.
☐	☐	3. Primary aim of the intervention. Do the programs included in the meta-analysis aim to: reduce crime, delinquency, overt problem behaviors, or victimization; improve justice system practices; or target an offender or at-risk population?
☐	☐	4. Literature search. Did the literature search include at least two sources and provide evidence that unpublished literature was sought in the search?
☐	☐	5. Primary outcomes. Does the meta-analysis report on at least one eligible outcome (defined below)?
☐	☐	6. Control groups. Do all studies included in the meta-analysis include an appropriate control, comparison, or counterfactual condition?
☐	☐	7. Reporting of results. Does the meta-analysis report effect sizes that represent the magnitude of the treatment effect?
☐	☐	8. Combining different types of relationships. If a quantitative synthesis of effect sizes is reported (i.e., a mean effect size is reported for multiple studies), do all effect sizes in the combination index the same type of *relationship*?
☐	☐	9. Publication date. Were at least 50% of the studies included in the meta-analysis published/available on or after 1980?
☐	☐	10. Age of samples. Are the included samples in the meta-analysis restricted to either adults or juveniles? Or, are mean effect sizes for adults and juveniles reported separately?

Part II. Quality Ratings

Step 1. Quality Ratings on Individual Items

Within each program and practice category, use the following criteria to rate each meta-analysis defined as eligible for consideration in Part II. *Note: prior to completing the quality coding of a meta-analysis, it is important to identify all other published copies of that meta-analysis or its protocol*. This will often be necessary when coding meta-analyses that were published in journal articles and Campbell or Cochrane Collaboration reviews (which will generally provide more detailed information that will be useful during coding).

Quality Rating (see detailed instructions for each item on following pages)

Rating (1-3)	
	A. Eligibility Criteria
	B. Comprehensive Literature Search
	C. Grey Literature Coverage
	D. Coder Reliability
	E. Methodological Quality
	F. Outlier Analysis
	G. Handling Dependent Effect Sizes
	H. Effect Size Reporting
	I. Weighting of Results
	J. Analysis Model
	K. Heterogeneity Attentiveness
	L. Publication Bias

INCLUSION CRITERIA

A. *ELIGIBILITY CRITERIA* rates the degree to which the meta-analysis provides a clear, detailed statement of the inclusion and exclusion criteria used to determine whether primary studies were eligible for inclusion in the final meta-analysis. For meta-analyses focused on intervention effects, these will generally involve a description of the eligible study participant populations, intervention types, comparison groups/study designs, crime/delinquency outcomes, and other study characteristics such as publication year or country of study. Meta-analyses with clearly defined eligibility criteria should outline their inclusion and exclusion criteria in enough detail to permit replication of study selection in a future meta-analysis.

Check	Points	Description
	3 =	*Clear PICO Criteria.* The meta-analysis clearly defines what types of studies were included or excluded. Specifically, the meta-analysis must outline what types of Participant populations, Interventions, Comparison groups/study designs, and Outcomes were eligible. The PICO criteria must be explicitly presented or CLEARLY implicit without any ambiguity. That is, some meta-analyses may not specify the eligible participant population but the eligible participants are clear from the nature of the programs, e.g., drug courts are clearly intended for substance using offenders.
	2 =	*Vague or Incomplete Criteria.* The meta-analysis makes vague mention of eligibility criteria, but not at a level of specificity that would permit replication of study selection. Or, the meta-analysis may provide an incomplete listing of eligibility criteria, e.g., include some but not all of the PICO eligibility criteria (e.g., describes eligible intervention and study designs, but not participants). Note that if a PICO criterion is not specified but the nature of that criterion is implicit (e.g., the meta-analysis is about sex offender programs but the eligible participants are not specified), the meta-analysis would be coded as a "3" above.
	1 =	*Cannot Tell.* The meta-analysis does not specify the criteria used to select studies eligible for inclusion.

LITERATURE COVERAGE

B. *COMPREHENSIVE LITERATURE SEARCH* rates the degree to which the meta-analysis conducted an exhaustive, comprehensive review of the literature in an attempt to identify all eligible studies. Ideally, a meta-analysis should be based on a comprehensive literature search with diverse search methods including some combination of searches of electronic bibliographic databases, web-searching (e.g., government agencies, professional organizations), hand-searches of journals, personal contact with researchers in the field, forward- and backward-citation searching, etc. The purpose of this rating is to give more weight to those meta-analyses that are based on comprehensive and diverse sources for identifying potentially eligible studies.

Check	Points	Description
	3 =	*Comprehensive Search.* The literature search was (a) comprehensive: included more than two sources of literature, AND (b) diverse: included some combination of literature sources such as searches of bibliographic databases, searches of Web sites, hand-searches of journals, contacting authors, forward-citation searching, backward-citation searching (harvesting references), etc.
	2 =	*Bibliographic Databases Only.* The literature search included more than two sources, but relied solely on searches of bibliographic databases (e.g., PsycINFO, PubMed, CSA, ERIC, Web of Science, IBSS). No alternative means of identifying literature are specified.
	1 =	*Two Sources Only.* The meta-analysis relies on only two sources of literature (this could be electronic databases or other types of literature sources).
		Note: If the meta-analysis does not include at least two literature sources, it is not eligible for inclusion on CrimeSolutions.gov

C. *GREY LITERATURE COVERAGE* assesses the extent to which a meta-analysis includes results from unpublished or "grey" literature sources. Grey literature is defined as "that which is produced on all levels of government, academia, business and industry in print and electronic formats, but which is not controlled by commercial publishers." Sources of grey literature or unpublished studies include dissertations, theses, government reports, technical reports, conference presentations, and other unpublished sources. Note that books and book chapters are NOT considered grey literature or unpublished research given that they are controlled by commercial publishers. A meta-analysis should always attempt to include grey literature due to consistent evidence that the nature and direction of research findings is often related to publication status. Meta-analyses that do not include grey literature may provide biased estimates of effects, and in many cases, may provide inflated estimates of intervention effects. This item presumes that grey literature exists in the research literatures relevant to CrimeSolutions.gov, and meta-analyses that fail to locate grey literature, therefore, have lower quality search strategies (i.e., the search did not adequately attempt to find grey literature).

Note also that a meta-analysis that includes only published research but requests unpublished data from primary study authors is not considered to include grey or unpublished research.

Check	Points	Description
	3 =	*Includes Grey Literature.* The meta-analysis includes grey literature/unpublished studies. Note that this refers to studies included in the actual systematic review/meta-analysis; it does not refer to studies that are listed in the reference list of the document but were not included in the actual synthesis of the research.
	2 =	*Attempted But Didn't Locate Grey Literature.* The meta-analysis attempted to include grey literature during the search procedure, but no unpublished reports were ultimately included in the meta-analysis.
		Note: If the literature search does not include an effort to locate unpublished studies, or is explicitly restricted to published literature, it is not eligible for inclusion on CrimeSolutions.gov

DATA COLLECTION PROCEDURES

D. *CODER RELIABILITY* assesses how the authors of the meta-analysis handled reliability of data extraction from the primary research reports. Ideally, two or more coders would extract all pieces of information from each eligible research report and reliability statistics would be used to assess coder reliability and/or consensus would be reached on all items. Sometimes double-coding is not feasible due to time or financial restraints, or authors may only double-code subsets of the entire meta-analysis dataset. The goal of this rating is to give higher scores to those meta-analyses that adequately assess and document reliability of study coding.

Check	Points	Description
	3 =	*Coder Reliability Attained.* The meta-analysis had (a) two or more coders coding at least some portion of the studies, and inter-coder reliability information is reported, OR (b) all studies were double-coded (and presumably any identified coding differences were resolved).
	2 =	*Moderate Coder Reliability.* The meta-analysis had (a) two or more coders coding at least some portion of the studies, but no information given on inter-coder reliability, OR (b) two or more coders but differences were resolved only on a subset of studies or variables.
	1 =	*Inadequate Coder Reliability.* The meta-analysis (a) does not report how many coders were used to extract data and there is no information on inter-coder reliability, OR (b) only had one coder.

STATISTICAL ANALYSES

E. *METHODOLOGICAL QUALITY* assesses the extent to which the authors of the meta-analysis were aware of and attentive to the methodological quality of the studies included in the meta-analysis. The goal is to give greater weight to those meta-analyses that indicate the authors were aware of and appropriately handled (if necessary) the effect of study quality on the meta-analysis findings. There are many ways a meta-analysis may assess methodological quality including but not necessarily limited to:

- calculating quality scores based on some criteria,
- categorizing studies into those with low, medium, high quality, etc.,
- using a "risk of bias" tool or some other quality checklist to assess methodological quality,
- summarizing the number of studies with desired methodological attributes such as random assignment, intent-to-treat analyses, low attrition, blinding of personnel or outcomes, allocation concealment, or measurement reliability.

In addition to assessing the methodological quality of included studies, it is important for a meta-analysis to address whether quality had any effect on the main findings (e.g., does it bias the results). There are several ways a meta-analysis may adjust for and address methodological quality, including but not necessarily limited to:

- presenting results split out separately by some quality characteristic (e.g., quality score, risk of bias score, study design, study attrition),
- including methodological variables in a multivariate analysis to statistically adjust findings for quality,
- reporting meta-regression or ANOVA results of the effect sizes by study design, risk of bias score, or some other quality characteristic,
- reporting the correlation between risk of bias or study quality and the effect sizes.

Check	Points	Description
	3 =	*Assessed and Addressed Quality.* The meta-analysis (a) explicitly describes the methodological quality of the included studies and (b) provides some evidence that either quality did not bias or influence the findings OR reports results split out separately by some quality characteristic. It is not sufficient for the meta-analysis to state that study quality was measured; it must also provide at least some description of the actual quality of the included studies, and provide evidence that quality did not bias the results or report the results separately or use statistical adjustments to ensure quality did not bias the results.
	2 =	*Vague Quality Assessment.* The meta-analysis vaguely mentions that the quality of included studies was assessed, but does not explicitly state the actual quality of the included studies. Or, the meta-analysis included reportedly poor quality studies, but provides no evidence that these poor quality studies were handled in a way so as not to bias the results.
	1 =	*No Quality Assessment or Cannot Tell.* The meta-analysis does not explicitly address the quality of the studies included or the potential for bias.

F. *OUTLIER ANALYSIS* assesses whether the meta-analysis checks for effect size outliers in the data. Note that this item refers to outlying effect sizes included as data points in the meta-analysis, not outlying data in the primary research studies that contributed to the meta-analysis. Extreme outliers can potentially distort the overall mean effect size estimate as well as the results of other analyses. An extreme effect size with a large weight (i.e., from a large study) is more problematic than an extreme effect size with a relatively small weight. Thus, a visual outlier on a forest plot with a very large confidence interval is less problematic than one with a small confidence interval. Note that it is possible for a meta-analysis to have a large amount of variability or heterogeneity in the effect sizes estimates, yet have no outliers. The degree of heterogeneity in a meta-analysis therefore does not inherently provide information about the presence of outliers. Although not all meta-analyses may have effect size outliers, this rating assumes that meta-analyses are of higher quality if they check for outliers and, if found, make some attempt to assess or control their influence.

Check	Points	Description
	3=	*Outliers Explicitly Mentioned.* The meta-analysis explicitly reports that distributions were checked for outlying effect sizes and/or reports how outliers were handled in the analysis (this includes establishing that there were none).
	2 =	*No Outliers Mentioned or Cannot Tell.* The meta-analysis does not mention outliers or report examining whether there were any outlying effect sizes.
	1 =	*Outliers Identified but not Addressed.* The meta-analysis explicitly reports outlying effect sizes; however, the meta-analysis does not report how, or if, they were handled in the analysis.

G. *HANDLING DEPENDENT EFFECT SIZES.* This item rates a meta-analysis based on its appropriate analysis of effect sizes. The most common methods assume that all effect sizes within a given analysis are statistically independent, meaning that only one effect size per study sample should be included within a given analysis. Acceptable methods for handling statistically dependent effect sizes include (a) selecting only one effect size per study sample based on some decision criteria; (b) creating one effect size per study sample by averaging all or a selected set of effect sizes for that sample; (c) using multivariate methods that allow inclusion of multiple effect sizes per study and account for the covariance between effect sizes within a study sample; (d) using robust standard error procedures that allow inclusion of multiple effect sizes per study sample; and (e) running a series of smaller meta-analyses based on subsets of effect sizes such that only one effect size per study sample is included within a given analysis. Meta-analyses that have potential statistical dependencies among effect sizes and ignore them or do not describe how they were handled are assumed to produce incorrect standard errors and erroneous results for the associated statistical significance tests and confidence intervals.

In some cases, even if the authors do not explicitly state how dependent effect sizes were handled, it may be possible to ascertain by examining forest plots or tables that list the studies included within a given analysis. If each of the studies contributing to a given analysis appear to be unique, it is reasonable to infer that independent effect sizes were used. In other cases the forest plot or other itemizations may show multiple effect sizes for a given study reference or citation, which may or may not be appropriate (e.g., some study reports provide data for multiple independent subsamples). In those instances, the coder will need to examine the text of the meta-analysis to determine if those apparent dependencies are explained.

Check	Points	Description
	3 =	*Appropriate Handling of Dependent Effect Sizes.* The meta-analysis does not include more than one effect size from the same respondent sample within any analysis, or appropriate statistical procedures were used to handle dependent effect size estimates.
	2 =	*Cannot Tell.* It is unclear whether the meta-analysis data include more than one effect size per study. The meta-analysis may include more than one effect size per study, but there is ambiguity as to whether or how those dependent effect sizes were handled. Or, the meta-analysis does not provide enough information about the included studies to assess whether multiple effect sizes were included from each study.
	1 =	*Inappropriate Handling of Dependent Effect Sizes.* The meta-analysis explicitly includes multiple effect sizes per study within a single analysis, and those dependencies are ignored in the analysis.

H. *EFFECT SIZE REPORTING* assesses whether the meta-analysis reported an aggregate mean effect size that synthesized (averaged) effect sizes across one or more sets of multiple studies, and whether the meta-analysis also provided estimates of precision around the point estimate(s). Preference is given to meta-analyses that report mean effect sizes for the key findings along with confidence intervals around the mean, or standard errors of the mean effect size that would allow calculation of the confidence interval around the mean.

Check	Points	Description
	3 =	*Mean Effect Size(s) Reported with Confidence Intervals.* The meta-analysis reports mean effect sizes with confidence intervals or standard errors for each analysis conducted, or at least for the key outcomes of interest. Note that if a key part of the analysis/findings is related to mean effect sizes within certain subgroups (e.g., quality subgroups, population subgroups, intervention type subgroups), then it is essential that the meta-analysis report confidence intervals or standard errors for those subgroups, or at minimum, provide results from statistical tests (generally meta-regression coefficients, or a Q-test for between-group heterogeneity) comparing the difference in means across those groups.
	2 =	*Mean Effect Size(s) Reported without Confidence Intervals.* The meta-analysis reports mean effect sizes with no confidence intervals or standard errors, or the meta-analysis does not report confidence intervals or standard errors for the main subgroups of interest. This also includes cases where the meta-analysis does not include confidence intervals or standard errors, but includes other numeric values that could be used to approximate confidence intervals or standard errors (e.g., exact p-values).
	1 =	*No Mean Effect Sizes Reported.* The meta-analysis does not report a mean effect size but reports individual study effect sizes only. Seek an expert meta-analyst to determine whether this meta-analysis is eligible for inclusion (see Identifying an Eligible Meta-Analysis #2).

I. *WEIGHTING OF RESULTS* assesses whether the meta-analysis uses appropriate weighting schemes when estimating mean effect sizes and in other analyses in order to give greater weight to the effect sizes estimated with more precision (e.g., based on larger samples). Appropriate weighting schemes include inverse variance weighting and sample size weighting. Inappropriate weighting schemes include weighting effect sizes by quality scores or some other arbitrary author-defined function, or using no weights at all.

Coders can reasonably assume appropriate weighting even if the meta-analysis does not explicitly state what weights were used, but does mention using statistical software specifically designed for meta-analysis. Some of the most common meta-analysis software procedures and macros that may be reported are:

- CMA (Comprehensive Meta-Analysis)
- Meta-Analyst
- Meta-Win
- RevMan
- R packages designed specifically for meta-analysis (e.g., metafor)*
- SAS macros designed specifically for meta-analysis*
- SPSS macros designed specifically for meta-analysis*
- Stata macros designed specifically for meta-analysis*

For those programs indicated with an asterisk (R, SAS, SPSS, Stata), it is not sufficient for the meta-analysis author to have used that program. Rather, the meta-analysis must explicitly state that analyses were conducted using macros in these programs that were ***developed specifically for meta-analysis***. Analyses conducted within other programs (CMA, Meta-analyst, Meta-Win, RevMan) can be assumed to indicate that appropriate weighting methods were used. In some cases a meta-analysis may use one of the above for the overall analysis of effect sizes but then use standard ANOVA or OLS regression for the moderator analyses. These can easily be spotted: they will report F-values (F-tests). If F-values are associated with the moderator analyses, they have been performed incorrectly. Note that this item does not require the meta-analysis to use specialized meta-analysis software; we list these here for easy reference as an indicator of likely correct procedure if the authors do not explicitly discuss weighting procedures.

Check	Points	Description
	3 =	*Appropriate Weighting Used.* The meta-analysis uses inverse variance weights or sample size weights in all analyses with effect sizes.
	2 =	*Cannot Tell if Weighting Used.* It is unclear whether weighting was used.
	1 =	*No Weighting, or Quality or Other Weighting.* The meta-analysis does not use inverse variance or sample size weighting, but instead does not use any weighting, or uses some other type of index, such as a quality rating, to weight some or all effect size estimates.

J. *ANALYSIS MODEL* rates a meta-analysis based on whether the authors recognized and addressed the issue of random effects versus fixed effect analysis models. In most cases, a random effects analysis model will be preferred, although there are instances in which a fixed effect model may be appropriate. Preference is generally given to meta-analysis that estimate random effects models because they provide results that are generalizable beyond those studies included in the meta-analysis; allow for variability in effects across different types of studies, populations, interventions, etc.; and provide more conservative estimates of the precision of mean effect sizes. Ideally, meta-analysis authors will report the type of analysis model used; if not, one can assume they estimated random effects models if they report that weighting functions used estimates of within- and between-study variance/variability; or if they estimated the tau-squared random effects variance component/tau-squared estimate of between-studies variability.

Broadly, a fixed effect model will use inverse variance weights that only incorporate within-study variance, and may be shown as:

$$w = \frac{1}{V}$$

or

$$w = \frac{1}{SE^2}$$

where V and SE are the estimates of the effect size variance and standard error, respectively (the calculations for V/SE will vary depending on the effect size metric being used).

Random effects models will use inverse variance weights that incorporate an additional variance component which indexes variability between effect sizes in a distribution of true effect sizes, and may be shown as:

$$w^* = \frac{1}{V}$$

or

$$w = \frac{1}{V + \tau^2}$$

where V is the estimate of the within-study effect size variance and τ^2 is the estimate of the between-studies variance component.

There are some situations in which a fixed effect model may be justifiable if the meta-analysis is not intended to generalize to studies beyond those included in the meta-analysis, or the author has a specific reason to expect no unexplainable variability in the effect sizes. The meta-analysis must provide a substantive or statistical justification for the use of a fixed effect model to receive the highest quality rating.

Check	Points	Description
	3 =	*Random Effects or Justification for Fixed Effect Model.* The meta-analysis uses at least one random effects model to estimate mean effect sizes, or provides a substantive or statistical justification (i.e., justified based on a non-significant p-value for the Q statistic) for the use of one or more fixed effect models. If the meta-analysis includes both random effects and fixed effect models, it would receive this highest quality rating.
	2 =	*Fixed Effect Model without a Justification.* The meta-analysis (a) does not use any random effects models, AND (b) explicitly mentions the use of a fixed effect model but does not provide justification for using that model.
	1 =	*Unclear Model.* The meta-analysis does not clearly specify whether a random effects or fixed effect model was used to estimate mean effect sizes, and it is not possible to determine based on the reported information.

K. *HETEROGENEITY ATTENTIVENESS* rates a meta-analysis on whether the authors were aware of and attentive to heterogeneity (i.e., variability) in the effect size estimates from the studies in the meta-analysis. This rating is specifically interested in assessing whether authors conducted analysis to examine the heterogeneity in the distribution of effect sizes and explain any observed heterogeneity. A meta-analysis will commonly report one or more of the following heterogeneity statistics:

- τ^2 — often reported as tau-squared, other names are random effects variance component, between-studies variance component, estimate of between-study variability, variance of the distribution of effect size estimates.
- τ — often reported as tau, random effects standard deviation component, between-studies standard deviation, standard deviation of the distribution of effect size estimates.
- Q — often reported as Q, χ^2, Cochran's Q, Q-total, Q-test for homogeneity, Q-test for heterogeneity. This Q refers to the weighted sum of squares of the effect sizes for all of the studies included in a meta-analysis.
- I^2 — often reported as I-squared, may also be called the percentage of variability in effect sizes due to heterogeneity or between-study differences.

Authors who observe heterogeneity in the effect sizes may then attempt to explain some of that variability using moderator analysis, typically through the use of meta-regression models or analysis of variance models specifically designed for meta-analysis data.

Check	Points	Description
	3 =	*Assessed and Addressed Heterogeneity.* The meta-analysis explicitly reports at least one heterogeneity statistic in the text of the document, and appropriately handles any observed heterogeneity. Note that it is not sufficient for the meta-analysis to include heterogeneity statistics in a table or text; the meta-analysis must at some point reference or discuss at least one heterogeneity statistic in the text. In some situations, there may be little variability and thus the authors conduct no additional moderator analysis. However, in the presence of heterogeneity, the authors should report having conducted some sort of moderator analysis in an attempt to explain some of that heterogeneity.
	2 =	*Heterogeneity Referenced, but Reported Inadequately.* The meta-analysis makes reference to effect size heterogeneity statistics or homogeneity/heterogeneity in the effect size, but does not report the values for any heterogeneity statistics in the text, does not report the results of any heterogeneity tests in the text, or does not present results for subgroups based on heterogeneity. This rating also includes meta-analyses that report heterogeneity statistics somewhere in the text or tables/figures, but do not examine further with moderator analyses if warranted. Meta-analyses that report moderator analyses but do not explicitly report or discuss heterogeneity statistics also would be coded here.
	1 =	*No Heterogeneity Statistics Referenced or Cannot Tell.* The meta-analysis does not make reference to effect size heterogeneity or report any heterogeneity statistics or tests.

L. *PUBLICATION BIAS* rates the extent to which a meta-analysis investigates the potential for publication bias in the sample of included studies. Publication bias broadly refers to the idea that published results are more likely to include large and/or statistically significant effects, whereas unpublished results are more likely to include null, small, or "negative" (i.e., opposite of what would be predicted) effects. The purpose of this rating is to assess whether the meta-analyst was aware of and sensitive to the possibility that publication bias could influence the results of their analysis. Note that it is not sufficient for a meta-analysis to discuss publication bias in terms of the literature search, e.g., whether unpublished results were included or an attempt was made to find them. This rating refers to whether the meta-analysis descriptively or statistically assessed the possibility of publication bias in the results (this will primarily be reported in the results section, although occasionally may be present in the methods section of the meta-analysis). There are several techniques meta-analysts may use to statistically assess the possibility of publication bias. The most common techniques include:

- funnel plot graphs or contour-enhanced funnel plot graphs
- regression-based tests for funnel plot asymmetry (Egger test, Peters test, Harbord test, Begg rank correlation)
- cumulative meta-analysis graphs
- selection modeling approaches
- trim and fill analysis
- statistical comparison of effect sizes for published and unpublished studies
- inclusion of publication status in multivariate analysis, e.g. meta-regression

Check	Points	Description
	3 =	*Publication Bias Analysis Reported.* The meta-analysis uses statistical techniques (graphs or statistical tests) to assess the possibility of publication bias. It is acceptable if the meta-analysis reports publication bias analyses were conducted, but does not provide the actual results from those analyses (e.g., does not include a funnel plot). To receive this rating the meta-analysis must explicitly indicate that some formal statistical or graphical technique (outlined above) was used to assess possible publication bias. It is not sufficient for the meta-analysis to present descriptive results separately by publication type, or to provide subgroup confidence intervals by publication type. Note that the Fail-Safe estimate of the number of left out null studies (often called the Failsafe-N) required to make the mean effect size non-significant or zero, and analogous techniques, are not acceptable as the sole assessment of the potential for publication bias.
	2 =	*Inadequate Publication Bias Reporting.* The meta-analysis makes reference to the potential for publication bias but does not conduct any graphical or statistical analysis to assess it; or the meta-analysis only reports a Failsafe-N value to assess publication bias. If a meta-analysis only provides a descriptive comparison of published/unpublished studies (e.g., presenting means effect sizes for these two groups, but conducted no additional statistical tests), it also would be coded in this category.
	1 =	*Publication Bias Not Mentioned.* The meta-analysis does not mention the possibility of publication bias in the data.

Step 2. Overall Meta-Analysis Quality Rating

Add up the scores on the individual quality items in each of the following four groups:

1. Item E (Methodological Quality): 1 item;
2. Items G–K (Main Analysis): 5 items;
3. Items A–C (Eligibility & Search): 3 items; and
4. Items D, F, & L (Coder Reliability, Outlier Analysis, Publication Bias): 3 items.

Use the following matrix to determine the low, medium, or high scoring range for each group of items and the associated summary value subtotal.

Item E (Methodological Quality)	Low Score=1	Medium Score=2	High Score=3	Subtotal
	4	8	12	
Items G–K (Main Analysis)	Low Score=5–8	Medium Score=9–12	High Score=13–15	Subtotal
	3	6	9	
Items A–C (Eligibility & Search)	Low Score=4–5	Medium Score=6–7	High Score=8–9	Subtotal
	2	4	6	
Items D, F, & L (Coder Reliability, Outlier Analysis, Publication Bias)	Low Score=3–4	Medium Score=5–7	High Score=8–9	Subtotal
	1	2	3	
				Total Summary Score

Sum the summary score subtotals to determine the total summary score and categorize the overall Meta-Analysis Quality Rating as follows:

High Quality
Total Summary Score 24–30

Medium Quality
Total Summary Score 17–23

Low Quality
Total Summary Score 10–16

Part III. Evidence Summary and Validity Ratings

Use the *Evidence Summary and Validity Spreadsheet* to record the following information.

Step 1. Evidence Summary

Now that the relevant mean effect sizes have been selected, summarize the key information about each keeping in mind that a single meta-analysis may provide mean effect sizes for more than one program or population.

Meta-Analysis
Record a label for the meta-analysis (e.g., Jones 2010).

Program Category
Categorize the mean effect size you are coding into the CrimeSolutions.gov program scheme.

Subgroup
If relevant, record the subgroup.

Macro (Tier 1) Outcome
Categorize the outcome into the appropriate Tier 1 category

Micro (Tier 2) Outcome
Categorize the outcome into the appropriate Tier 2 category

Use in Final Rating
Indicate whether the outcome is to be used in the final evidence rating. For outcomes not to be used in the final evidence rating, the effect sizes and validity should still be recorded so that they may be included in the text description on the CrimeSolutions.gov Web site.

Type of ES

md –	unstandardized mean difference
d –	standardized mean difference (also *g* and *smd*)
r –	correlation coefficient
phi –	phi coefficient
LOR –	log odds ratio
OR –	odds ratio
LRR –	log risk ratio
RR –	risk ratio
Other –	other type of effect size (please note the type of effect size)

Unit on Which ES Is based

P – Persons

O – Other (e.g., incidents, location restricted crime, etc. for place-based studies)

Mean ES

Record the mean effect size for the listed outcome, in the effect size metric as specified in the "Type of ES" code.

Confidence Intervals

Record the lower and upper confidence intervals, standard errors, or both (if reported). Note that if a confidence interval other than 95% is reported, please consult an expert meta-analyst.

Standard Error

If confidence intervals are not reported, record the standard error of the mean effect size, which can be used to automatically calculate a 95% confidence interval for the mean effect size.

K

Record the number of individual effect sizes that are averaged into the mean effect size. This is typically the number of studies represented in that mean, with each study contributing one effect size. If some studies contribute more than one effect size from the same participant sample to the mean (dependent effect sizes) note both the number of effect sizes and the number of studies.

Model

FE – fixed effect model

RE – random effects or mixed effects model

Cannot tell

Significance

Yes – mean effect size is significantly different from the null value with alpha=.05 (i.e., $p \leq .05$).

No – mean effect size is not statistically significant at alpha=.05.

If statistical significance is not reported, seek an expert meta-analyst to determine the statistical significance of the mean effect size. Listed below are the null values associated with the most commonly reported effect sizes; thus, 95% confidence intervals that do not include these null values indicate statistically significant effects.

Effect size	Null value	Effect size	Null value
md	0	LOR	0
d	0	OR	1
r	0	LRR	0
phi	0	RR	1

Direction

Record whether the treatment or control group is favored.

TX – treatment group is favored (e.g., lower recidivism, less drug use)

CT – control group is favored

Equal – mean effect size is exactly zero (or one, in the case of odds ratios or risk ratios).

Step 2. Validity Ratings

For each mean effect size for each selected outcome, rate the internal validity and statistical conclusion validity of the research contributing to the mean effect size at issue on the following items.

A. *Internal Validity* refers to the extent to which the research design is free from threats that potentially bias the effect estimate. Randomized control trials (RCTs) have the strongest inherent internal validity and this item uses the extent to which the mean effect size is based on results from randomized controlled trials to assess the overall internal validity of the mean effect size being coded.

Recall that all ratings in this section are based on individual mean effect sizes, not on the entire meta-analysis. If the mean effect size for only RCT studies has been selected in the previous step, rate only the internal validity of that effect size. In cases where you cannot tell the proportions of RCTs/non-RCTs in the meta-analysis, always default to the lowest rating of 1.

Check	Points	Description
	3 =	*High Internal Validity.* For this rating, <u>at least one of the following criteria</u> must be met (i.e., not all criteria need to be met, but at least one must be met). If any of these criteria are met, then the mean effect size recorded above is the combined effect size (i.e., do not code the RCT and non-RCT mean effect sizes separately): (a) At least 60% of the studies included in the mean effect size are RCTs. (b) The mean effect size is covariate adjusted to estimate the effect size expected if all studies were RCTs. That is, the difference between non-RCTs and RCTs was statistically controlled in the analysis with results that produced a covariate adjusted mean effect size representing RCT outcomes. There must be at least 5 RCTs represented in that analysis to qualify under this criterion. (c) There are at least 5 RCTs, the mean effect sizes for the RCTs and non-RCTs are reported separately, and <u>at least one of the following conditions apply</u>: 1) A statistical test is reported that shows no statistically significant difference between the mean effect size for the RCT and non-RCT studies. 2) The mean effect sizes for the RCTs and non-RCTs both fall within an approximate fixed effect 95% confidence interval around the mean effect size for both combined. (If necessary, compute the combined mean as the weighted average of the RCT and non-RCT means, with each weighted by the number of studies contributing to the respective mean—see the formula below for assistance calculating the weighted average). The confidence intervals are defined for this combined mean standardized mean difference effect sizes as follows, based on the number of studies in the combined mean:

Check	Points	Description

High Internal Validity (con't.).

- 10 studies or fewer: ± .12 around the combined mean
- 11–20 studies: ± .09 around the combined mean
- 21–30 studies: ± .07 around the combined mean
- 31–50 studies: ± .06 around the combined mean
- 50 studies or more: ± .05 around the combined mean

3) The mean effect size for the RCTs is tested for statistical significance, the mean effect size for the combined RCTs and non-RCTs is in the same direction and is also tested for statistical significance, and the results are the same in both cases (both significant or both non-significant).

Notes:

A. This code refers to the studies summarized in individual effect size means, so if possible consider only the studies included in that mean when making this rating. If no information specific to those studies is available, it is permissible to use any relevant information provided about the broader set of studies from which those contributing to the mean effect size at issue were drawn so long as there is no overt indication that the subset may be unrepresentative of that broader set.

B. Regression-discontinuity designs should be treated as the equivalent of RCTs for the purposes of this item.

C. For interventions that do not lend themselves to individual or group-level assignment to conditions (e.g., area studies of crime), RCT refers to random assignment of the appropriate unit, e.g., random assignment of areas or neighborhoods to the treatment or comparison conditions.

2 = *Medium Internal Validity.*

For this rating, none of the criteria above for a high rating must apply and at least one of the following criteria must be met. If the mean effect size at issue is from a subset of the studies in the meta-analysis, there also must be no evidence at the aggregate level of significant differences between RCT and non-RCT studies, or mean RCT effect sizes and mean non-RCT effect sizes that fall outside the confidence levels above.

At least 30% (but not as many as 60%) of the studies included in the mean effect size are RCTs.

Note: If you cannot determine the proportion of RCTs and non-RCTs, the internal validity is automatically coded as low.

1 = *Low Internal Validity.*

For this rating, none of the criteria above for high or medium ratings must apply and at least one of the following criteria must be met:

Check	Points	Description
		Low Internal Validity (cont'd). (a) Fewer than 30% of the studies included in the mean effect size are RCTs. (b) Both RCTs and non-RCTs are included and the exact proportion of each cannot be determined. (c) The mean effect size at issue is from a subset of the studies in the meta-analysis and combines RCTs and non-RCTs. At the aggregate level (the full set of studies from which this subset was drawn) mean effect sizes for RCTs vs. non-RCTs were shown to be significantly different or did not fall within the confidence limits for the combined effect size identified above.

If it is necessary to compute the weighted average effect size for the RCT and non-RCT means, use the following calculation:

$$\overline{ES}_{RCT\ \&\ non-RCT} = \frac{(ES_{RCT} * n_{RCT}) + (ES_{non-RCT} * n_{non-RCT})}{(n_{RCT} + n_{non-RCT})}$$

Where ES_{RCT} is the mean effect size for the RCT studies, $ES_{non-RCT}$ is the mean effect size for the non-RCT studies, n_{RCT} is the number of studies used to calculate the RCT mean effect size, and $n_{non-RCT}$ is the number of studies included in the non-RCT mean effect size.

Part IV. Final Evidence Ratings

Step 1. Selecting the Best Available Evidence (Summative Scores)

For *each* mean effect size coded above, fill in the grid below.

1. Record the overall quality rating for the meta-analysis from which the effect size was extracted from Part II (Step 2) in the first row of the grid. In the far right column, record the value for the row (1, 2, or 3 points).
2. Record the internal validity rating for the effect size in the second row of the grid. In the far right column, record the value for the row (1, 2, or 3 points).
3. Sum the two scores from the grid to compute to Summative Score for the mean effect size.

Summative Score Calculation

Rating	Low	Medium	High	Total
Overall Quality Rating (Part II, Step 2)	☐ 1 point	☐ 2 points	☐ 3 points	
Internal Validity Rating (Part III, Step 2)	☐ 1 point	☐ 2 points	☐ 3 points	
			Summative Score	

Select the Mean Effect Sizes to Carry Forward

1. Effect sizes with scores of 2 or 3 should not be retained for further evidence rating; all other mean effect sizes should be retained for final evidence ratings.

Step 2. Statistical Significance of the Best Available Evidence

For those mean effect sizes selected as sources of evidence in the previous step: rate the outcome findings using information from the summative scores calculated in the previous step, along with information about the direction and statistical significance of the mean effect size.

Check	Class	Description
	1	*Strong Evidence of a Positive Effect* • Statistically significant mean effect size favoring the intervention • Summative Score = 6
	2	*Moderate Evidence of a Positive Effect* • Statistically significant mean effect size favoring the intervention • Summative Score = 4 or 5
	3	*Negative Effect* • Statistically significant mean effect size favoring the comparison condition • Summative Score = 4, 5, or 6
	4	*Non-significant or Null Effect* • The mean effect size is not statistically significant • Summative Score = 4, 5, or 6
	5	*Insufficient Evidence* • The mean effect size is statistically significant favoring the intervention or not statistically significant • Summative Score = 2 or 3

Step 3. Final Ratings Based on Best Available Evidence

 A single icon is used to identify programs that have been evaluated with only one meta-analysis.

 A multiple studies icon is used to represent a greater extent of evidence supporting the evidence rating. The icon depicts programs that have more than one meta-analysis in the evidence base demonstrating effects in a consistent direction.

Evidence for Programs or Practices Based on a Single Meta-analysis (Single Icon)

For each mean effect size selected as a source of evidence from this one meta-analysis in the previous steps, classify the findings **for the corresponding macro-level (Tier 1) outcome construct** using the decision rules for the Evidence Rating below. Multiple micro-level (Tier 2) mean effect sizes judged to represent the same macro-level (Tier 1) outcome domain are treated as supporting their single common outcome construct.

 *Effective*

- Effect size(s) in Class 1

 *Promising*

- Effect size(s) in Class 2

 *No Effects*

- Effect size(s) in Class 3 or 4

Insufficient Evidence

- Effect size(s) in Class 5—does not receive an icon; results only reviewed in the text write-up

Remember! Each macro-level (Tier 1) outcome from the meta-analysis must be rated separately:

Outcome 1 Rating: _____
Outcome 2 Rating: _____
Outcome 3 Rating: _____
…
Outcome *n* Rating: _____

Evidence for Programs or Practices Based on Multiple Meta-analyses (Stacked Icon)

When multiple meta-analyses provide mean effect sizes for the Evidence Ratings for a given program or practice category, all of the effect sizes from those meta-analyses carried forward from the Summative Score ratings will enter into the final Evidence Ratings. Recall that multiple mean effect sizes in the same macro-level (Tier 1) outcome domain are all used to represent the same common outcome construct. Classify the outcome findings **for each macro-level (Tier 1) outcome construct from each meta-analysis within a program or practice category separately**. In some cases, multiple meta-analyses provide information about the same outcome. An additional step is needed to arrive at a final evidence rating for that outcome: outcome findings are combined from multiple meta-analyses using the process below. CrimeSolutions.gov does not limit the number of eligible meta-analyses that may be included.

Outcome 1	Class 1	Class 2	Class 3	Class 4	Class 5
Effect Size 1	☐	☐	☐	☐	☐
Effect Size 2	☐	☐	☐	☐	☐
Effect Size 3	☐	☐	☐	☐	☐
Effect Size 4	☐	☐	☐	☐	☐

Outcome 2	Class 1	Class 2	Class 3	Class 4	Class 5
Effect Size 1	☐	☐	☐	☐	☐
Effect Size 2	☐	☐	☐	☐	☐
Effect Size 3	☐	☐	☐	☐	☐
Effect Size 4	☐	☐	☐	☐	☐

Each class of outcomes is weighted as follows:

- Class 1 = 3
- Class 2 = 1
- Class 3 = -3
- Class 4 = 0
- Class 5 is not counted

Ratings for the same outcome from multiple meta-analyses are summed and averaged. The final evidence rating for the outcome is assigned based on this score.

 *Effective*

- The averaged points ≥ 1.50

 *Promising*

- The averaged points ≥ 0.50 and ≤ 1.49

 *No Effects*

- The averaged points ≤ 0.49

Insufficient Evidence (no icon is assigned, results provided in text write-up only)

- All effect sizes are in Class 5
- If the combination of effect sizes doesn't fit into Effective, Promising, or No Effects as outlined above, then the evidence rating defaults here.