Appendix 2. Clearinghouse ratings systems

The California Evidence-Based Clearinghouse (CEBC) for Child Welfare

Website: http://www.cebc4cw.org/ratings/scientific-rating-scale/

CEBC uses a scientific rating scale with ratings from 1 to 5 to indicate the strength of the research evidence supporting a practice or program. A rating of 1 represents a practice with the strongest research evidence, and a rating of 5 represents a concerning practice that appears to pose substantial risk to children and families. Some programs do not currently have strong enough research evidence to be rated on the CEBC's scientific rating scale and are classified as NR - (Not able to be Rated).

Specific criteria for each rating are presented below:

Well Supported by Research Evidence

a. There is no case data suggesting a risk of harm that: i) was probably caused by the treatment; and ii) the harm was severe or frequent.

b. There is no legal or empirical basis suggesting that compared to its likely benefits, the practice constitutes a risk of harm to those receiving it.

c. The practice has a book, manual, and/or other available writings that specify components of the service and describes how to administer it.

d. Multiple Site Replication: At least two rigorous RCTs in different usual care or practice settings have found the practice to be superior to an appropriate comparison practice. The RCTs have been reported in published, peer-reviewed literature.

e. In at least one RCT, the practice was shown to have a sustained effect at least one year beyond the end of treatment.

f. Outcome measures must be reliable and valid, and administered consistently and accurately across all subjects.

g. If multiple outcome studies have been published, the overall weight of the evidence supports the benefit of the practice.

Supported by Research Evidence

a. There is no case data suggesting a risk of harm that: i) was probably caused by the treatment; and ii) the harm was severe or frequent.

b. There is no legal or empirical basis suggesting that compared to its likely benefits, the practice constitutes a risk of harm to those receiving it.

c. The practice has a book, manual, and/or other available writings that specifies the components of the practice protocol and describes how to administer it.

d. At least one rigorous RCT in usual care or a practice setting has found the practice to be

superior to an appropriate comparison practice. The RCT has been reported in published, peer-reviewed literature.

e. In at least one RCT, the practice was shown to have a sustained effect of at least six months beyond the end of treatment.

f. Outcome measures must be reliable and valid, and administered consistently and accurately across all subjects.

g. If multiple outcome studies have been published, the overall weight of evidence supports the benefit of the practice.

Promising Research Evidence

a. There is no case data suggesting a risk of harm that: a) was probably caused by the treatment; and b) the harm was severe or frequent.

b. There is no legal or empirical basis suggesting that compared to its likely benefits, the practice constitutes a risk of harm to those receiving it.

c. The practice has a book, manual, and/or other available writings that specifies the components of the practice protocol and describe how to administer it.

d. At least one study utilising some form of control (e.g., untreated group, placebo group, matched wait list study) has established the practice's benefit over the control, or found it to be comparable to a practice rated a 1, 2, or 3 on this rating scale or superior to an appropriate comparison practice. The study has been reported in published, peer-reviewed literature.

e. If multiple outcome studies have been conducted, the overall weight of evidence supports the benefit of the practice.

Evidence Fails to Demonstrate Effect

a. Two or more RCTs have found the practice has not resulted in improved outcomes, when compared to usual care. The studies have been reported in published, peer-reviewed literature.

b. If multiple outcome studies have been conducted, the overall weight of evidence does not support the benefit of the practice. The overall weight of evidence is based on the preponderance of published, peer-reviewed studies, and not a systematic review or meta-analysis. For example, if there have been three published RCTs and two of them showed the program did not have the desired effect, then the program would be rated a "4 - Evidence Fails to Demonstrate Effect."

Concerning Practice

a. If multiple outcome studies have been conducted, the overall weight of evidence suggests the intervention has a negative effect upon clients served; and/or

b. There is case data suggesting a risk of harm that: i) was probably caused by the

The California Evidence-Based Clearinghouse (CEBC) for Child Welfare

treatment; and ii) the harm was severe or frequent; and/or

c. There is a legal or empirical basis suggesting that compared to its likely benefits, the practice constitutes a risk of harm to those receiving it.

NR. Not able to be Rated

a. There is no case data suggesting a risk of harm that: i) was probably caused by the treatment; and ii) the harm was severe or frequent.

b. There is no legal or empirical basis suggesting that compared to its likely benefits, the practice constitutes a risk of harm to those receiving it.

c. The practice has a book, manual, and/or other available writings that specifies the components of the practice protocol and describes how to administer it.

d. The practice is generally accepted in clinical practice as appropriate for use with children receiving services from child welfare or related systems and their parents/caregivers.

e. The practice does not have any published, peer-reviewed study utilising some form of control (e.g., untreated group, placebo group, matched wait list study) that has established the practice's benefit over the placebo, or found it to be comparable to or better than an appropriate comparison practice.

f. The practice does not meet criteria for any other level on the CEBC Scientific Rating Scale.

National Resource Center for Community-Based Child Abuse Prevention (CBCAP)

Website: http://friendsnrc.org/

Programs are rated according to the following criteria:

Emerging Programs and Practices

Programmatic Characteristics

a. The program can articulate a theory of change which specifies clearly identified outcomes and describes the activities that are related to those outcomes. This may be represented through a program logic model or conceptual framework that depicts the assumptions for the activities that will lead to the desired outcomes.

b. The program may have a book, manual, other available writings, training materials, OR may be working on documents that specifies the components of the practice protocol and describes how to administer it.

c. The practice is generally accepted in clinical practice as appropriate for use with children and their parents/caregivers receiving child abuse prevention or family support services.

Research & Evaluation Characteristics

a. There is no clinical or empirical evidence or theoretical basis indicating that the practice

constitutes a substantial risk of harm to those receiving it, compared to its likely benefits.

b. Programs and practices may have been evaluated using less rigorous evaluation designs that have no comparison group. This includes using "pre-post" designs that examine change in individuals from before the program or practice was implemented to afterward, without comparing to an "untreated" group. OR - an evaluation may be in process with the results not yet available.

c. The program is committed to and is actively working on building stronger evidence through ongoing evaluation and continuous quality improvement activities. For additional information on evaluation and developing logic models, visit the FRIENDS Evaluation Toolkit and Logic Model Builder at: http://www.friendsnrc.org/outcome/toolkit/index.htm

Promising Programs and Practices

Programmatic Characteristics

a. The program can articulate a theory of change which specifies clearly identified outcomes and describes the activities that are related to those outcomes. This is represented through presence of a program logic model or conceptual framework that depicts the assumptions for the activities that will lead to the desired outcomes.

a. The program may have a book, manual, other available writings, and training materials that specifies the components of the practice protocol and describes how to administer it. The program is able to provide formal or informal support and guidance regarding program model.

b. The practice is generally accepted in clinical practice as appropriate for use with children and their parents/caregivers receiving services child abuse prevention or family support services.

Research & Evaluation Characteristics

a. There is no clinical or empirical evidence or theoretical basis indicating that the practice constitutes a substantial risk of harm to those receiving it, compared to its likely benefits.

c. At least one study utilizing some form of control or comparison group (e.g., untreated group, placebo group, matched wait list) has established the practice's efficacy over the placebo, or found it to be comparable to or better than an appropriate comparison practice, in reducing risk and increasing protective factors associated with the prevention of abuse or neglect. The evaluation utilised a quasi-experimental study design, involving the comparison of two or more groups that differ based on their receipt of the program or practice. A formal, independent report has been produced which documents the program's positive outcomes.

d. The local program is committed to and is actively working on building stronger evidence through ongoing evaluation and continuous quality improvement activities. Programs continually examine long-term outcomes and participate in research that would help

solidify the outcome findings.

e. The local program can demonstrate adherence to model fidelity in program or practice implementation.

Supported Programs and Practices

Programmatic Characteristics

a. The program articulates a theory of change which specifies clearly identified outcomes and describes the activities that are related to those outcomes. This is represented through the presence of a detailed logic model or conceptual framework that depicts the assumptions for the inputs and outputs that lead to the short, intermediate and long-term outcomes.

b. The practice has a book, manual, training, or other available writings that specifies the components of the practice protocol and describes how to administer it.

c. The practice is generally accepted in clinical practice as appropriate for use with children and their parents/caregivers receiving child abuse prevention or family support services.

Research & Evaluation Characteristics

a. There is no clinical or empirical evidence or theoretical basis indicating that the practice constitutes a substantial risk of harm to those receiving it, compared to its likely benefits.

b. The research supporting the efficacy of the program or practice in producing positive outcomes associated with reducing risk and increasing protective factors associated with the prevention of abuse or neglect meets at least one or more of the following criterion:

- At least two rigorous RCTs (or other comparable methodology) in highly controlled settings (e.g., university laboratory) have found the practice to be superior to an appropriate comparison practice. The RCTs have been reported in published, peer-reviewed literature.

OR

- At least two between-group design studies using either a matched comparison or regression discontinuity have found the practice to be equivalent to another practice that would qualify as supported or well-supported; or superior to an appropriate comparison practice.

c. The practice has been shown to have a sustained effect at least one year beyond the end of treatment, with no evidence that the effect is lost after this time.

d. Outcome measures must be reliable and valid, and administered consistently and accurately across all subjects.

e. If multiple outcome studies have been conducted, the overall weight of evidence supports the efficacy of the practice.

f. The program is committed and is actively working on building stronger evidence through ongoing evaluation and continuous quality improvement activities.

g. The local program can demonstrate adherence to model fidelity in program implementation.

Well Supported Programs and Practices

Programmatic Characteristics

a. The program articulates a theory of change which specifies clearly identified outcomes and describes the activities that are related to those outcomes. This is represented through the presence of a detailed logic model or conceptual framework that depicts the assumptions for the inputs and outputs that lead to the short, intermediate and long-term outcomes.

b. The practice has a book, manual, training or other available writings that specify components of the service and describes how to administer it.

c. The practice is generally accepted in clinical practice as appropriate for use with children and their parents/caregivers receiving child abuse prevention or family support services.

Research & Evaluation Characteristics

a. Multiple Site Replication in Usual Practice Settings: At least two rigorous RCTs or comparable methodology in different usual care or practice settings have found the practice to be superior to an appropriate comparison practice. The RCTs have been reported in published, peer-reviewed literature.

b. There is no clinical or empirical evidence or theoretical basis indicating that the practice constitutes a substantial risk of harm to those receiving it, compared to its likely benefits.

c. The practice has been shown to have a sustained effect at least one year beyond the end of treatment, with no evidence that the effect is lost after this time.

d. Outcome measures must be reliable and valid, and administered consistently and accurately cross all subjects.

e. If multiple outcome studies have been conducted, the overall weight of the evidence supports the effectiveness of the practice.

f. The program is committed and is actively working on building stronger evidence through ongoing evaluation and continuous quality improvement activities.

g. The local program can demonstrate adherence to model fidelity in program implementation.*

Programs and Practices Lacking Support or Positive Outcomes/ Undetermined/ Concerning/Harmful Effects

Programmatic Characteristics

National Resource Center for Community-Based Child Abuse Prevention (CBCAP)

a. The program is not able to articulate a theory of change which specifies clearly identified outcomes and describes the activities that are related to those outcomes.

b. The program does not have a book, manual, other available writings, training materials that describe the components of the program.

Research & Evaluation Characteristics

a. Two or more RCTs have found the practice has not resulted in improved outcomes, or has had harmful effects when compared to usual care.

OR

b. If multiple outcome studies have been conducted, the overall weight of evidence does NOT support the efficacy of the practice.

OR

c. No evaluation has been conducted. The program may or may not have plans to implement an evaluation.

Social Programs that Work (SPW) (Coalition for Evidence-Based Policy)

Website: http://www.evidencebasedprograms.org/

Description of Rating System

The Coalition for Evidence Based Policy use a "Top Tier Evidence" system to identify and validate interventions for inclusion in their Social Programs that Work clearinghouse. For each viable program, their search the literature and contact experts to identify all well-conducted randomised trials of the intervention (in addition to those initially brought to their attention). An Advisory Panel of nationally-recognized, evidence-based researchers and former public officials, decides which interventions to identify as Top Tier or Near Top Tier.

Top Tier

The standard used to evaluate candidates for the Top Tier, based on the Congressional legislative language, is: "Interventions shown in well-conducted randomised controlled trials, preferably conducted in typical community settings, to produce sizeable, sustained benefits to participants and/or society."

In applying this standard, the Checklist For Reviewing a Randomized Controlled Trial is used, which closely tracks guidance from the U.S. Office of Management and Budget (OMB), National Academies, and other respected research organisations, and reflects well-established principles on what constitutes a high-quality trial (e.g., adequate sample size, low sample attrition, valid outcome measures, intention to treat analysis). It also addresses the importance of replication in establishing strong evidence – namely, demonstration of effectiveness in at least two well-conducted trials, or one large multi-site trial.

The main focus for each candidate intervention is on assessing whether there is strong evidence that the intervention's effects are sizeable and sustained. However, in some cases,

Social Programs that Work (SPW) (Coalition for Evidence-Based Policy)

reviewers might also take into account such factors as the intervention's cost and ease of implementation (e.g., cases where the cost is exceptionally low).

Over time, short case summaries are developed illustrating the reasoning used in applying the above standard and guidance to particular studies, thus building a body of additional guidance for reviewers and applicants that is grounded in case-by-case decisions. (This approach – using actual case decisions to grow the body of guidance over time – has been long used by the Food and Drug Administration in its well-established procedures for reviewing randomised controlled trials of pharmaceutical drugs.)

Near Top Tier

The standard used to evaluate candidates for Near Top Tier is: Interventions shown to meet all elements of the Top Tier standard in a single site, and which only need one additional step to qualify as Top Tier – a replication trial to confirm the initial findings and establishing that they generalise to other sites.

The purpose of this category is to help grow the body of Top Tier interventions, by enabling policymakers and others to identify particularly strong candidates for replication trials from among the many interventions backed by more preliminary evidence, and thereby maximise the chances of a positive replication that would qualify the intervention as Top Tier.

For each viable program, the literature is searched and experts are contacted to identify all other high quality randomised trials of the intervention (in addition to those initially brought to the attention of the reviewers). Also, for interventions being considered for Top Tier or Near Top Tier on the basis of a limited number of well-designed and implemented randomised trials, the literature of high-quality non-randomised studies of the intervention is checked, to look for any patterns of effects that differ from those in the trials (possibly suggesting problems in generalisability) or for any adverse intervention effects.

Blueprints

Website: http://www.colorado.edu/cspv/blueprints/criteria.html

The selection criteria used by Blueprints reflect the level of confidence necessary for recommending that communities use programs with reasonable assurances that they will prevent violence and other behavioural problems when implemented with fidelity. Blueprints Model Programs are not intended to be a comprehensive list of programs that work, but rather reflect a selection of programs with strong research designs for which there is good evidence of their effectiveness. There is no implication that programs not on this list are necessarily ineffective. Chances are that there are a number of good programs that have just not yet undergone the rigorous evaluations required to demonstrate effectiveness.

Selection Criteria

There are several important criteria considered by Blueprints when reviewing program effectiveness. Three of these criteria are given greater weight: evidence of deterrent effect with a strong research design, sustained effect, and multiple site replication. Blueprints Model Programs must meet all three of these criteria, while Promising Programs must meet at least the first criterion.

Blueprints

Evidence of deterrent effect with a strong research design

This is the most important of the selection criteria.

Providing sufficient quantitative data to document effectiveness in preventing or reducing targeted behaviours requires the use of evaluative designs that provide reasonable confidence in the findings (e.g., experimental designs with random assignment or quasi-experimental designs with matched control groups). When random assignment cannot be used, the Blueprints Advisory Board considers studies that use control groups matched as closely as possible to experimental groups on relevant characteristics (e.g., gender, race, age, socioeconomic status, income) and studies with control groups that use statistical techniques to control for initial differences on key variables. As carefully as experimental and control groups are matched, however, it is impossible to determine if the groups may vary on some characteristics that have not been matched or controlled for and that are related to program outcome. Random assignment, therefore, is believed to be the most rigorous of methodological approaches.

At a minimum, the following issues need to be addressed:

1. Sample sizes must be large enough to provide statistical power to detect at least moderate sized effects. Selection of participants must be made in a manner that avoids bias. For example, a self-selecting sample that relies on volunteer participants might be more motivated to make change, thus introducing a plausible alternative explanation for outcomes that are achieved. An adequate description should report the characteristics of the sample, the selection process, and pre-test differences on relevant variables between the treatment and control conditions.

2. Sample sizes and losses must be reported through all follow-up periods, and tests that rule out differential attrition should be conducted.

3. Tests to measure outcomes must be administered fairly, accurately and consistently to all study participants. The instruments used to measure outcomes should be demonstrated to be reliable and valid. Measurements of actual behaviour are required for Blueprints, not attitudes or intent. More than one report of behaviour is preferable in instances where the same person both delivers the intervention and provides a measure of the outcome. When multiple measures of outcomes are used in a study, the intervention should significantly influence the most important outcomes and influence the others in the expected direction.

4. Analyses should be appropriately designed. They should be done at the same level as the randomisation and, following an "intent to treat" approach, should include all participants originally assigned to treatment and control conditions. Secondary analyses can be performed to determine the effectiveness of a program at differing levels of implementation and dosage. Two-tailed tests of significance are preferred since they represent the most conservative of tests.

School-based evaluations

Evaluations of school-based programs, with schools as the unit of analysis, typically require multiple schools per condition to perform a main effects analysis with sufficient power to detect effects. Since meeting this criterion requires a complex and costly evaluation, it would eliminate most existing school-level studies from consideration in the Blueprints Series.

Therefore, school-based evaluations that use experimental or quasi-experimental designs with relatively few schools, but more than one in each condition, are considered in the Blueprints Series if they meet an additional burden of proof. They must demonstrate consistency across effects and across replications with multiple measures from different sources. The theoretical rationale should be well developed, and there should be a rigorous evaluation of theory with evidence that the results are consistently in line with the expectations (i.e., there are changes in the risk and protective factors which mediate the changes in outcomes). Outcomes should be robust, with at least moderate effect sizes. Evidence that the benefits of the program outweigh the costs is helpful. Evaluations with multiple schools are most desirable and should be encouraged among funders and researchers.

Sustained effect

Designation as a Blueprints Model Program requires a sustained effect at least one year beyond treatment, with no subsequent evidence that this effect is lost.

A program may be identified as promising without meeting the sustainability criterion. In some cases, programs may not have conducted longer-term follow-ups. In other cases, programs will have performed long-term follow-ups and found no enduring effects. If program effects disappear at a later time period, Blueprints may qualify the program for only the period of time in which it was found to be effective, stating the loss of enduring effects at the point at which they were found. While these programs may not show enduring effects for 12 months or longer on specifically measured outcomes, in some cases they can provide meaningful benefits to youth, schools, and communities. For example, even if benefits don't last, delaying the onset of alcohol and drug use to a later age would improve the safety of youth during a highly vulnerable period of their lives. And since early onset of youth problems often leads to more serious problems later, delaying onset with temporary improvements may have payoffs at older ages

Multiple site replications

Becoming a Blueprints Model Program requires at least one high-quality replication with fidelity demonstrating that the program continues to be effective. This criterion does not need to be met to qualify as a promising program.

Some projects may be initially implemented as a multisite single design (i.e., several sites are included in the evaluation design). Although not as valuable as independent replications, these designs can check for overall main effects and sources of variation across sites.

Replication dismantling designs will also be considered. If a program has been implemented and evaluated as a component within a number of different programs (multiple component studies) and has also been implemented and evaluated alone, it is possible that the multiple component studies might meet the replication criterion. There must be a total of three studies, including the standalone program evaluation and two additional multiple component studies. All must be well designed with positive effects and with no negative effects.

Additional Factors

In the selection of Blueprints Model Programs, two additional factors are considered: whether a program conducted an analysis of mediating factors and whether a program is cost effective.

Analysis of mediating factors

The Blueprints Advisory Board looks for evidence that change in the targeted risk or protective

Blueprints

factor(s) mediates the change in problem behaviours. This evidence clearly strengthens the claim that participation in the program is responsible for the change in behaviour, and it contributes to the theoretical understanding of the causal processes involved.

Costs versus benefits

Program costs should be reasonable and should be less or no greater than the program's expected benefits.

Strengthening America's Families (SAF): Effective Family Programs for Prevention of Delinquency

Website: http://www.strengtheningfamilies.org/

Description of Rating System

Numerous criteria were used to rate and categorise programs. The criteria included: theory, fidelity of the interventions, sampling strategy and implementation, attrition, measures, data collection, missing data, analysis, replications, dissemination capability, cultural and age appropriateness, integrity and program utility.

Each program was rated independently by reviewers, discussed and a final determination made regarding the appropriate category. The following categories were used:

Exemplary I

This indicates the program has evaluation of the highest quality with an experimental design with a randomised sample and replication by an independent investigator other than the program developer. Outcome data from the numerous research studies show clear evidence of program effectiveness.

Exemplary II

This indicates the program has evaluation of the highest quality with an experimental design with a randomised sample. Outcome data from the numerous research studies show clear evidence of program effectiveness.

Model

This indicates the program has research of either an experimental or quasi-experimental design with few or no replications. Outcome data from the research project(s) indicate program effectiveness but the data are not as strong in demonstrating program effectiveness.

Promising

This indicates the program has limited research and/or employs non-experimental designs. Evaluation data associated with the program appears promising but requires confirmation using scientific techniques. The theoretical base and/or some other aspect of the program is also sound.

Programs rated as Exemplary programs are those that are well-implemented, are rigorously evaluated, and have consistent positive findings (integrity ratings of "A4 "or "A5 "). Model programs are those that have consistent integrity ratings of "A3" and "A4" and Promising programs are those that have mixed integrity ratings but demonstrate high integrity ratings in

at least 3-4 of the following categories.

Theory: the degree to which the project findings are based in clear and well-articulated theory, clearly stated hypotheses, and clear operational relevance.

1 = no information about theory or hypotheses specified

2 = very little information about theory and hypotheses specified

3 = adequate information about theory and hypotheses specified

4 = nearly complete information about theory and hypotheses specified

5 = full and complete information about theory and hypotheses specified

Fidelity of interventions: the degree to which there is clear evidence of high fidelity implementation, which may include dosage data.

1 = no or very weak evidence that most treatment participants received the full intervention

2 = weak evidence that most treatment participants received the full intervention

3 = some evidence that most treatment participants received the full intervention

4 = strong evidence that most treatment participants received the full intervention

5 = very strong evidence that nearly all treatment participants received the full intervention

Sampling strategy and implementation: the quality of sampling design and implementation.

1 = no control group; unspecified sample size or inadequate sample size

2 = inappropriate control group included or no attempt at random assignment; inadequate sample size

3 = inappropriate control group included or no attempt at random assignment; adequate sample size

4 = control group included; random assignment at individual or other level (e.g., school); adequate sample size

5 = control group included; random assignment at individual or other level (e.g., school); more than adequate sample size

Attrition: evidence of sample quality based on information about attrition.

1 = no data on attrition or very high attrition

2 = high attrition

3 = moderate attrition

4 = acceptable retention

5 = high retention

Measures: the operational relevance and psychometric quality of measures used in the evaluation, and the quality of supporting evidence.

1 = no or insufficient information about measures

2 = poor choice of measures; low psychometric qualities

3 = adequate choice of measures; mixed quality

4 = relevant measures with good psychometric qualities

5 = highly relevant measures with excellent psychometric qualities

Missing data: the quality of implementation of data collection (e.g., amount of missing data).

1 = high quantity of missing data

2 = somewhat high quantity of missing data

3 = average amount of missing data

4 = some missing data

5 = no or almost no missing data

Data collection: way data collected in terms of bias or demand characteristics and haphazard manner.

1 = very biased manner of data collection with high demand characteristics; data collected in haphazard manner without any standardization

2 = somewhat biased manner of data collection with some demand characteristics; data collected in haphazard manner without any standardization

3 = relatively unbiased manner of data collection; standardized method of data collection

4 = anonymous or confidentiality ensured in data collection; standardized method of data collection

5 = anonymous or confidentiality ensured in data collection; standardized method of data collection; ethnic group or gender match between data collectors and participants specified

Analysis: the appropriateness and technical adequacy of techniques of analysis, primarily statistical.

1 = no analyses reported; all analyses inappropriate or do not account for important factors

2 = some but not all analyses inappropriate or left out important factors

3 = mixed in terms of appropriateness and technical adequacy

4 = appropriate analyses but not cutting edge techniques

5 = proper, state-of-the-art analyses conducted

Other plausible threats to validity (excluding attrition): the degree to which the evaluation design and implementation addresses and eliminates plausible alternative hypotheses concerning program effects. The degree to which the study design and implementation warrants strong causal attributions concerning program effects.

1 = high threat to validity or no ability to attribute program effects

2 = threat to validity and difficult to attribute program effects

3 = somewhat of threat to validity and mixed ability to attribute effects to the program

4 = low threat to validity and ability to attribute effects to the program

5 = no or very low threat to validity and high ability to attribute effects to the program

Replications: the exact or conceptual reproduction of both the intervention implementation and evaluation.

1 = no replication.

2 = one self-replication.

3 = two or more self-replications.

4 = one or two replications by independent evaluators.

5 = three or more replications by independent evaluators producing similar results.

Dissemination capability: program materials developed including training in program implementation, technical assistance, standardized curriculum and evaluation materials, manuals, fidelity instrumentation, videos, recruitment forms, etc.

1 = Materials, training and technical assistance not available; in case of model that requires no curriculum (i.e., therapeutic models), training/qualified trainers and technical assistance not available.

2 = Materials available but of low quality or very limited in scope; training/qualified trainers and technical assistance either not available or limited.

3 = Materials of sufficient quality with limited technical assistance and/or training/qualified trainers.

4 = High quality materials, limited technical assistance and/or training/qualified trainers or vise versa.

5 = High quality materials, technical assistance readily available and training/qualified trainers readily available.

Cultural and age appropriateness

1 = no claim of culturally or age appropriate materials targeted for specific populations.

2 = claim of cultural or age appropriate materials but no of validation.

3 = age specific but not culturally appropriate or vice versa with some face validation.

4 = some materials validation materials presented.

5 = specialised materials, culturally and age appropriate, developed and evaluated or existing validated materials targeting population used.

Integrity: the overall level of confidence that the reviewer can place in project findings based on research design and implementation.

1 = no confidence

2 = weak, at best some confidence in results

3 = mixed, some weak, some strong characteristics

4 = strong, fairly good confidence in results

Strengthening America's Families (SAF): Effective Family Programs for Prevention of Delinquency

5 = high confidence in results, findings fully defensible

Utility: the overall usefulness of project findings for informing prevention theory and practice. This rating is anchored according to the following categories, and combines the strength of findings and the strength of evaluation.

1 = The evaluation produced clear findings of null or negative effects for a program with well-articulated theory and program design, the study provides support for rejecting the program as a replication model.

2 = The evaluation produced findings that were predominately null or negative, though not uniform or definitive.

3 = The evaluation produced ambiguous findings because of inconsistency in result or methods weaknesses that do not provide a strong basis for programmatic or theoretical contributions.

4 = The evaluation produced positive findings that demonstrate the efficacy of the program in some areas, or support the efficacy of some components of the program.

5 = The evaluation produced clear findings supporting the efficacy of well-articulated theory and program design, the study provides support for the program as a replication model

Office of Juvenile Justice and Delinquency Prevention (OJJDP)

Website: http://www.ojjdp.gov/mpg/ratings.aspx

The evidence ratings used by the OJJDP are based on the evaluation literature of specific prevention and intervention programs. The overall rating is derived from four summary dimensions of program effectiveness:

- The conceptual framework of the program
- The program fidelity
- The evaluation design
- The empirical evidence demonstrating the prevention or reduction of problem behaviour; the reduction of risk factors related to problem behaviour; or the enhancement of protective factors related to problem behaviour.

Programs are classified into three categories that are designed to provide the user with a summary knowledge base of the research supporting a particular program. A brief description of the rating criteria is provided below.

Exemplary

In general, when implemented with a high degree of fidelity these programs demonstrate robust empirical findings using a reputable conceptual framework and an evaluation design of the highest quality (experimental).

Effective

In general, when implemented with sufficient fidelity these programs demonstrate adequate

Office of Juvenile Justice and Delinquency Prevention (OJJDP)

empirical findings using a sound conceptual framework and an evaluation design of the high quality (quasi-experimental).

Promising

In general, when implemented with minimal fidelity these programs demonstrate promising (perhaps inconsistent) empirical findings using a reasonable conceptual framework and a limited evaluation design that requires causal confirmation using more appropriate experimental techniques.

SAMHSA's National Registry of Evidence-based Programs and Practices

Website: http://nrepp.samhsa.gov/ReviewQOR.aspx

Quality of Research

SAMHSA's National Registry of Evidence-based Programs and Practices Quality of Research ratings are indicators of the strength of the evidence supporting the outcomes of the intervention. Higher scores indicate stronger, more compelling evidence. Each outcome is rated separately because interventions may target multiple outcomes (e.g., alcohol use, marijuana use, behaviour problems in school), and the evidence supporting the different outcomes may vary.

SAMHSA uses specific standardised criteria to rate interventions and the evidence supporting their outcomes. All reviewers who conduct reviews are trained on these criteria and are required to use them to calculate their ratings.

Criteria for Rating Quality of Research

Each reviewer independently evaluates the Quality of Research for an intervention's reported results using the following six criteria:

- Reliability of measures
- Validity of measures
- Intervention fidelity
- Missing data and attrition
- Potential confounding variables
- Appropriateness of analysis

For each outcome, reviewers use a scale of 0.0 to 4.0, with 4.0 being the highest rating given, to rate each criterion listed above. Then a mean score is calculated, and reported as an overall rating for each outcome. It is this overall rating that is reported in the current review of parenting programs.

A more detailed description of rating criteria is provided below.

1. Reliability of Measures: Outcome measures should have acceptable reliability to be interpretable. "Acceptable" here means reliability at a level that is conventionally accepted by experts in the field.

0 = Absence of evidence of reliability or evidence that some relevant types of reliability

(e.g., test-retest, inter-rater, inter-item) did not reach acceptable levels.

2 = All relevant types of reliability have been documented to be at acceptable levels in studies by the applicant.

4 = All relevant types of reliability have been documented to be at acceptable levels in studies by independent investigators.

2. Validity of Measures: Outcome measures should have acceptable validity to be interpretable. "Acceptable" here means validity at a level that is conventionally accepted by experts in the field.

0 = Absence of evidence of measure validity, or some evidence that the measure is not valid.

2 = Measure has face validity; absence of evidence that measure is not valid.

4 = Measure has one or more acceptable forms of criterion-related validity (correlation with appropriate, validated measures or objective criteria); OR, for objective measures of response, there are procedural checks to confirm data validity; absence of evidence that measure is not valid.

3. Intervention Fidelity:- The "experimental" intervention implemented in a study should have fidelity to the intervention proposed by the applicant. Instruments that have tested acceptable psychometric properties (e.g., inter-rater reliability, validity as shown by positive association with outcomes) provide the highest level of evidence.

0 = Absence of evidence or only narrative evidence that the applicant or provider believes the intervention was implemented with acceptable fidelity.

2 = There is evidence of acceptable fidelity in the form of judgment(s) by experts, systematic collection of data (e.g., dosage, time spent in training, adherence to guidelines or a manual), or a fidelity measure with unspecified or unknown psychometric properties.

4 = There is evidence of acceptable fidelity from a tested fidelity instrument shown to have reliability and validity.

4. Missing Data and Attrition: Study results can be biased by participant attrition and other forms of missing data. Statistical methods as supported by theory and research can be employed to control for missing data and attrition that would bias results, but studies with no attrition or missing data needing adjustment provide the strongest evidence that results are not biased.

0 = Missing data and attrition were taken into account inadequately, OR there was too much to control for bias.

2 = Missing data and attrition were taken into account by simple estimates of data and observations, or by demonstrations of similarity between remaining participants and those lost to attrition.

4 = Missing data and attrition were taken into account by more sophisticated methods that model missing data, observations, or participants, OR there were no attrition or missing data needing adjustment.

5. Potential Confounding Variables: Often variables other than the intervention may account for the reported outcomes. The degree to which confounds are accounted for affects the

SAMHSA's National Registry of Evidence-based Programs and Practices

strength of causal inference.

0 = Confounding variables or factors were as likely to account for the outcome(s) reported as were the hypothesized causes.

2 = One or more potential confounding variables or factors were not completely addressed, but the intervention appears more likely than these confounding factors to account for the outcome(s) reported.

4 = All known potential confounding variables appear to have been completely addressed in order to allow causal inference between the intervention and outcome(s) reported.

6. Appropriateness of Analysis: Appropriate analysis is necessary to make an inference that an intervention caused reported outcomes.

0 = Analyses were not appropriate for inferring relationships between intervention and outcome, OR sample size was inadequate.

2 = Some analyses may not have been appropriate for inferring relationships between intervention and outcome, OR sample size may have been inadequate.

4 = Analyses were appropriate for inferring relationships between intervention and outcome. Sample size and power were adequate.

Promising Practices Network (PPN)

Website: http://www.promisingpractices.net/criteria.asp

How programs are considered

The PPN reviews any program for which there is evidence of a positive effect. A formal application is not required to submit a program for consideration. PPN relies on publicly available information for the review of a program's effectiveness. PPN are interested in programs as they were designed and evaluated — programs do not have to have been replicated or be currently in existence for inclusion. Also, even if the specific goal of the program does not address an indicator, but the evaluation shows a positive effect, PPN will include the program under the indicator for which the evidence indicates effectiveness

Evidence Levels

Proven and Promising Programs

Programs are generally assigned either a "Proven" or a "Promising" rating, depending on whether they have met the evidence criteria below. In some cases a program may receive a Proven rating for one indicator and a Promising rating for a different indicator. In this case the evidence level assigned will be Proven/Promising, and the program summary will specify how the evidence levels were assigned by indicator.

Other Reviewed Programs

Some programs on the PPN site are identified as "Other Reviewed Programs". These are programs that have not undergone a full review by PPN, but evidence of their effectiveness has been reviewed by one or more credible organizations that apply similar evidence criteria. Other Reviewed Programs may be fully reviewed by PPN in the future and identified as Proven

or Promising, but will be identified as Other Reviewed Programs in the interim.

Evidence Criteria

Proven Program

Program must meet all of these criteria to be listed as "Proven":

a. Type of Outcomes Affected - Program must directly impact one of the indicators used on the site

b. Substantial Effect Size - At least one outcome is changed by 20%, 0.25 standard deviations, or more

c. Statistical Significance - At least one outcome with a substantial effect size is statistically significant at the 5% level

d. Comparison Groups - Study design uses a convincing comparison group to identify program impacts, including randomised-control trial (experimental design) or some quasi-experimental designs

e. Sample Size - Sample size of evaluation exceeds 30 in both the treatment and comparison groups

f. Availability of Program Evaluation Documentation - Publically available.

Promising Program

Program must meet at least all of these criteria to be listed as "Promising":

a. Type of Outcomes Affected - Program may impact an intermediary outcome for which there is evidence that it is associated with one of the PPN indicators

b. Substantial Effect Size - Change in outcome is more than 1%

c. Statistical Significance - Outcome change is significant at the 10% level (marginally significant)

d. Comparison Groups - Study has a comparison group, but it may exhibit some weaknesses, e.g., the groups lack comparability on pre-existing variables or the analysis does not employ appropriate statistical controls

e. Sample Size - Sample size of evaluation exceeds 10 in both the treatment and comparison groups

f. Availability of Program Evaluation Documentation - Publically available.

Not Listed on Site

If a program meets any of these conditions it will not be listed on the site:

a. Type of Outcomes Affected - Program impacts an outcome that is not related to children or their families, or for which there is little or no evidence that it is related to a PPN indicators (such as the number of applications for teaching positions)

b. Substantial Effect Size - No outcome is changed more than 1%

c. Statistical Significance - No outcome change is significant at less than the 10% level

d. Comparison Groups - Study does not use a convincing comparison group. For example, the use of before and after comparisons for the treatment group only

Promising Practices Network (PPN)

e. Sample Size - Sample size of evaluation includes less than 10 in the treatment or comparison group

f. Availability of Program Evaluation Documentation - Distribution is restricted, for example only to the sponsor of the evaluation.

Currently, PPN does not require programs to do the following:

- Be currently implemented in some location and provide technical assistance or support.
- Have been replicated numerous times. (While PPN recognise the importance of program replication and fidelity to program success, they believe there is value to including information about programs that have successfully improved outcomes for children and families but have not been replicated.)
- Have articulated as program goals the outcomes they impact. (For example, if a program was designed to reduce violence, but met the criteria for a proven program because it reduced drug use, PPN would list the program as a "proven" program under the drug use reduction indicator, even though the program did not intend to reduce drug use.)
- Evaluation to have appeared in a peer-reviewed journal. Nor do PPN count as "Proven" every evaluation that has been published in a peer-reviewed journal.