The Collaborative Outcome Data Committee's Guidelines for the Evaluation of Sexual Offender Treatment Outcome Research

Part 2: CODC Guidelines

2007-03

Collaborative Outcome Data Committee (in alphabetical order):

Anthony Beech, Guy Bourgon, R. Karl Hanson, Andrew J. R. Harris, Calvin Langton, Janice Marques, Michael Miner, William Murphy, Vernon Quinsey, Michael Seto, David Thornton, Pamela M. Yates

Cat. No.: PS3-1/2007-3E
ISBN No.: 978-0-662-46069-5

Table of contents

Introduction **1**

Identifying the design **3**

I. Administrative control of independent variables **4**

1. Defining treatment 6

2. Defining comparison 10

3. Miscellaneous incidental factors 13

II. Experimenter expectancies **16**

4. Experimenter investment in outcome 17

5. Blinding in data management 20

III. Sample Size **23**

6. Sample size of treatment group(s) 25

7. Sample size of comparison group(s) 28

8. Sample size of institutions: Cross-institutional designs 31

IV. Attrition **34**

9. Subject selection 37

10. Program attrition 41

11. Intent-to-treat analysis 44

12. Attrition in follow-up 47

V. Equivalence of groups **50**

13. A priori equivalence of groups 52

14. Adequacy of search for pre-existing differences 54

15. Findings on group equivalence 61

VI. Outcome variables **64**

16. Length of follow-up 66

17. Validity and reliability of recidivism information 69

18. Equivalence of follow-up 72

VII. Correct comparisons conducted 75

19. Data dredging 76

20. Effectiveness of statistical procedures to control bias 79

21. Computation of least bias comparison 83

VIII. Global rating **87**

Study quality rating guide summary sheet ……. 89

References **91**

Committee members **93**

Introduction

The following Collaborative Outcome Data Committee (CODC) Guidelines are for evaluating the quality of sexual offender treatment outcome studies. A high quality study is one where there is a high degree of confidence that the effect of treatment was estimated with minimal bias. These Guidelines should be helpful for readers and reviewers of the professional literature, as well as for researchers designing new studies or evaluating existing programs. Background information can be found in *The Collaborative Outcome Data Committee's Guidelines for the Evaluation of Sexual Offender Treatment Outcome Research (CODC Guidelines), Part 1: Introduction and Overview.*

Direction for using CODC Guidelines with existing studies

Step 1 – Identify the type of design. Using the flow chart provided, identify the method of subject assignment that most closely resembles the study under consideration. This initial classification is important because certain ratings only apply to specific designs.

Step 2 – Identify the outcome variable. The choice of outcome variable has consequences for rating study quality. The following guidelines are based on sexual recidivism as the outcome variable - a low base rate event. If general violence or any criminal recidivism is used as the outcome criterion, it is possible to obtain equivalent statistical power with shorter follow-up periods and smaller sample sizes because the base rate will be higher. As well, evaluators would want to use different control variables depending on the outcome of primary interest. The predictors of general violence and any criminal recidivism are not identical to the predictors of sexual recidivism.

Step 3 – Rating the individual items. The individual ratings address either confidence (e.g., sample size) or bias (e.g., equivalence of follow-up). The confidence and bias ratings are scored on three-point scales (little/some/high confidence; considerable/some/negligible bias).

The item should be rated first on the information presented in the article. On the next page, the item can be rated a second time based on new information obtained from other sources, or by new analyses. We encourage raters to obtain such information when it would improve the quality of the ratings or improve the quality of the study. Examples of new information would be treatment manuals, or correspondence with the study's authors indicating that the person conducting the data analysis was blind to outcome. Examples of new analyses would be re-organizing the presented data into an intent-to-treat analysis, or conducting new analyses from the original raw data.

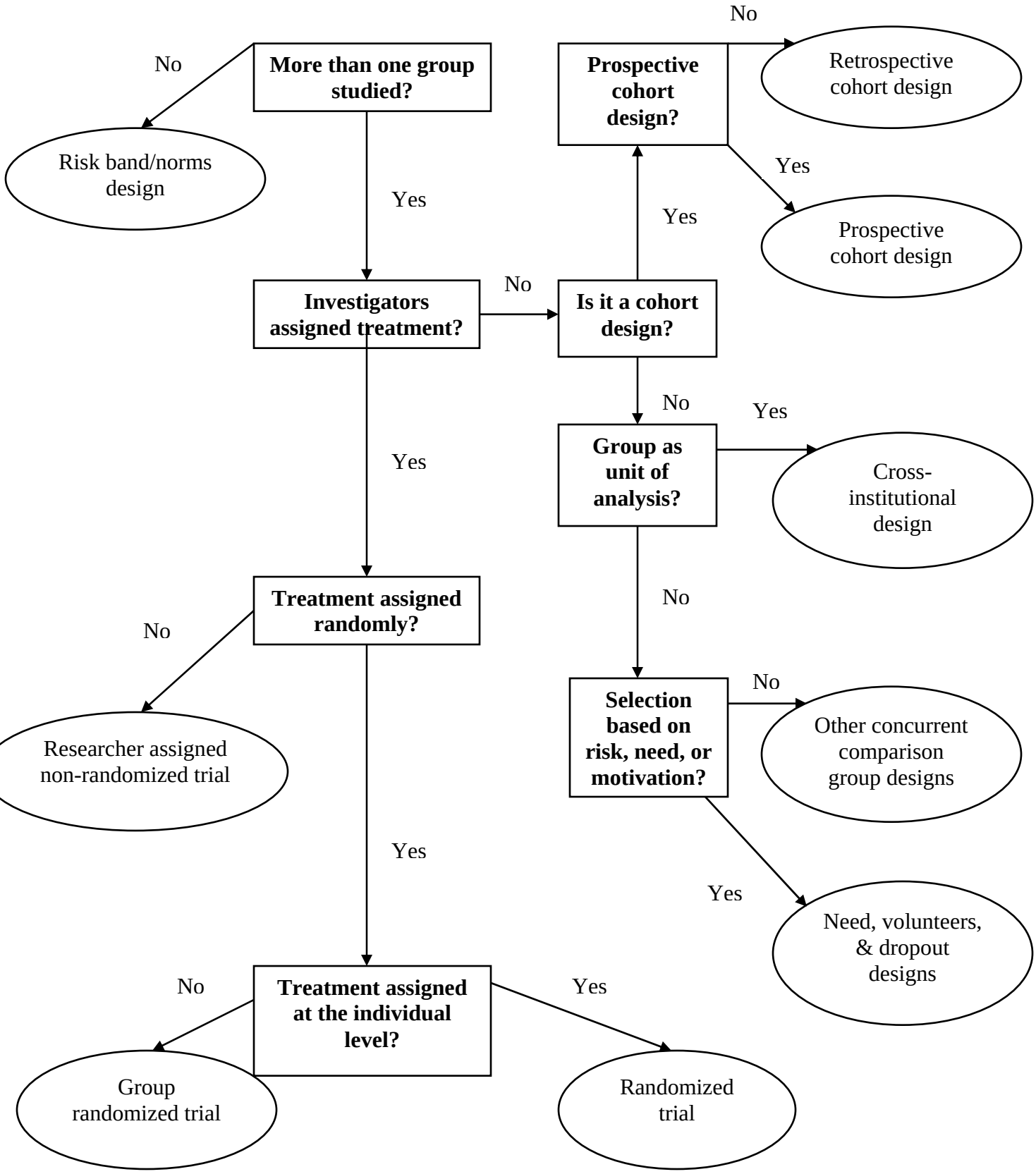
Step 4 – Form an overall quality rating. The final quality rating is based on overall ratings of "bias" and "confidence". Reviewers use their own professional judgement to determine overall ratings of "bias" and "confidence". Although the overall ratings would be expected to be proportional to the "total scores", we do not recommend any specific numeric algorithm to

arrive at the overall judgements of “bias” and “confidence”. Once overall bias and confidence ratings have been determined, however, explicit criteria are provided that translate these ratings into a four-point scale of study quality: rejected, weak, good, and strong.

The CODC Guidelines consist of 21 individual items, three summary ratings of confidence, quantity of bias, and direction of bias, and one overall rating of quality. One item (#8) is specific to cross-institutional designs and is only rated for this type of study. The criteria for rating some of the other items vary depending on how treatment and comparison groups were selected. The table below identifies the designs, the items, and the design-specific considerations that are used to assess that factor.

Design	Item
Cross-institutional designs	8. Sample size of institutions: Cross-institutional designs *This item is only rated for cross-institutional designs (see page 31).*
Random allocation designs	6. Sample size of treatment group(s) *The required sample size is smaller for random allocation designs than for other designs (see page 25).* 7. Sample size of comparison group(s) *The required sample size is smaller for random allocation designs than for other designs (see page 28).* 13. A priori equivalence of groups *Specific considerations of the allocation procedures are required to assess this item (see page 53).*
Risk band/norms designs	13. A priori equivalence of groups *Specific considerations of the validity of the norms or risk bands are required to assess this item (see page 54).*
Cohort designs	13. A priori equivalence of groups *Specific considerations of possible cohort effects are required to assess this item (see page 55).*

Identifying the design



Adapted from Zaza et al, 2000

I. Administrative control of independent variables

Administrative control of independent variables refers to the content and integrity of the treatment and comparison conditions. Without knowledge of the treatment given, it is impossible to know what was evaluated. Evaluation of vague, unarticulated interventions can be of interest to administrators narrowly concerned about the effectiveness of a specific treatment at a specific site. Researchers often have loftier ambitions, however, wishing to generalize the results to other settings and samples. Applying the results to other sites requires a clear definition of the treatment delivered. This would normally require treatment manuals and clear descriptions of what happened to the comparison group.

In addition to clearly defining the content of treatment, it is important to judge whether the treatment was implemented as intended (treatment integrity). This can be accomplished by training therapists and by "manipulation checks" that monitor the actual delivery of treatment (e.g., observing what goes on in actual treatment sessions). In prospective studies, investigators need to make decisions regarding how "rigid" treatment protocols are to be. A study may require that therapists rigidly adhere to a manual or may allow some deviations based on participant needs, risk level or progress. If many adjustments are allowed, it becomes difficult to describe the treatment program clearly. On the other hand, real clinical situations may dictate some flexibility (e.g., a community treatment program may need to intensify treatment if an individual's risk level increases and puts potential victims at risk). Although it may be good clinical practice, this flexibility can result in significant deviations from the treatment protocol or even treatment crossover. For example, if control participants are more likely to show increased risk and, therefore, need interventions for safety reasons, the original design of comparing a treatment group to a no-treatment condition is weakened. Investigators need to establish a priori the criteria that justify deviation from the treatment protocol.

Retrospective studies present significant challenges when attempting to understand the interventions given. Available documentation is often scarce. If therapists and clients could be interviewed, it is unlikely that their accounts would contain the details desired by researchers. Even if extensive documentation is available, the official version of a treatment program may differ from what actually happened.

Integrity issues are also relevant to the comparison group(s). To evaluate potential differences found between the treated and comparison groups, knowledge of the policies, procedures associated with, and experiences of the comparison group is necessary. It is common for sexual offenders to be exposed to certain services, legal provisions, and/or special supervisory practices even if they did not receive the treatment that is being evaluated. It may be that untreated offenders garner special attention because they did not receive treatment. To critically evaluate a study, it is important to know about these special conditions and how they might affect the recidivism of the comparison group(s).

The CODC Guidelines contain three items to assess the administrative control of independent variables. The first item, defining treatment, assesses the confidence that offenders in fact received the treatment as described, and whether the treatment could be replicated. Relevant information includes such details as the length of treatment, number of sessions, number of hours in treatment, theoretical orientation, content of the treatment program (e.g., treatment manual), location of treatment, and procedures and/or measures to ensure the integrity of treatment. Ratings on this item reflect the coder's confidence, based on available information, that the treatment was delivered as described, with integrity, and could be replicated by others.

The second item, defining comparison, assesses the confidence that offenders in the comparison condition(s) were exposed to the conditions as described (and not something else) and that these conditions could be replicated. Relevant information includes the services, supervision practices, any special considerations, and likelihood of receiving other forms of treatment. Ratings on this item reflect the coder's confidence that the comparison group was exposed to the conditions described and were not exposed to any special services or practices.

The third item, miscellaneous incidental factors, assesses the bias that may be introduced from exposure to extraneous factors not specifically related to the treatment program or comparison procedures *per se*. Relevant information includes the physical location of treatment (e.g., isolated secure mental hospital, jail/prison), and potential differences in supervision practices between treated and untreated offenders. Ratings on this item reflect the coder's assessment of potential bias that may be introduced resulting from extraneous factors that could differentially affect the outcome of the group(s).

1. Defining treatment

Concept: The general concept is whether or not the treatment could be replicated. Was the treatment program described in sufficient detail to re-create it? How confident are you that the program described was the program delivered to the participants? There are two factors to consider: content and integrity. Was there sufficient information to deliver the same treatment, including content, intensity (i.e., frequency of contact), dosage, and style of delivery? What procedures were put in place to ensure that the participants actually received the intended interventions? The more information describing what facilitators and participants did in treatment, the higher the rating of confidence that treatment was delivered as described.

Indicators: Potential indicators that would increase confidence that the treatment is replicable and was given as intended would include such things as a comprehensive treatment manual, specific information on the duration and frequency of sessions, including the overall duration of treatment, training and supervision of facilitators, and manipulation checks verifying that treatment was delivered as intended (e.g., the number of sessions actually attended by participants). When most or all of these indicators are missing and general statements of theoretical orientation (e.g., cognitive-behavioural approach) are used to describe treatment, there is little confidence. When there are some indicators, such as a manual but little training or no manipulation check, one has some confidence that the treatment is replicable and given as intended.

Cross-institutional design considerations: Cross-institutional designs are unique in that they use institutions, rather than offenders, as the unit of analysis. They require the researcher to sample a number of institutions and assess specific, theoretically grounded institutional factors to evaluate their impact on recidivism. The selection and operational definition of institutional factors is critical and should be well-defined constructs theoretically related to recidivism.

Rating	Description
0	*There is no treatment manual; content of treatment is vaguely described. Information may be provided identifying the duration of the program or frequency and length of sessions, but the total number of hours of therapist-client contact is not specified. It is unknown if or how facilitators deviated from the program. Facilitator training is informal and there is little or no supervision. No manipulation check was done to verify adherence to the program.*
1	*Content and structure of the treatment program can be identified via documentation such as a program agenda that identifies session topics, handouts, and/or goals. Information identifies the total number of hours, session frequency and duration of the program. There are formal training and supervision procedures in place. Manipulation check procedures are either informal or considered part of supervision.*
2	*A comprehensive treatment manual is available. There were formal training procedures used and facilitators provided treatment under supervision. Manipulation checks verified that treatment was delivered as described and intended.*

1. Defining treatment

This item is concerned with the level of confidence that the treatment group actually received the treatment as described. Information to consider includes details on length of treatment, number of sessions, amount of hours in treatment, theoretical orientation, content of the treatment program (e.g., treatment manual), and the procedures/measures employed to ensure its integrity.

Treatment information extracted from the study:			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence that treatment was delivered as described*	**1** *Some confidence that treatment was delivered as described but could be more convincing / have some reservations*	**2** *High confidence that treatment was delivered as described*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Defining treatment		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that Rx was delivered as described*	**1** *Some confidence that Rx was delivered as described but could be more convincing / have some reservations*	**2** *High confidence that treatment was delivered as described*
Reason(s) for rating		

2. Defining comparison

Concept: The general concept is the extent to which the comparison condition could be replicated. The goal of this item is to determine the coder's confidence that the author(s) described in sufficient detail the conditions of the comparison group to re-create it and that these conditions were in fact those to which the comparison group was exposed. Similar to the previous item, there are two factors to consider. One factor is the quality of the description of the content of the comparison condition. This includes the services that comparison offenders were expected to receive. The second factor is the integrity of the comparison condition: Did they actually receive these and only these services? If the comparison group was provided an alternative treatment or different supervision practice, evaluate the available information about the alternative interventions.

Indicators: Potential indicators that would increase confidence that the conditions of the comparison group are replicable and as described would include the following: clear descriptions of the services they received; the existence of supervision policies and procedures; and manipulation checks verifying that these participants did not receive any special, different, or treatment services. When most or all of these indicators are missing or there is only a general statement that the comparison group did not receive the treatment under investigation, there is little confidence. One has confidence that the conditions of the comparison group are replicable when there are some indicators, such as a description of the content of the comparison condition, as well as verification that participants did not receive any additional or unplanned treatment or supervision services.

Risk band/norm designs: In these studies, the comparison group is a specific normative sample. Consequently, the rating for the risk band/norm designs would typically be "2 – high confidence". The ratings could be less than 2, however, if the norms are believed to be unreliable, and researchers should use other normative samples if attempting to replicate the study.

Rating	Description
0	*Comparison group is simply described as offenders who did not receive the treatment in question and no other information is presented regarding the conditions to which they were exposed.*
1	*Some description of the comparison condition, including supervision practices and/or activities while incarcerated, but no information is presented ensuring that these subjects did not receive treatment or other differential services not otherwise specified.*
2	*A clear and detailed description of the comparison condition including standard supervision practices or typical activities while incarcerated and manipulation check evidence that verified comparison subjects did not receive treatment or other differential services not otherwise specified.*

2. Defining comparison

This item is concerned with the procedures provided to the comparison group(s) and the level of confidence that the conditions of this group are replicable and as described. Information to consider includes the services and/or supervision they received, including equivalent or similar interventions provided to the treatment group, and the integrity of the comparison group (e.g., this group did not receive treatment). If the comparison group was provided an alternative treatment, evaluate the available information about the alternative treatment.

Comparison information extracted from study			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence that comparison group was exposed as described*	**1** *Some confidence that the comparison group received services/supervision as described but could be more convincing/ have some reservations*	**2** *High confidence that comparison group was exposed to conditions as described*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Defining comparison		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that comparison group was exposed as described*	**1** *Some confidence that the comparison group received service /supervision as described but could be more convincing / have some reservations*	**2** *High confidence that comparison group was exposed to conditions as described*
Reason(s) for rating		

3. Miscellaneous incidental factors

Concept: The general concept is whether or not there were any miscellaneous incidental factors that may introduce bias into the results by differential exposure of the treatment and comparison groups to factors that may be related to recidivism. The goal of this item is to assess the magnitude and direction of bias that results from these factors. Important differential incidental factors may be present during the provision of treatment or comparison conditions (e.g., treatment occurred in a secure remote mental health hospital whereas comparison subjects were incarcerated in general population of a maximum-security jail) or after the provision of services (e.g., treated offenders were exposed to different community supervision practices than the comparison offenders).

Indicators: Potential indicators that would increase the potential for bias would be substantially different locations or different supervision practices while incarcerated or while in the community, particularly practices that may affect recidivism.

Cross-institutional design considerations: Miscellaneous incidental factors are the variables of interest in cross-institutional designs as it is these variables that are evaluated. It is particularly important that the researcher vigorously explored and tested alternative hypotheses, as it is unlikely that the researcher assessed all incidental factors.

Rating	Description
0	*The settings and services delivered to the treatment and comparison groups differ on at least one factor that would be expected to differentially affect recidivism rates. For cross-institutional designs, these incidental factors cannot be separated from other features of interest.*
1	*The treatment and comparison groups differ on at least one factor that may have a relationship with recidivism. The effect of this incidental factor is unknown. For cross-institutional designs, this incidental factor cannot be fully separated from other features of interest. The effect of this incidental factor is unknown.*
2	*There are no incidental factors inherent in the provision of services or during follow-up that would be expected to differentially influence the recidivism of the groups. The important elements of the treatment and comparison settings are the same.*

3. Miscellaneous incidental factors

This item is concerned with miscellaneous incidental factors that could introduce bias into the results. It is important to be aware of, and to assess, the similarity of incidental factors between the treatment and comparison groups. Examples of incidental factors are the physical location of treatment (e.g., isolated secure mental hospital, jail/prison), and potential differences in supervision practices between treated and untreated offenders.

Information on miscellaneous incidental factors extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Miscellaneous incidental factors			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

II. Experimenter expectancies

Experimenter expectations can have a significant effect on the results of even randomized clinical trials (see Zaza et al., 2000; Juni et al., 1999), which is why double-blind procedures are often used. Double-blinding cannot be directly applied to sexual offender treatment because it is not realistic to expect that subjects and experimenters will not know who receives treatment. Some reviewers even consider it advantageous for the experimenter to be involved in both program development and delivery as it may lead to higher levels of treatment integrity (see Gendreau, Goggin, & Smith, 1999). Nevertheless, those with a vested interest in a program are at risk of making decisions favourable to demonstrating positive treatment outcome.

Blinding is also important in data collection and analysis. Knowledge of the group to which an offender belongs may inadvertently influence assessment and coding decisions. In sex offender treatment outcome studies, it is feasible and desirable for data coders and managers to be blind to the subjects' group membership and outcome.

The CODC Guidelines have two items that assess experimenter expectancies. The first item, experimenter investment in outcome, assesses whether bias is introduced by the experimenters through their involvement in the research and delivery of services. Bias is most likely to be introduced when the experimenter has a vested interest in the outcome of the evaluation, and has the potential to influence day-to-day decisions concerning program delivery and data collection. Experimenters would be expected to have a vested interest in the outcome when they are evaluating their own programs, or programs to which they are closely aligned (e.g., through institutional affiliation). Relevant information includes any special efforts or activities that may have been introduced due to the experimenter's involvement in treatment and other procedures for the comparison group, and whether an experimenter's expectancies could influence data collection. Ratings of the magnitude and direction of bias are required.

The second item, blinding in data management, assesses the degree to which the data managers may have introduced bias because they know group assignment as they are collecting and analyzing the data (particularly coding recidivism information).

4. Experimenter investment in outcome

Concept: The general concept is the extent to which the experimenter has a vested interest or a stake in the treatment and the extent of the experimenter's capacity to influence the outcome of the evaluation (e.g., their level of involvement in the day-to-day delivery of services). The goal is to assess the bias that may be introduced by their involvement in the treatment and its evaluation.

Indicators: Potential indicators of bias introduced by experimenter involvement can occur when the investigator has a vested interest or stake in the treatment program, and the evaluator is directly involved in the day-to-day operations of the program and delivery of services. Although experimenter involvement can affect treatment integrity (assessed by the item defining treatment), this item assesses the influence that the experimenter may have on the results of the evaluation by day-to-day decisions that may introduce small amounts of bias in the results (e.g., admission decisions, and additional interventions that enhance motivation and compliance that have not been described or noted). Some bias may be introduced when the researcher evaluates his/her own work.

Rating	Description
0	*Experimenter(s) has a vested interest in the program and evaluation, and is directly involved in day-to-day administration of the program, including such things as admission decisions, attrition, and direct provision of treatment.*
1	*Experimenter is part of the evaluation team that is linked indirectly via a consulting or supervisory role to the operation of the program. The primary investigator is not involved in the direct delivery of services. It is unclear if their role would affect day-to-day decisions. A rating of 1 would be given for evaluations conducted by a separate section of the same organization (e.g., research or audit branch of the same government department delivering the program).*
2	*Experimenter is a third-party evaluator of program. The primary investigator is not involved in the management or delivery of services. There is no reason to believe that the experimenter has a vested interest in the outcome, either through personal involvement or institutional affiliation.*

4. Experimenter investment in outcome

This item is concerned with the potential bias that may result from the experimenter's vested interest and/or direct involvement in the management and delivery of services, particularly when the experimenter is directly responsible for day-to-day decisions.

Information on experimenter involvement extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Experimenter investment in outcome			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

5. Blinding in data management

Concept: The general concept is the degree to which those responsible for the management and coding of data knew, or could reasonably know, group membership and outcome status (i.e., recidivism) of study participants, and whether this knowledge potentially introduced bias.

Indicators: Potential indicators of bias in data management procedures can occur when the data coders/managers had a priori knowledge of the offender's group membership and/or recidivism status during data collection periods. Although not critical for simple transcription or photocopying of information, it is crucial when rating or coding decisions are required. For example, prior knowledge of the offender being in the treatment group may influence coding decisions during retrospective assessment of risk. Ideally, the person responsible for the collection, management, and analysis of the data would be blind to group membership and outcome status. When no information is provided concerning how the data were collected and analyzed, this item should not be rated (i.e., "insufficient information to evaluate").

Rating	Description
0	*Researchers/data managers had a priori knowledge of group membership or recidivism status during data procedures requiring decisions or judgements.*
1	*It is unclear if researchers/data managers had a priori knowledge of group membership or recidivism status during data procedures requiring decisions or judgements.*
2	*Researchers/data managers were blind to group membership and recidivism status during data collection procedures requiring decisions or judgements.*

5. Blinding in data management

This item is concerned with whether those who collected and recorded data on the study's participants could introduce bias into the results. Data such as risk assessment information and recidivism are critical in the evaluation of the effect of treatment. Note any information that would indicate whether or not the person(s) who gathered, assessed or recorded the information knew the treatment status and/or recidivism status of the participants.

Information on blinding in data management extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Blinding in data management			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** Some bias likely introduced in the results	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

III. Sample size

Confidence in the results of a study increases with sample size. The confidence interval of an effect size decreases as the sample size increases, as does the statistical power to detect pre-existing differences between groups. For random assignment studies, there is the expectation that the error between the groups will cancel out (i.e., be random). Consequently, there is less demand on studies using random assignment than there is for other designs to ensure there are no pre-existing differences between groups. When participants are well matched on risk prior to randomization, there is even less need to check for potential pre-existing difference than there is in other random design studies. For all studies, however, it is desirable to have large sample sizes. Even well conceived studies break down, and adequate statistical power allows reviewers to explore potential deviations from the intended protocol.

For non-random assignment designs, there is a strong need to verify the pre-test equivalence of the treatment and comparison groups. As sample sizes decrease, it becomes increasingly difficult to detect significant pre-existing differences (i.e., biases inherent in the composition of the groups). To obtain equal levels of confidence, non-random allocation designs require larger sample sizes than random assignment designs. The smaller sample sizes for random assignment studies, however, are only appropriate when the random assignment procedure is well implemented. Given a breakdown in the random assignment process, reviewers need to examine the equivalence of the groups, and the usual (larger) sample sizes are required.

Statistical power is optimized when the overall study sample is equally split between groups (i.e., 50% of the total study sample). The loss of power is minimal with unequal sample sizes, however, provided that the split is between 30% and 70% of the total study sample. For example, the maximum power in a study with an overall sample size of 600 occurs when there are 300 subjects in the treatment group and 300 subjects in the comparison group. The loss of power is minimal when one group has 180 subjects (30% of the total study sample of 600) and the other group has 420 subjects. When the deviation is outside this range, the power to detect statistical differences between the groups is reduced (Cohen, 1988).

In cross-institutional designs, the unit of analysis is institutions rather than individual offenders. These designs are unique in that sample size also refers to the number of institutions. As these designs define and assess constructs/characteristics of institutions, it is important that the number of institutions is sufficiently large to provide variation in the construct of interest. One's confidence in the results increases as more institutions are sampled.

The CODC Guidelines contain three items to assess sample size. The first item assesses the confidence that the sample size of the treatment group is sufficiently large to detect significant pre-existing differences that may influence outcome. Sample size specifically refers to the number of participants for which follow-up information has been obtained. Confidence ratings are based on the sample size required to have sufficient power (i.e., 70% power) to detect small

(i.e., d = .20), medium (i.e., d = .50), and large effects (i.e., d = .80)(Cohen, 1988). Criteria specific to random and non-random allocation designs are presented.

The second item assesses the confidence that the sample size of the comparison group is sufficiently large to detect significant pre-existing differences that may influence outcome. Confidence rating criteria are identical to the previous item.

The third item is specific to cross-institutional designs and is not intended for use with other designs. It assesses the confidence that the sample size of institutions is sufficiently large to provide variation in the construct of interest and detect significant differences between institutions. Where there are less than six institutions sampled, there is little confidence. With more than eight institutions, one has confidence that the sample is sufficiently large to evaluate and examine differences in recidivism rates. Confidence increases with the number of institutions in which the construct is present and the number of institutions in which the construct is absent. Confidence would generally be higher given 50-50 splits (four versus four) than extreme splits (one versus seven).

For cross-institutional designs, statistical power is still determined by the number of offenders, not the number of institutions. Consequently, it is useful to rate the sample size of treatment group and sample size of comparison group items along with sample size of institutions item.

6. Sample size of treatment group(s)

Concept: The general concept is the adequacy of the sample size to detect significant pre-existing differences between groups. The goal of this item is to assess the coder's confidence that the sample size is sufficient to provide 70% power to detect small (d = .20), medium (d = .50), and/or large effects (d = .80) between groups. Confidence increases when there is a high probability that the sample size is large enough to detect such differences (i.e., power). For well-implemented random allocation designs, the assignment of subjects minimizes the probability of pre-existing differences between groups. Therefore, confidence rating criteria for this type of design are the sample sizes required to detect medium (d = .50) and large effects (d = .80) with 70% power. For all other designs, demonstrating the pre-existing equivalence of the groups is crucial. Therefore, criteria for the confidence rating of these designs are the sample sizes required to detect small (d = .20) and medium effects (d = .50) with 70% power. The values were taken from Cohen's (1988) power table 2.3.5 (two tailed α of .05).

Indicators: To rate this item, the design must be identified and the overall study sample size and the sample size of the treatment group with follow-up information must be known. There is some flexibility in the ratings, as the sample sizes described below are based on 50% of the overall N in the study. As long as the overall N is double the criteria noted below and the treatment group represents between 30% and 70% of the overall N, then that confidence rating is appropriate.

Rating	Description
0	*For random allocation designs, treatment sample size less than 20.* *For all other designs, treatment sample size less than 50.*
1	*For random allocation designs, treatment sample size greater that 20 but less than 50.* *For all other designs, treatment sample size greater than 50 but less than 300.*
2	*For random allocation designs, treatment sample size of 50 or more.* *In all other designs, treatment sample size of 300 or more.*

6. Sample size of treatment group(s)

This item is concerned with the sample size of the treatment group with follow-up information and the level of confidence that this sample size provides sufficient power to detect differences. Decisions concerning the adequacy of the sample size should be based on the sample size with follow-up information.

Overall study sample size:	*N =*
Sample size of treatment group with follow-up information:	*n =*
Percentage of study subjects in treatment group:	*%*

Confidence rating			
--	**0**	**1**	**2**
Insufficient information to evaluate	*Little confidence that the sample is large enough to provide sufficient power to detect differences*	*Some confidence that the sample is large enough to provide sufficient power to detect differences*	*High confidence that the sample is large enough to provide sufficient power to detect differences*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Sample size of treatment group(s)		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the sample is large enough to provide sufficient power to detect differences*	**1** *Some confidence that the sample is large enough to provide sufficient power to detect differences*	**2** *High confidence that the sample is large enough to provide sufficient power to detect differences*
Reason(s) for rating		

7. Sample size of comparison group(s)

Concept: The general concept is the adequacy of the sample size to detect significant pre-existing group differences that may affect outcome. The goal of this item is to assess the coder's confidence that the sample size is sufficient to provide 70% power to detect small (d = .20), medium (d = .50), and/or large effects (d = .80) between groups. Confidence increases when there is a high probability that the sample size is large enough to detect such differences (i.e., power). For well-implemented random allocation designs, the assignment of subjects minimizes the probability of pre-existing differences between groups. Therefore, the confidence rating criteria for this type of design are the sample sizes required to detect medium (d = .50) and large effects (d = .80) with 70% power. For all other designs, demonstrating the pre-existing equivalence of groups is crucial. Therefore, criteria for the confidence rating of these designs are the sample sizes required to detect small (d = .20) and medium effects (d = .50) with 70% power. The values were taken from Cohen's (1988) power table 2.3.5 (two tailed α of .05).

Indicators: To rate this item, the design must be identified and the overall study sample size and the sample size of the comparison group with follow-up information must be known. There is some flexibility in the ratings, as the sample sizes described below are based on 50% of the overall N in the study. As long as the overall N is double the criteria noted below and the comparison group represents between 30% and 70% of the overall N, then that confidence rating is appropriate.

Risk band/norm designs: The comparison group in these designs are normative samples taken from other studies. If these norms are reasonable, a rating of "2 – high confidence" is appropriate.

Rating	Description
0	*For random allocation designs, comparison sample size less than 20.* *For all other designs, comparison sample size less than 50.*
1	*For random allocation designs, comparison sample size greater than 20 but less than 50.* *For all other designs, comparison sample size greater than 50 but less than 300.*
2	*For random allocation designs, comparison sample size of 50 or more.* *In all other designs, comparison sample size of 300 or more.*

7. Sample size of comparison group(s)

This item is concerned with the sample size of the comparison group with follow-up information, and the level of confidence that this sample size provides sufficient power to detect differences. Decisions about the adequacy of sample size should be based on the sample size with follow-up information.

Overall study sample size:	*N =*
Sample size of comparison group with follow-up information:	*n =*
Percentage of study subjects in comparison group:	*%*

Rating

--	0	1	2
Insufficient information to evaluate	*Little confidence that the sample is large enough to provide sufficient power to detect differences*	*Some confidence that the sample is large enough to provide sufficient power to detect differences*	*High confidence that the sample is large enough to provide sufficient power to detect differences*

Reason(s) for rating:

Is it worth looking for additional information or new analyses? If yes, complete next page.

Information: Sample size of comparison group(s)		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the sample is large enough to provide sufficient power to detect differences*	**1** *Some confidence that the sample is large enough to provide sufficient power to detect differences*	**2** *High confidence that the sample is large enough to provide sufficient power to detect differences*
Reason(s) for rating		

8. Sample size of institutions: Cross-institutional designs

Concept: As the unit of analysis in cross-institutional designs is the institution rather than the offender, the number of institutions sampled is critical. The sample size should be sufficient so that more than one institution represents variation in the construct of interest.

Indicators: When making the judgement of sample size, it is worth considering both the number of institutions and the number of subjects represented by the institutions. More institutions are needed if the number drawn from each institution is small (see previous items).

Rating	Description
0	*Data were collected and the analysis was based on less than six institutions/sites.*
1	*Data were collected and the analysis was based on six to eight institutions/sites.*
2	*Data were collected and the analysis was based on more than eight institutions/sites.*

8. Sample size of institutions: Cross-institutional designs

This item is concerned with the sample size of the institutions and the level of confidence that this sample size provides sufficient variability in the construct of interest to detect differences. Information about the number of institutions sampled and the sample size of each should be recorded.

<table>
<tr><td colspan="2">Overall number of institutions sampled:</td><td colspan="2">N =</td></tr>
<tr><td colspan="2">Sample size of each institution with follow-up information:</td><td>n₁ =
n₂ =
n₃ =
n₄ =
n₅ =</td><td>n₆ =
n₇ =
n₈ =
n₉ =
n₁₀ =</td></tr>
<tr><td colspan="4">Confidence rating</td></tr>
<tr><td>-
Insufficient information to evaluate</td><td>0
Little confidence that the sample is large enough to provide sufficient variation in the construct of interest</td><td>1
Some confidence that the sample is large enough to provide sufficient variation in the construct of interest</td><td>2
High confidence that the sample is large enough to provide sufficient variation in the construct of interest</td></tr>
<tr><td colspan="4">Reason(s) for rating:</td></tr>
</table>

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Sample size of institutions: Cross institutional designs		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the sample is large enough to provide sufficient variation in the construct of interest*	**1** *Some confidence that the sample is large enough to provide sufficient variation in the construct of interest*	**2** *High confidence that the sample is large enough to provide sufficient variation in the construct of interest*
Reason(s) for rating		

IV. Attrition

Attrition of participants is an important consideration when evaluating study quality. One needs to consider how many participants were lost and at what point in the investigation the attrition occurred. As attrition increases, the potential for bias increases. Specifically, attrition first occurs during the subject selection process. At this stage, study participants, particularly potential treatment participants, are identified. There are certain offenders who will not be considered by the investigator(s) for inclusion (e.g., offenders who are presently psychotic). Additionally, there are offenders who refuse to participate or outright reject treatment (e.g., refusing to participate because "he is not a sexual offender").

It is extremely difficult to interpret the recidivism rate of these "refusers" in comparison to the treatment group. If the recidivism rate of these "refusers" is similar to the comparison group and is higher than the treatment group, this would appear to support the efficacy of treatment. It appears that the "refusers" could have reduced their recidivism potential by participating in treatment. If, however, the "refusers" were included with the treatment group in an intent-to-treat design, then this would reduce the apparent treatment effect. Conversely, if the "refusers" demonstrate a low recidivism rate (similar to the treatment group and lower than the comparison group), then it would appear that treatment was not necessary, and could even have had a detrimental effect for some offenders. Paradoxically, including these "refusers" with the treatment group would increase the apparent efficacy of treatment. Although the empirical evidence suggests that the observed recidivism rates of those who refuse treatment are similar to those who start treatment (Hanson et al., 2002), the most easily interpreted studies are ones in which the treatment is accepted by most of the offenders to whom it is offered, and when there are few potential "refusers" buried in the comparison group.

The population of "treatable" offenders is limited to offenders who would be considered for participation (i.e., meet some selection criteria) and offenders who would consider participating in the study or treatment. It is this subset that provides the basis for evaluating the effectiveness of treatment. Narrow selection criteria, however, can lead to an artificially homogeneous sample, thereby limiting the generalizability of the results (Westen, Novotny, & Thompson-Brenner, 2004). In addition, a highly restrictive selection criterion increases the difficulty in constructing an equivalent comparison group.

The amount of bias is related to the equivalence of the selection criteria for the treatment and comparison groups (e.g., high risk, motivated, willing to consider participating, have sufficient time left in sentence to complete treatment). Clear and specific information on selection criteria for those offenders interviewed for potential participation in the study (either in treatment and/or comparison groups, if applicable) allows the reader to estimate the potential effects of selection biases.

In addition to initial selection bias, offenders may drop out of treatment at various points in time. For example, some offenders may drop out after being accepted into treatment but prior to actually starting treatment, and others may quit after attending one or more treatment sessions. We refer to this loss as program attrition.

There is strong evidence that treatment dropouts are at higher risk to reoffend than are treatment completers (Hanson et al., 2002). Consequently, information about treatment dropouts should be routinely recorded, and treatment dropouts should be included with the treatment group (intent-to-treat). As well, researchers should record the reasons for attrition: a) termination from treatment by the treatment providers due to misbehaviour within or outside the program; b) voluntary withdrawal (e.g., lack of interest/motivation); and c) benign administrative reasons (e.g., program cancelled). In designing prospective studies, investigators need to consider methods to limit attrition during treatment and be aware of the risk of differential attrition. For example, differential attrition can occur when high risk, more difficult participants drop out of a rigorous treatment program while high-risk participants in the control group do not drop out because little is required of them. Although some of the bias introduced by attrition can be handled through intent-to-treat analyses, high rates of attrition may be better described as program implementation failure rather than a test of treatment effectiveness. Prospective studies should include measures designed to establish and sustain offender participation.

Loss of participants during follow-up is also a threat to validity. If criminal justice records alone are used to detect recidivism, all participants should be able to be tracked with equal effectiveness. One potential problem with using official records is the deletion of older, inactive records (Hanson & Nicholaichuk, 2000). When the treatment and comparison groups are selected from different time periods or different jurisdictions, then the methods of record retention may introduce bias. For example, in retrospective designs, the reverse matching of existing records in which groups are selected from existing records may introduce bias because only certain offenders would still have criminal history records (typically the young, criminally active offenders). An additional problem is that institutional review boards may not allow investigators to collect or retain data on individuals who withdraw from the study. The likelihood of differential attrition rates during follow-up is greatest when follow-up data are collected directly from participants. As well, researchers may have a much closer relationship with those in active treatment as compared to those in the comparison group. Good prospective studies need to have explicit methods for tracking participants and maintaining participants in the protocol (Shadish, Cook, & Campbell, 2002).

The CODC Guidelines contain four items assessing the potential bias introduced by attrition. The first item assesses the potential bias of the explicit criteria used for subject selection in the treatment and comparison groups. Relevant information includes criteria used to identify the pool of potential offenders, including those offenders who refuse participation. If the criteria used create an expectation that one or more groups would be different on important variables, then this introduces bias. Finally, if selection criteria are strict or idiosyncratic, it will be

difficult to generalize the results to other populations. Ratings of the estimated magnitude and the direction of bias (i.e., favours treatment effectiveness or decreases the effectiveness of treatment) are required.

The second item assesses the potential bias introduced by program attrition. This item is concerned with attrition that occurs after the condition is assigned or offered to the offender but prior to completion (i.e., dropouts). Relevant information includes the number of, reasons for, and characteristics of, the offenders who dropped out early (i.e., after agreeing to participate but prior to formally starting the program) and later (i.e., after formally starting the program). Ratings of the estimated magnitude and direction of bias are required. This item is not concerned with the analytic procedures used (e.g., intent-to-treat), but is only concerned with the number of, reasons for, and characteristics of the "dropouts". If the number of dropouts is sufficiently large (greater than 50%), the study is rated as an "implementation failure".

The third item, intent-to-treat, assesses how the investigator(s) handle the aforementioned attrition when estimating the effects of treatment. Bias can be introduced when the rate of program attrition is more than minimal and when these participants are excluded from the analysis. In all cases, an intent-to-treat analysis is the least biased approach (i.e., include the "dropouts" when calculating the overall recidivism rate for the treatment group). It is important to record any information provided on the attrition group (e.g., risk, proportion of overall group, recidivism rates, reasons for attrition), and whether or not this group was included in the treatment group in the estimation of treatment effectiveness. Ratings of the estimated magnitude and direction of the bias, specifically in relation to the intent-to-treat issue, are required.

The fourth and final item in the attrition section is attrition in follow-up. Relevant information includes the number of participants lost due to lack of, or insufficient, follow-up information. Bias due to attrition during follow-up may be introduced when there are differential attrition rates between groups, or when the overall rates are high in both groups. Ratings of the estimated magnitude and direction of the bias due to attrition in follow-up are required.

9. Subject selection

Concept: This item concerns the bias introduced by the criteria by which offenders are considered for a treatment program. All programs set limits on who is accepted into (and/or denied) treatment based on such factors as perceived needs (e.g., deviant sexual preferences) and capacity to benefit (e.g., motivation, lack of active psychosis). Such limits can potentially introduce bias and limit generalizability. This item is not directly concerned with the study's design or procedures by which subjects are assigned to research groups as this aspect is addressed in item 13 (i.e., evaluating the a priori equivalence of groups).

Selection criteria for treatment may be general and broad (e.g., has a sexual offence conviction on record) or specific and narrow (e.g., self-referred, undetected pedophiles treated at private clinics). The narrower and more specific the criteria, the greater the likelihood that bias is introduced and that the results will not generalize to the general population of sexual offenders.

The inclusion/exclusion selection criteria for treatment may be strongly related to risk (e.g., deviant sexual preferences, high risk offenders, psychopaths) or minimally related to risk (e.g., time left in sentence, active psychosis, language, literacy). The stronger the selection criteria's relationship to risk, the more important it is to consider the impact of the inclusion/exclusion criteria on the characteristics of the research groups. For example, when the exclusion criteria have clear links to risk (e.g., psychopaths), a substantial amount of bias may be introduced if the identical exclusion criteria are not used for the comparison group (e.g., the comparison group may contain psychopaths).

Indicators: In order to evaluate this item, identify the selection criteria for admission into treatment (including inclusion and exclusion criteria). When these criteria are general and broad, there is a greater chance that the treatment and comparison group are comparable, particularly when identical criteria are used for both groups. When the selection criteria are narrow, in addition to limiting generalizability, this increases the importance of using identical criteria for the comparison group.

Another indicator is the exclusion/inclusion criteria's relationship to risk. Examples of selection factors that are clearly linked to risk include restricting treatment to those with deviant sexual preferences, selecting only "manageable" clients, or excluding offenders with personality disorders. Selection factors that have a weak or indirect relationship with risk include time remaining in sentence, active psychosis and responsivity factors such as motivation, language, and literacy. When the proportion of offenders excluded is small, clear links to risk are needed before substantial bias would be introduced. When the exclusion rate is large (e.g., only illiterate offenders were admitted into treatment), then substantial bias could be introduced by factors that have weak or indirect relationships with recidivism risk.

Finally, for studies in which the experimenter did not determine who received treatment (e.g., cohort studies where treatment was implemented over time), consider the proportion of the comparison group that would have been treated if treatment had been available to them (e.g., by estimating from the refusal/dropout rate of the treatment group).

Subject Selection Rating

Rating	Description
0	*The selection criteria between the two groups were not identical AND the exclusion criteria for treatment is directly linked to risk.* **OR** *The selection criteria between the two groups were not identical AND the relative proportion of excluded offenders from treatment is large in relation to the comparison group.*
1	*Selection criteria between the two groups were not identical. The proportion of excluded offenders is small, and the exclusion criteria for treatment would be expected to have minimal relationship to risk.* **OR** *Selection criteria are identical but unusually narrow, such that the treatment and comparison groups are not representative of other samples to which the intervention may be applied.*
2	*Selection criteria for both groups are identical. Inclusion and exclusion criteria are general, broad, and applicable to both treatment and comparison groups.*

9. Subject selection

This item is concerned with the requirements for "admission to treatment", including the inclusion and exclusion criteria. Bias increases as these criteria become narrower, and the use of these criteria to select comparison subjects becomes more difficult. Bias also increases as the number of potential "refusers" included in the comparison group increases.

Information on subject selection extracted from study (inclusion and exclusion criteria)		
	Treatment group	*Comparison group*
The number of all possible subjects	*N =*	*N =*
The number of subjects who were willing to consider and appropriate for inclusion	*N =*	*N =*

Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*

Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*

Reason(s) for rating

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Subject selection		
What additional information is desired and why?		
If new information was obtained, specify		
	Treatment group	*Comparison group*
The number of all possible subjects	*N =*	*N =*
The number of subjects who were willing to consider and appropriate for inclusion	*N =*	*N =*

Revised bias rating		
0 *Introduces a considerable amount of bias*	***1*** *Some bias likely introduced in the results*	***2*** *An expectation of negligible bias in the results*

Revised direction of bias			
? *Cannot assess the direction of bias*	***+1*** *Bias likely increases the magnitude of treatment effectiveness*	***0*** *No bias expected*	***-1*** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

10. Program attrition

Concept: The general concept is the bias introduced by attrition that occurred between the point at which an offender shows a willingness to participate in treatment (or to being assigned a specific study condition) and completion of that condition. Attrition of study participants happens when, after showing a willingness to participate and being accepted to participate, offenders change their mind and withdraw from the study. Some may drop out prior to attending any sessions (e.g., changing their minds once they learn that they have to move to a different institution). They may also withdraw from treatment (or the study) after having begun treatment for a variety of reasons (e.g., voluntary termination, termination by treatment providers, benign administrative reasons). The goal of this item is to assess the magnitude and direction of bias that results from this attrition. Bias increases as the proportion of dropouts increases and/or as the proportion of dropouts in one group differs from the proportion of dropouts in the other group. A program should be considered an implementation failure when the percentage of dropouts is too high.

Indicators: Potential indicators that would decrease bias can occur when a high proportion of participants completed the condition to which they were assigned. To evaluate the proportion of dropouts, it is necessary to know the number of subjects that considered participating, the number who withdrew their participation early on (i.e., from initial signs of interest to initial stages of treatment), reasons for withdrawal, and the number who withdrew later (i.e., started treatment or study participation but did not complete). When the percentage of those starting and completing is high, there is a decrease in bias.

Implementation failure: Dropouts are those who initially showed an interest in treatment or participation in the study (e.g., did not initially reject treatment/participation from the onset) but did not complete treatment or study participation. When the dropout rate is 50% or higher, it is considered an implementation failure.

Rating	Description
0	*The program is an implementation failure (i.e., dropout rate of 50% or higher)* **OR** *Between 51% and 59% completed treatment.* *The dropout rate is between 41% and 49%.*[1]
1	*Between 60% and 79% completed treatment.* *The dropout rate is between 21% and 40%.* [1]
2	*At least 80% completed treatment/participation.* *The dropout rate is 20% or less.* [1]

[1] These criteria were taken from Thomas et al. (2004).

10. Program attrition

This item is concerned with the attrition of participants after they had demonstrated an interest in participation (either in treatment or as a study participant) but prior to completion of the treatment or experimental condition (i.e., dropouts). Bias increases as the percentage of dropouts increases. If there is 50% or higher dropout rate, the bias in the results is too great, regardless of how the investigator handles the attrition when estimating treatment effectiveness, and the program is considered an implementation failure.

Information on program attrition extracted from study	*Treatment group*	*Comparison group*
The number of subjects who were willing to consider and appropriate for inclusion (See previous item)	*N =*	*N =*
Number of subjects who withdrew prior to finishing one month of treatment	*N =*	*N =*
Number of subjects withdrew during treatment	*N =*	*N =*
Number of subjects completing treatment	*N =*	*N =*
DROPOUT RATE		

Implementation failure? (i.e., dropout rate of 50% or higher). Provide reasons. *Bias rating = 0*	*No*	*Yes*

Bias rating			
-- *Insufficient information to evaluate*	**0** *A considerable amount of bias: Implementation failure*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*

Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Program attrition			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Rea*son(s) for rating*			

11. Intent-to-treat analysis

Concept: The general concept is how the investigator(s) handled program attrition when estimating the effectiveness of treatment. In all studies, investigators should use an intent-to-treat analysis to calculate effect sizes. However, as attrition rates increase, the bias of an intent-to-treat analysis increases too. The goal of this item is to assess the magnitude and direction of bias that stems from how the investigator handles attrition when calculating the effects of treatment.

Indicators: There are two important indicators to consider. One indicator is the amount of program attrition as assessed in the previous item. The second indicator is whether the investigator(s) used intent-to-treat analysis or if dropouts were handled in some other fashion (e.g., excluded from the analysis or included in the comparison group). Bias can be introduced when there is more than a minimal number of dropouts and when dropouts are excluded from the analysis. Bias is minimized when there is a small number of dropouts and when intent-to-treat analysis is performed.

Rating	Description
0	*Program attrition rating of 0 (i.e., dropout rate exceeds 41%), whether or not intent-to-treat analyses are used.* ***OR*** *Program attrition rating of 1 (i.e., dropout rate between 21% and 40%) and dropouts are excluded from the estimation of effect size.*
1	*Program attrition rating of 1 (i.e., dropout rate is between 21% and 40%) and dropouts are included (intent-to-treat) in the analysis.* ***OR*** *Program attrition rating of 2 (i.e., dropout rate is 20% or less) and dropouts excluded from effect size.*
2	*Program Attrition rating of 2 (i.e., dropout rate is 20% or less) and the effect size is calculated with an intent-to-treat analysis.*

11. Intent-to-treat analysis

This item is concerned with how the investigator(s) handled program attrition when estimating the effectiveness of treatment. Bias can be introduced when there is more than a minimal number of dropouts and when dropouts are excluded from the analysis. It is important to record any information regarding program attrition (e.g., risk, proportion to overall group, recidivism rates, reasons for attrition) and whether or not these dropouts were included in the treatment group in the estimation of treatment effectiveness.

Information on intent-to-treat extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Intent-to-treat analysis			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

12. Attrition in follow-up

Concept: The general concept is the bias introduced by attrition of subjects during follow-up. Typically, subjects are lost due to missing or insufficient follow-up information. In cohort studies and in retrospective designs, it is critical to be aware of potential attrition in follow-up due to the deletion of older, inactive records, different methods of record retention, and the reverse matching of existing records. Systematic and differential biases may be introduced simply through the differential availability of records between groups, time periods and jurisdictions. The goal of this item is to assess the magnitude and direction of bias that stems from attrition during follow-up.

Indicators: Potential indicators of decreased bias occur when almost all participants, including all dropouts, had adequate follow-up information. Potential indicators that would increase bias include high rates of attrition in follow-up information, missing or inadequate follow-up information for dropouts, or if subjects were selected for treatment or comparison condition based on the availability of records.

Rating	Description
0	*Follow-up information available for less than 70% of subjects in either group* *OR* *The difference in attrition rate between groups is greater than 10%.*
1	*Follow-up information available for a minimum of 70% of subjects in each group* *AND* *The difference in attrition rate between groups is 10% or less.*
2	*Follow-up information available for a minimum of 90% of subjects in each group.*

12. Attrition in follow-up

This item is concerned with attrition during follow-up. There are two potential sources of bias that may be introduced. One is missing follow-up information due to inadequate records or problems in finding the participants' records. Another problem (more serious) is when there are different practices/policies of record retention (e.g., old records are destroyed under certain conditions) and/or retrieval processes (e.g., data unavailable on participants who withdrew from study).

Information on attrition in follow-up extracted from study			
Percentage of treatment subjects with follow-up:		*Percentage of comparison subjects with follow-up:*	
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Attrition in follow-up			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

V. Equivalence of groups

Given that the major threat to validity of research is the lack of equivalence between groups, it is important for researchers to examine the extent of pre-existing differences between the treatment and comparison groups. Not all differences are important; the differences that matter are differences in recidivism potential. The ultimate outcome of interest is recidivism, but it is not possible to measure recidivism at intake. This is a special problem in sexual offender treatment outcome research. For many other disorders (e.g., depression, physical disability), it is possible to convincingly equate the groups at pre-treatment. Without direct access to the problem of interest, however, researchers are forced to rely on "proxy measures" to estimate equivalence (Shadish et al., 2002).

Researchers can never prove pre-existing equivalence between the treatment and comparison groups when the proxy measures have imperfect correlations with outcome. Given that we are unable to predict recidivism with complete accuracy, it would always be possible that any observed differences in recidivism could be attributed to undetected differences in risk potential at intake. Confidence in the equivalence of groups increases when: a) the methods of subject assignment minimize the probability of systematic differences; and b) the control variables account for a large proportion of the variance in the outcome.

A careful examination of pre-treatment equivalence requires thorough evaluation of the recidivism risk of each offender in the treatment and comparison groups. These risk assessments are most credible when conducted by competent evaluators (who are blind to outcome) using the best available approaches. Although there is some debate about how best to assess recidivism risk, most evaluators use one or more actuarial risk tools along with some consideration of factors external to the actuarial scheme. An alternate approach is to consider a list of empirically validated risk factors. Whatever method is used, it is important to show that the method used for assessing offender risk has validity in the research sample. The stronger the association with recidivism, the greater confidence that most of the relevant risk factors have been considered. The optimal control measures will vary with the outcome criteria used. The factors associated with sexual recidivism are not identical to the factors associated with violent or other criminal recidivism. Researchers should emphasize the control variables most closely associated with the outcome of interest.

Although there has been considerable progress in risk assessment for adult male sexual offenders, much less is known about risk assessment with female sexual offenders or juvenile sexual offenders. Consequently, researchers conducting treatment outcome studies with these populations would have little confidence in their choice of control variables.

As the equivalence of groups is critical to obtaining an unbiased estimation of the effectiveness of treatment, the CODC Guidelines have three items to assess this factor. The first item examines the a priori equivalence of groups based upon the study's design and procedures for subject assignment and evaluates if this introduces bias. A careful review of the procedures

used to recruit and retain participants in the two groups can provide an expectation of the equivalence of the groups. For example, with random assignment procedures we can expect there to be little, if any, a priori differences between the groups. As with all items assessing bias, there is a rating of the magnitude and the direction of bias. Given the importance of determining a priori equivalence of the groups, additional questions are provided to help with the ratings of random assignment, cohort, and risk band designs.

The second item addresses the adequacy of the search for pre-existing differences between groups and evaluates the coder's confidence that the search was adequate to identify important pre-existing differences. Relevant information includes the measures used and other information collected to evaluate the equivalence of the groups. This item does not take into account the results of the comparison; rather, it evaluates the tools and methods used to compare the groups. For example, irrespective of the results, more confidence can be drawn from groups compared on validated risk assessment instruments than on personality measures with little or no relationship with sexual recidivism risk.

The third item, findings on group equivalence, examines the results of this search for pre-existing differences, and evaluates the potential bias of the similarity/differences between groups. In this item, the coder examines the data specifically related to recidivism potential that were presented to demonstrate group equivalence. Whenever there is a difference between groups, it is this difference that can introduce bias and threaten the validity of the results. This item rates the data presented and not the statistical procedure(s) used to control for differences (a separate item assesses the use of statistical control procedures). Coders are required to assess the magnitude and direction of bias based on the group equivalence data presented.

13. A priori equivalence of groups

Concept: The general concept is the extent to which the treatment and comparison group would be expected to be equivalent based on the study's design and the procedures used to assign participants to each condition. Evidence of the equivalence of groups is not evaluated in this item (this is evaluated by a separate item). The factors to consider are the type of design and the assignment procedures (e.g., random assignment). Ideally, the assignment procedures provide expectations that the two groups would be equivalent. This occurs in well-implemented random allocation designs where the randomization process is one in which (a) the allocation sequence is unpredictable; and (b) those doing the actual allocation have no knowledge of the assignments made, known as allocation concealment (Altman et al., 2001). Schultz, Chalmers, Hayes, and Altman's (1995) meta-analytic review of randomized clinical trials found that studies with inadequate allocation concealment show significantly larger treatment effects than those with adequate concealment.

On the other hand, many studies do not employ random assignment. In these studies, it is important to compare the selection criteria for the treatment group and the criteria for inclusion in the comparison group. In this manner, one may evaluate the potential differences between the groups based strictly on the assignment/allocation procedures of the study.

Indicators: Potential indicators that would decrease bias are random assignment, allocation concealment, and equal probability that any one subject would be assigned to either condition. Potential indicators that would increase bias include situations in which subjects selected for one group had little to no chance of being in the other condition.

Rating	Description
0	*The design and allocation procedures are likely to result in offenders with different risk relevant characteristics being systematically placed in different groups.*
1	*The design and allocation procedures are likely to result in systematic differences between groups, but the characteristics on which they differ are not directly related to risk (e.g., language, location).* OR *It is possible that participants with certain risk relevant characteristics (e.g., risk, need) were more likely to be in one group than the other.*
2	*Due to the design and allocation procedures, all participants had an equal probability of being in either the treatment or comparison condition.*

To aid in the determination of a priori differences, separate checklists are provided for three types of studies: a) random assignment; b) risk band/norm; and c) cohort designs. These checklists are not intended to be separate items; instead, they highlight the special concerns associated with different forms of subject assignment. If evaluating one of these three designs, evaluators should complete the appropriate checklist prior to completing the rating for a priori equivalence of groups.

Randomized trial designs - Randomization procedures

Concept: Well-implemented random assignment reduces the possibility of pre-existing group differences. To the extent that there are problems with the randomization process, the likelihood of bias increases. Consequently, it is important to consider the integrity of the random assignment process.

Considerations: There are three specific components to examine. First, the randomization process must in fact be truly random, with each subject having an equal chance to be assigned to each condition. Alternate assignment, assignment based on day of the week, birth date or file identification number are processes that are not truly random. Such procedures are subject to experimenter influence and can alter the equivalence of groups. Second, the experimenter should be blind to group allocation at the point of assignment; otherwise group assignment may be biased. Examples of allocation concealment procedures are the use of sealed, sequentially numbered opaque envelopes or contacting a central office, which is unaware of subject characteristics, for group allocation. Third, the randomization procedures can break down or change during the course of the study due to unforeseen events (e.g., a change in policies, systemic pressures to treat certain individuals), resulting in pre-existing differences between groups.

Allocation is random?	
Allocation procedures resulted in predictable or anticipated group assignment such as alternating assignment, or by day of the week. The probability of subjects being assigned to each group was not random.	**0**
Allocation procedures described resulted in truly random assignment, such as use of random numbers.	**1**
Allocation concealment?	
Allocation was not concealed. The experimenter knew or could have anticipated group assignment and subconscious bias may have influenced subject assignment, eligibility or efforts to have a subject participate.	**0**
The procedures ensured that the allocation was concealed to the experimenter through the use of sealed, sequentially numbered, opaque envelopes, or a central office assigned allocation without knowledge of subject characteristics.	**1**
Consistent allocation procedures?	
The allocation procedures were altered during the course of the experiment.	**0**
The allocation procedures were followed consistently without change for the duration of the study.	**1**

Note: For random allocation designs, the above considerations will assist in evaluating a priori equivalency of groups.

Risk band/norm designs – Validity of norms

Concept: The general concept is the extent to which there are pre-existing differences between the treatment group and the norms or risk bands to which they are compared. With these designs, there is always an a priori expectation of some pre-existing group differences and, therefore, some bias is expected. It is impossible to demonstrate equivalence of groups on all factors.

Considerations: There are two issues to consider. The first issue is the accuracy of the assessment of the risk measure. A good measure should have high predictive accuracy with little or no missing data. It is critical that the assessment of risk for the treatment group is as complete as possible when comparing to risk bands or norms. Second, the equivalence of the normative population to the treatment sample should be examined for possible cohort effects or jurisdiction effects. For example, the recidivism rates of the normative group (i.e., comparison group) or the treatment group may be influenced by cohort or jurisdiction factors (e.g., comparing recent norms to a 1980s cohort of treated offenders).

Accuracy of assessment?	
The reliability of the assessment was not evaluated (or inadequate) and/or the assessments were incomplete due to missing information.	**0**
The actuarial risk assessment of the treatment group is completed with no missing information and high rater reliability was established.	**1**
Accuracy of norms?	
The norms (risk bands) have some empirical support (e.g., normative sample is based on less than 1000 offenders and/or the normative sample does not contain offenders from that country), but confidence in the norms is less than strong.	**0**
The norms (risk bands) have substantial empirical support, have been cross-validated, and the characteristics of the normative group are similar to that of the treatment group.	**1**
Cohort or jurisdiction effects?	
There are cohort differences and/or possible jurisdictional differences between the normative group and treatment group.	**0**
The treatment group is likely a good representation of the normative group in terms of cohort (time) and jurisdiction (geographical location).	**1**

Note: For risk band/norm designs, the above considerations will assist in evaluating a priori equivalence of groups.

Cohort designs – Cohort effects

Concept: The general concept is the influence of cohort on the results of the study. Because offenders are from different time periods, there could be systemic changes that occurred over time and/or changes in offender characteristics that affect re-offending rates. With these designs, researchers need to pay special attention to possible cohort effects.

Considerations: There are two specific considerations. First, the systemic cohort effects on recidivism must be checked. Examining whether variation in release date is associated with differences in recidivism (with the study sample and with the population from which they were drawn) can allow the researcher to assess systemic cohort effects. Second, possible cohort effects on the changing profile of offenders must also be examined. For example, it may be that the characteristics of one cohort are different from the characteristics of another. As it is difficult to assess all possible factors associated with recidivism, and because populations and systems change and evolve over time, it is critical that cohort designs examine possible changes in study populations, and when feasible, control for such systematic variation.

Cohort effects - System changes?	
The research did not evaluate systemic changes in the cohorts OR *Cohort effects were found after evaluating systemic changes in the study sample and/or the population from which they were drawn.*	**0**
The researcher evaluated systemic changes in the cohorts by examining both the study sample and the population from which they were drawn. Cohort effects were not found after a thorough search.	**1**
Cohort effects - Offender changes?	
The research did not evaluate changes in offender characteristics OR *Cohort effects were found after evaluating offender characteristics in the study sample and/or the population from which they were drawn.*	**0**
The researcher evaluated changes in offender characteristics in both the study sample and in the population from which they were drawn. Cohort effects were not found after a thorough search.	**1**

Note: For cohort designs, the above considerations will assist in evaluating a priori equivalence of groups. How the researcher addresses cohort effects should be considered in the item effectiveness of statistical procedures to control bias.

13. A priori equivalence of groups

This item is concerned with a priori expectations of the equivalence of groups based on design characteristics and subject selection criteria. For example, there is little expectation of a priori differences in large random assignment studies. However, if the criteria for inclusion in one group are different from the other group(s), there are some a priori expectations that the groups are different on variables related to outcome. For this item, it is the expectations of group equivalence, not the empirical evidence that is evaluated.

Information on a priori equivalence of groups extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: A priori equivalence of groups			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

14. Adequacy of search for pre-existing differences

Concept: The general concept is how and with what measures did the investigator(s) attempt to evaluate the equivalence of the groups prior to treatment. In other words, how adequate a job did the investigator(s) do to examine the extent to which pre-existing group differences could account for observed differences in outcome? The goal of this item is to determine the coder's confidence that the author(s)' approach, method, tools and/or information used to compare pre-existing differences were sufficient to identify possible and important differences.

Indicators: Potential indicators that would increase confidence that the search was sufficient would include the use of validated risk assessment instruments, information gathered on other factors related to recidivism, the reliability of the information and the amount of missing information. This item assesses confidence in the information gathered to examine pre-existing group differences, not the actual results of this comparison (which is done in the next item).

Rating	Description
0	*No validated risk assessment instrument was used. Groups may have been assessed on many factors but few, if any, were related to recidivism potential. There was a significant amount of information missing.*
1	*Groups were assessed on one validated risk measure or on a substantial number of factors related to recidivism potential that demonstrated at least moderate predictive accuracy.*
2	*Groups were assessed on at least one validated risk measure, on additional information gathered on other factors related to recidivism potential external to the risk measures, and reliability of the information is demonstrated. There is little, if any, missing information.*

14. Adequacy of search for pre-existing differences

This item examines how the author(s) attempted to evaluate pre-existing differences between the groups. The assessor is attempting to determine the level of confidence that these measures, information and/or assessed factors could accurately detect potential pre-existing differences. The reliability and the amount of missing information influence confidence. This item assesses confidence in the information gathered to examine pre-existing group differences, and not the actual results of this comparison (which is rated by the next item).

Information on search for pre-existing differences extracted from study			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence that the procedures, measures and/or information used are empirically linked to outcome variable*	**1** *Some confidence that the search for pre-existing differences was thorough, reliable and valid*	**2** *High confidence that the search for pre-existing differences was thorough, reliable and valid*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Adequacy of search for pre-existing difference		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the procedures, measures, and/or information used are empirically linked to outcome variable*	**1** *Some confidence that the search for pre-existing differences was thorough, reliable and valid*	**2** *High confidence that the search for pre-existing differences was thorough, reliable and valid*
Reason(s) for rating		

15. Findings on group equivalence

Concept: The general concept is the extent to which the data support the equivalence of the treatment and comparison groups. The goal of this item is to assess the magnitude and direction of bias based on the pre-existing characteristics of the groups. In this item, it is the evidence that is evaluated, regardless of how the investigators analyzed their data. To rate this item, the study must have presented data on the two groups that are related to recidivism potential. If the information is insufficient to determine the extent of equivalence between the groups, this item is not rated. In general, a confidence rating of zero on the previous item implies that this item cannot be rated. However, if there are clear differences on risk relevant variables, this item can be rated even when the search for pre-existing differences was minimal.

Indicators: Indicators are based on the magnitude of the differences between the treatment and comparison group on risk potential. The groups are considered "essentially equal" when the probability level of the observed differences is greater than .40. The groups are "significantly different" when the probability of the observed differences is less than .05 or the effect size between groups is moderately large (i.e., $d \geq .50$). When the two groups are essentially equal on the risk measures and on the other information related to recidivism potential, there is an expectation that pre-existing differences would result in a negligible bias on the effect size. On the other hand, bias would be expected when the groups are significantly different on the risk measure(s) or if there are statistically significant differences on variables predictive of recidivism.

Rating	Description
0	*The groups are significantly different on the risk measure(s) ($p < .05$)* OR *There is no risk measure and there are statistically significant differences on some of the variables related to recidivism potential.*
1	*The groups are not significantly different on risk measure(s), and there are group differences on variables that are weakly or indirectly related to risk.* OR *The groups are not significantly different on risk measure(s), but there is some uncertainty in the results due to small sample size or marginal statistical significance ($.05 < p < .40$).* OR *A risk measure was not used, and there are no significant differences on a range of validated risk factors.*
2	*The groups are essentially equal on the risk measure(s) ($p \geq .40$).* AND *The groups are essentially equal on almost all of the other variables external to the risk measure(s).*

15. Findings on group equivalence

This item is concerned with evidence presented in the study that attempts to demonstrate the equivalence of groups. The assessor must evaluate these results to determine the equivalence of the groups and the extent to which the pre-existing characteristics of the groups may introduce bias in the estimation of treatment effectiveness.

Information and results regarding the equivalence of groups extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Date on group equivalence			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

VI. Outcome variables

The outcome variables are of critical importance in the evaluation of treatment effectiveness. All sexual offender treatment outcome studies should evaluate sexual recidivism, but it is also valuable to examine the broader category of "serious" recidivism, which includes both sexual and violent offences. Violent offences are of significant public concern and it is not uncommon for sexual crimes to result in convictions for non-sexual violent offences (e.g., forcible confinement). We also recommend that evaluations measure general (any) recidivism.

There are three aspects of recidivism information that need to be considered: a) the adequacy of the length of follow-up for detecting recidivism; b) the information sources that provide data on the outcome variable(s); and c) the equivalence in follow-up information between the treatment and comparison groups. The length of follow-up and sources of recidivism information play a key role in the degree of confidence placed in the results. Differences in the recidivism information introduce bias.

One reason that sex offender treatment outcome studies are difficult to conduct is that the ultimate outcome of interest – sexual recidivism – is a low frequency event. The observed sexual offence recidivism rates are approximately 10% to 15% after five years and 20% after 10 years (Harris & Hanson, 2004). In addition, there is often substantial lag time between the commission of the offence and the time that the offence is detected (if ever) and recorded in databases available to researchers. Most researchers propose that there should be minimum follow-up periods before attempting to assess sexual offence recidivism rates. The recidivism detected in very short follow-up periods may have more to do with arbitrary features of the criminal justice system than with the characteristics of the offenders or the treatment that they have received. There is no clear agreement on the minimum follow-up period required for a credible study, but figures such as 3-5 years have been proposed.

Low base rates present less of a problem when researchers are interested in broader outcome criteria than only sexual offence recidivism. If any recidivism were the primary outcome criterion, it would be possible to obtain an adequate sample of recidivists using smaller sizes and shorter timeframes than those proposed in these Guidelines.

The validity and reliability of the recidivism information also plays a role in the level of confidence in the study's results. A considerable amount of confidence can be placed in the results when the source of the information is credible (e.g., official records), the search is thorough (e.g., no cases are missed during the collection of the data), and when the study uses multiple sources of information (e.g., national and state criminal records, police reports, and probation files). The more valid and reliable the source(s) of recidivism information, the more confidence in the results.

It is important that the follow-up time is the same for the treatment and comparison groups. Such differences would be expected in cohort designs, but all studies need to consider the equivalence of the follow-up periods. Statistical control procedures employed due to inequality of follow-up time have limited utility. If the differences in the follow-up times are significant (e.g., $p < .05$), then researchers have an obvious problem. However, when the sample sizes are small, the differences would have to be very large to be detected. It is prudent for all studies to control for potential differences in follow-up times using either a) fixed follow-up times (e.g., five years) or b) survival analysis. Researchers should note, however, that survival analysis might still introduce bias when the proportional hazard assumption is not met or the shape of the survival curve changes over time.

The CODC Guidelines evaluate three aspects of outcome data. One item assesses confidence in the length of follow-up to detect sexual offence recidivism. Relevant information includes the length and the range of follow-up for each group. Confidence refers to the adequacy of the follow-up length to detect recidivism.

The second item assesses confidence in the validity and reliability of the recidivism information. Relevant information includes describing the thoroughness of records, whether multiple sources of information were used to detect recidivism, and the credibility of the sources of information. It is also desirable to check how many offenders have been lost to follow-up due to prolonged incapacitation, deportation or death. Confidence refers to the validity and reliability of the recidivism data.

The third and final item of this section assesses the bias that may be introduced by the equivalence of follow-up. Relevant information includes any difference in the length or range of follow-up, the sources of the recidivism information, or other important differences between the groups in the gathering of recidivism information. Ratings of the magnitude and the direction of bias are required.

16. Length of follow-up

Concept: The general concept is whether the length of follow-up for the study sample is sufficiently long to detect recidivism. The goal of this item is to assess confidence that the length of follow-up is adequate to provide accurate recidivism information.

Indicators: This item requires data on the average follow-up period for the study sample. Confidence increases as the average follow-up period increases. The follow-up times were based on using sexual offence recidivism or violent offence recidivism as the outcome criterion. If more frequent forms of recidivism are used (e.g., any new conviction, parole violations), it would be possible to obtain a sufficient number of recidivists with shorter follow-up times. Consequently, studies using the most common forms of recidivism can receive a rating of "2- high confidence" based on follow-up times of less than five years.

Rating	Description
0	Average follow-up period less than three years (36 months) for the study sample.
1	*Average follow-up period of three years (36 months) but less than five years (60 months).*
2	*Average follow-up period of five years (60 months) or more.*

16. Length of follow-up

This item is concerned with the length of follow-up and the level of confidence that it is adequately long to detect recidivism. Information about the length of follow-up of the entire study sample and of each group should be recorded.

Length of follow-up information extracted from study	*Average (range) follow-up*
Treatment group	
Comparison group	
Study sample	

Confidence rating

--	**0**	**1**	**2**
Insufficient information to evaluate	*Little confidence that the length of follow-up is sufficiently long to detect recidivism*	*Some confidence that the length of follow-up is sufficiently long to detect recidivism*	*High confidence that the length of follow-up is sufficiently long to detect recidivism*

Reason(s) for rating:

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Length of follow-up		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the length of follow-up is sufficiently long to detect recidivism*	**1** *Some confidence that the length of follow-up is sufficiently long to detect recidivism*	**2** *High confidence that the length of follow-up is sufficiently long to detect recidivism*
Reason(s) for rating		

17. Validity and reliability of recidivism information

Concept: Relevant information includes describing the thoroughness of records, whether multiple sources of information were used to detect recidivism, and the credibility of the sources of information. Confidence refers to the validity and reliability of the recidivism data.

Indicators: Confidence is increased when the recidivism information is gathered from official records (national or local). When official sources are supplemented with information from other credible sources, confidence is increased. Examples of credible sources of recidivism information would include police, child welfare or child protection agencies.

Rating	Description
0	*Recidivism information is self-reported only or from institutional records that would pertain to a small select sample of recidivists (e.g., readmission to a specific group home or halfway house).*
1	*Single source of recidivism information, either national or local records or information from other credible sources (e.g., child welfare, police investigations)*
2	*Multiple sources of recidivism information. One source must be official criminal history records (either national or local). The second source may be from another credible source (e.g., child welfare, police investigations)*

17. Validity and reliability of recidivism information

This item is concerned with the quality of the recidivism information and the level of confidence that this information is accurate. Record all source(s) of recidivism information.

Sources of recidivism information extracted from study			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence that the sources of follow-up information are valid and reliable indicators of recidivism*	**1** *Some confidence that the sources of follow-up information are valid and reliable indicators of recidivism*	**2** *High confidence that the sources of follow-up information are valid and reliable indicators of recidivism*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Validity and reliability of recidivism information		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the sources of follow-up information are valid and reliable indicators of recidivism*	**1** *Some confidence that the sources of follow-up information are valid and reliable indicators of recidivism*	**2** *High confidence that the sources of follow-up information are valid and reliable indicators of recidivism*
Reason(s) for rating		

18. Equivalence of follow-up

Concept: The general concept is the extent to which recidivists have an equal likelihood of being detected in the treatment and comparison groups. Factors to consider are the sources of recidivism information and the length of follow-up. Differences in sources of information or follow-up time would introduce bias.

Indicators: Potential indicators that would increase bias are significantly different follow-up periods between groups, or if the sources of recidivism information were different. Bias is decreased when recidivism information was obtained from the same sources and the length of the follow-up period was the same for each group (both in terms of central tendency and variability). The least bias is expected when the follow-up period is fixed for all participants.

Risk band/norm designs: These designs would usually receive a rating of "0 – considerable bias" because the outcome criteria used to create the norms would typically be different from the outcome criteria used to detect recidivism in the treatment group. It could be other than zero if both the study and the norms used the same outcome criteria in the same jurisdiction.

Rating	Description
0	*The sources of information for each group are different.* *OR* *The length of follow-up between groups is different and these differences would be expected to have a substantial impact on overall recidivism rates (e.g., three years versus five years) and survival analysis was not used.*
1	*The sources of information for each group are identical.* *AND* *The length of follow-up between groups is different but <u>either</u>* *a) These differences would not be expected to make an impact on overall recidivism rates (e.g., 21 years versus 23 years)* *OR* *b) The length of follow-up between groups is substantially different and survival analysis was used.*
2	*Follow-up length for each group is fixed and equivalent* *AND* *The sources of recidivism information for each group are identical.*

18. Equivalence of follow-up

This item is concerned with bias introduced by a lack of equivalence in the recidivism information (length, sources of information). If, for example, the treatment group has a shorter follow-up period than the comparison group, this would introduce bias favouring the treatment group. Another example is when follow-up information is gathered from state records for the comparison group and self-report was used for the treatment group. These differences in follow-up sources can result in bias.

Information and results regarding the equivalence of follow-up information extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Equivalence of follow-up			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

VII. Correct comparisons conducted

In every study, researchers must make a number of decisions when calculating the overall effect of treatment. For example, the researcher must choose which one of multiple outcome criteria will be used (e.g., arrests or convictions, the length of the follow-up) and which subjects will be included in the analysis. They must also decide which variables to employ if using statistical procedures to control for group/individual differences in recidivism potential, as well as the specific options employed (e.g., replace with mean values, or use pair wise or list wise deletion of cases with missing values). Often these decisions cannot be fully determined in advance due to unpredictable factors such as attrition and missing information; nevertheless, these decisions can affect confidence in the results and potentially introduce bias beyond that introduced by the factors previously described.

As the CODC Guidelines define high quality studies as those with the least amount of bias and the greatest degree of confidence, there are three items that examine the treatment effect calculation. The first item assesses data dredging, which refers to exploratory analyses of a data set with little a priori rationale. Such exploratory post hoc analyses enhance the probability of Type 1 error (incorrect conclusion that there is an effect). Effect size(s) based on such "data dredging" influences the confidence in the reported effect size.

The second item assesses the effectiveness of statistical procedures to control bias. Because differences between the treatment and comparison groups are common, even in designs that aim to minimize the possibility of such differences (e.g., random assignment), researchers attempt to control for any observed differences in the calculation of the effect size through statistical procedures. Many of these methods are statistically sophisticated and require decisions regarding specific options (e.g., handling of missing data). There does not appear to be a consensus on the best approach nor how effective these approaches are to limit bias in samples that have considerable pre-existing differences. This item assesses confidence in these procedures when estimating treatment effectiveness. That is, this item assesses the level of confidence in the statistical procedures used to minimize any bias inherent in the study (e.g., design weaknesses, subject selection factors, pre-existing group differences, and differences in follow-up).

Finally, the third item, called computation of least bias comparison, examines the potential bias in the overall effect size calculated and presented by the researcher(s). The effect size of treatment is determined by which group of "treated" offenders is compared to which "comparison" group. Within each study, it is possible to calculate an effect of treatment based on subjects that would introduce the least amount of bias possible with a reasonable amount of confidence (i.e., sample size). To do so, one must consider a number of factors that have been described earlier such as subject attrition (e.g., refusers, dropouts), sample size, and validity and reliability of the follow-up information. The comparison with the least amount of bias and most confidence would generally involve a comparison with the largest number of subjects who were assigned to each condition with the most thorough, valid and reliable information.

19. Data dredging

Concept: The general concept is whether the comparisons reported are based on a priori theoretical reasoning supported by research, and whether the investigation was specifically designed to evaluate/measure these comparisons. Another factor to consider is the overall number of comparisons, particularly the number of post-hoc comparisons as the probability of a Type 1 error increases with increased numbers of comparisons. The bias introduced by multiple comparisons cannot be addressed by using conservative p values (e.g., $p < .001$) because the data of interest are the group differences, not the statistical significance of these differences. The goal of this item is to assess the coder's confidence that the comparisons were planned a priori and not the result of "data dredging".

Indicators: Potential indicators that would decrease confidence include a large number of comparisons, analyses that are post hoc driven, or where little a priori rationale is provided for conducting these comparisons. Potential indicators that would increase confidence include studies in which few comparisons were conducted, the analyses were determined a priori, and these comparisons are reflected in other inherent features of the study, such as treatment goals, treatment methods or subject selection criteria.

Rating	Description
0	*Effect size(s) was one of a large number of post hoc comparisons and those comparisons were not planned.*
1	*It is unclear if the primary comparisons were planned a priori or post hoc. These comparisons appear reasonable, but it is possible that they were developed post hoc.*
2	*Comparisons (effect sizes) were planned a priori. The rationale for these comparisons is reflected in the study's design (e.g., subject selection factors or treatment goals). The number of comparisons required to test the hypotheses were minimal.*

19. Data dredging

This item is concerned with the a priori planning and the number of comparisons conducted to measure effectiveness of treatment, and the level of confidence that the estimation of treatment effectiveness is not due to chance. Record all comparisons/effect sizes.

Estimation of treatment effectiveness and mediators extracted from study			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence in the results as the comparisons were developed post hoc*	**1** *Some confidence in the results as the comparisons may have been developed post hoc*	**2** *High confidence in the results as the comparisons were planned a priori*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Data dredging		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence in the results as the comparisons were developed post hoc*	**1** *Some confidence in the results as the comparisons may have been developed post hoc*	**2** *High confidence in the results as the comparisons were planned a priori*
Reason(s) for rating		

20. Effectiveness of statistical procedures to control bias

Concept: The general concept is whether the researchers employed statistics to control for pre-existing differences between groups, and if these procedures would be effective. There are three factors to consider. One factor is the equivalence of groups. When pre-existing differences are large (expected or demonstrated), the confidence in the effectiveness of statistical control procedures to control bias is significantly reduced. When the differences are minimal, confidence that statistical control procedures can effectively control bias increases. The second factor to consider is the quality of the statistical control variables. The quality of the statistical control variables refers to their reliability and predictive relationship to outcome. When control variables are not risk relevant, confidence in statistical control analyses to control bias is low. On the other hand, when the control variables are reliable and are significantly related to outcome, confidence increases. The third factor to consider is the credibility of the statistical procedures employed. There are a number of different statistical analyses that could be employed, and the researcher must choose one. For example, some statistical procedures require normal distribution of scores yet it may be clear that the scores are not normally distributed.

Indicators: Three previously scored items provide information that can assist and guide the rating of this item. When considering group differences, the coder can review the scores on a priori expectation equivalence of groups (item 13) and the findings on group equivalence (item 15). For the quality of the control variables used in the analyses, item 14 (adequacy of search for pre-existing differences) can provide useful information. In this item, the coder must consider these factors as well as the statistical procedures and control variables used. There is little confidence when no statistical control procedures were used, when there are large differences between groups, either expected a priori or statistically demonstrated, or when the control variables were not risk relevant (e.g., ethnicity, language, location). Note, however, that in alternate treatment designs, researchers would want to also control for variables that would be related to offenders' capacity to benefit from the interventions given (i.e., responsivity variables), even when these variables may have no direct relationship to risk (e.g., language).

Rating	Description
0	*There were no controls for risk in the initial design (e.g., risk-based matching, random assignment) and post hoc statistical controls were not used* *OR* *Statistical control procedures used but were likely insufficient to control bias due to large group differences, or the statistical control variables were inadequate.*
1	*Controls for risk were inherent in the initial design (e.g., risk-based matching, random assignment) and post hoc statistical controls were not used.* *OR* *Statistical control procedures employed and plausibly effective, but there is some uncertainty about their ability to control bias.*
2	*Statistical control procedures employed and effective as there were minimal group differences, and the control variables used were adequate to control potential bias.*

20. Effectiveness of statistical procedures to control bias

This item is concerned with the statistical methods used to control for pre-existing differences when estimating the effectiveness of treatment and the level of confidence in these statistical procedures. Note the statistical methods used and the variables for which the analyses controlled.

Statistical procedures used in the study to estimate effects of treatment			
Confidence rating			
-- *Insufficient information to evaluate*	**0** *Little confidence that the statistical procedures were adequate to control for group differences*	**1** *Some confidence that the statistical procedures may have been adequate to control for group differences*	**2** *High confidence that the statistical procedures were adequate to control for group differences*
Reason(s) for rating:			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Effectiveness of statistical procedures to control bias		
What additional information is desired and why?		
If new information was obtained, specify		
Revised confidence rating		
0 *Little confidence that the statistical procedures were adequate to control for group differences*	**1** *Some confidence that the statistical procedures may have been adequate to control for group differences*	**2** *High confidence that the statistical procedures were adequate to control for group differences*
Reason(s) for rating		

21. Computation of least bias comparison

Concept: The general concept is whether the researcher chooses to base the effect of treatment on the subjects and outcome variable(s) that would least likely introduce bias and result in the greatest confidence. Often, when calculating the effect of treatment, a researcher must make a number of decisions regarding which subjects are to be included and which of multiple outcomes to use. These decisions should consider all of the previous factors that can introduce bias (e.g., attrition, group equivalence, outcome measures) and alter confidence in the results (e.g., sample size). It is the researcher's goal to generate the estimate of treatment effectiveness that has the least bias, and which garners the greatest confidence.

Indicators: Potential indicators are all the previous items such as missing data, sample size, attrition, etc., as well as other pertinent factors such as the composition of the comparison group, how the researchers addressed missing data, and what options are employed during statistical analyses.

Rating	Description
0	*The researchers included and/or excluded certain subjects from the analyses that introduces bias in the estimation of treatment effectiveness* *OR* *The researchers chose an outcome measure that is reasonably expected to introduce bias.*
1	*The comparison selected was plausible, but there was some uncertainty about whether the optimal comparison/statistical analysis was selected.*
2	*The researchers, using the most complete, valid and reliable outcome measure, included as many subjects assigned to each condition as possible when calculating the effects of treatment.*

21. Computation of least bias comparison

This item is concerned with the groups (and the subjects within each group used) and the measures used to compute the effectiveness of treatment and any potential bias this places on the results. In most cases, there are a number of potential comparisons. Identify the comparison used by the researchers as well as your assessment of the least biased and most reliable comparison available.

Information and results regarding the effectiveness of treatment extracted from study			
Bias rating			
-- *Insufficient information to evaluate*	**0** *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*
Direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

Is it worth looking for additional information or new analyses? If yes, complete next page.

Additional information: Computation of least bias comparison			
What additional information is desired and why?			
If new information was obtained, specify			
Revised bias rating			
0 *Introduces a considerable amount of bias*	**1** *Some bias likely introduced in the results*	**2** *An expectation of negligible bias in the results*	
Revised direction of bias			
? *Cannot assess the direction of bias*	**+1** *Bias likely increases the magnitude of treatment effectiveness*	**0** *No bias expected*	**-1** *Bias likely decreases the magnitude of treatment effectiveness*
Reason(s) for rating			

VIII. Global rating

The global rating is an overall summary of study quality. The global rating is a form of structured judgement, based on the individual items and summary ratings of confidence and bias. To assist in the global rating, it is useful to clearly identify the main findings of the study (effect sizes, sample sizes). We also recommend that coders review the reasons for their decisions as they transfer their scores on the individual items to the summary sheet. Coders then assess their level of confidence in the study based upon how well potential threats to internal validity were assessed and addressed. Next, coders consider the magnitude and direction of bias that is reflected in the main findings (effect sizes). Although the summary confidence and bias ratings should correlate with ratings on individual items, we have not proposed a specific algorithm to translate the individual ratings to summary ratings. Once the summary rating of bias and confidence are completed, however, explicit directions are provided to produce a four-category global rating of study quality: strong, good, weak, and rejected.

a) STRONG: High confidence that the study has minimal bias in estimating the effectiveness of sexual offender treatment. It is well-designed and well-executed with convincing results. The study may have minor problems, but these problems are unlikely to influence the main conclusions or to change the direction of the observed effects.

b) GOOD: High confidence that the study has no more than a small amount of bias (intermediate rating). Reasonable efforts have been made to address threats to validity, but much remains unknown.

c) WEAK: Some confidence that the study has no more than a small amount of bias. The study has significant flaws, but is of possible relevance to the question of treatment effectiveness. Weak evidence at best.

d) REJECTED: Low confidence in the results, *or* considerable bias. The study has multiple significant flaws. The procedures used would be expected to introduce considerable bias, or the study lacks important information required to eliminate plausible alternate explanations for the findings.

GLOBAL RATING

EFFECT SIZE	
0 *Cannot calculate effect size: No further ratings required.*	**1** *Effect size calculated*
Least bias estimate of effect size extracted from study.	
Details of effect size (i.e., description of subjects & treatment, data used to calculate).	

Confidence

Rate confidence in the study's internal validity. Confidence refers to the validity and reliability of the effect size as a measure of this study's treatment (implemented and provided to the sample as described) impact on the outcome measure employed.

0 **Little or no confidence**	**1** **Some confidence**	**2** **High confidence**
Reason(s) for rating		

Bias

Rate the amount and direction of bias inherent in the effect size calculated above. Take into account the design and specific factors in the study. Provide reasons for the rating.

0	***Considerable Bias***	***1***	***Some Bias***	
-	*Decreasing effect of treatment*	**-**	*Decreasing effect of treatment*	**2** *No bias expected (direction: 0)*
+	*Increasing effect of treatment*	**+**	*Increasing effect of treatment*	
?	*Unknown direction on results*	**?**	*Unknown direction on results*	
Reason(s) for ratings				

GLOBAL STUDY QUALITY RATING

0 **Rejected**	**1**	**2** **Good**	**3** **Strong**
*Confidence Rating of **0*** ***OR*** *Bias Rating of **–0** or **+0** or **?0***	*Confidence Rating of **1*** AND *Bias Rating of **–1** or **+1** or **?1***	*Confidence of **1** and Bias of **2*** *OR* *Confidence of **2** and Bias of **–1, +1,** or **?1***	*Confidence Rating of **2*** ***AND*** *Bias Rating of **2***

Study Quality Rating Guide Summary Sheet (1 of 2)

Confidence *Bias*	*Little confidence* **0** *Considerable bias*	*Some confidence* **1** *Some bias*	*High confidence* **2** *Negligible bias*
Direction of Bias	**+** *Increases Treatment*	**-** *Decreases Treatment*	**?** ***Unknown Direction***

Administrative Control of Independent Variables

1. Defining Treatment	***Confidence***		
2. Defining Comparison	***Confidence***		
3. Miscellaneous Incidental Factors		***Bias***	***Direction***

Experimenter Expectancies

4. Experimenter Involvement		***Bias***	***Direction***
5. Blinding in Data Management		***Bias***	***Direction***

Sample Size

6. Sample Size of Treatment	*Confidence*		
7. Sample Size of Comparison	***Confidence***		
8. Cross Institution Design Sample Size	***Confidence***		

Attrition

9. Subject Selection		***Bias***	***Direction***
10. Program Attrition		*Bias*	*Direction*
11. Intent-to-treat		*Bias*	*Direction*
12. Attrition in Follow-up		*Bias*	*Direction*

Study Quality Rating Guide Summary Sheet (2 of 2)

Equivalence of Groups

13. A Priori Equivalence of Groups		***Bias***	***Direction***
14. Adequacy of Search for Differences	*Confidence*		
15. Findings on Group Equivalence		*Bias*	*Direction*

Outcome Variables

16. Length of Follow-up	*Confidence*		
17. Validity/Reliability of Recidivism	*Confidence*		
18. Equivalence of Follow-up		*Bias*	*Direction*

Correct Comparisons Conducted

19. Data Dredging	*Confidence*		
20. Effectiveness of Statistical Controls	*Confidence*		
21. Compute Least Bias Comparison		*Bias*	*Direction*

GLOBAL RATING

Effect Size and N's			
	Global Confidence		
(**0** = Little/No,	**1** = Some,	**2** = High)	
	Global Quantity of Bias		
(**0** = Considerable,	**1** = Some,	**2** = Negligible)	
	Global Direction of Bias		
(**?** = Unknown,	**+** = Increases Rx	**-** = Decreases Rx)	
Global Rating			
(0 = Rejected, 1 = Weak, 2 = Good, 3 = Strong)			

References

Altman, D. G., Schultz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzche, P. C., & Lang, T., for the CONSORT Group (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663-694.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collaborative Outcome Data Committee. (2007). Guidelines for the Evaluation of Sexual Offender Treatment Outcome Research (CODC Guidelines), Part 1: Introduction and Overview. *User Report 2007-02. Ottawa: Public Safety and Emergency Preparedness Canada.*

Gendreau, P., Goggin, C., & Smith, P. (1999). The forgotten issue in effective correctional treatment : Program implementation. *International Journal of Offender Therapy and Comparative Criminology, 43*, 180-187.

Hanson, R. K., Gordon, A., Harris, A. J. R., Marques, J. K., Murphy, W., Quinsey, V. L., & Seto, M. C. (2002). First report of the Collaborative Outcome Data Project on the effectiveness of psychological treatment of sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 14,* 169-194.

Hanson, R.K., & Nicholaichuk, T. (2000). A cautionary note regarding Nicholaichuk et al. (2000). *Sexual Abuse: Journal of Research and Treatment, 12*, 289-293.

Harris, A.J.R., & Hanson, R.K. (2004). Sex offender recidivism : A simple question. Corrections Users Report No. 2004-03: Public Safety and Emergency Preparedness Canada.

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*, 1054-1060.

Schultz, K.F., Chalmers, I., Hayes, R.J., & Altman, D. J. (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association, 273,* 408-412.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Thomas, H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing,* 2, 91-99.

Westen, D., Novotny, C.M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings and reporting in controlled clinical trials. *Psychological Bulletin, 130,* 631-663.

Zaza, S., Wright-De Agüero, L.K., Briss, P.A., Truman, B.I., Hopkins, D.P., Hennessy, M.H., et al. (2000). Data collection instrument and procedure for systematic reviews in the *Guide to Community Preventive Services*. *American Journal of Preventive Medicine, 18*, 44-74.

Committee Members

Anthony R. Beech, Ph.D., is a Professor in Criminological Psychology in the Centre for Forensic and Family Psychology, School of Psychology, University of Birmingham, U.K., and a Fellow of the British Psychological Society. a.r.beech@bham.ac.uk

Guy Bourgon, Ph.D., is a Research Officer with Public Safety Canada and Adjunct Research Professor in the Psychology Department of Carleton University, Ottawa, Canada. Guy.Bourgon@ps.gc.ca

R. Karl Hanson, Ph.D., is a Senior Research Officer with Public Safety Canada and Adjunct Professor in the Psychology Department of Carleton University, Ottawa, Canada. Karl.Hanson@ps.gc.ca

Andrew J. R. Harris, Ph.D., is Senior Research Manager, Research Branch, Correctional Service of Canada. He does not profess. HarrisAJ@csc-scc.gc.ca

Calvin M. Langton, Ph.D., is an Assistant Professor in the Department of Psychiatry, University of Toronto, Canada, and Honourary Research Fellow in the School of Community Health Sciences, University of Nottingham, UK. calvin.langton@utoronto.ca

Janice Marques, Ph.D., is a consulting psychologist who recently retired from the California Department of Mental Health. She was President of ATSA when this collaborative project was launched in 1998. jkmarques@sbcglobal.net

Michael H. Miner, Ph.D., is an Associate Professor at the Program in Human Sexuality, Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, MN. miner001@umn.edu

William Murphy, Ph.D., is a Professor in the Department of Psychiatry, University of Tennessee Health Science Center, Memphis, Tennessee. wmurphy@utmem.edu

Michael Seto, Ph.D., is a psychologist in the Law and Mental Health Program, Centre for Addiction and Mental Health, and an Associate Professor in the Department of Psychiatry and Centre of Criminology at the University of Toronto. Michael_Seto@camh.net.

Vernon Quinsey, is Professor and Head of Psychology at Queen's University, Kingston, Ontario. vernon.quinsey@queensu.ca

David Thornton, Ph.D., is the treatment director at Sand Ridge Secure Treatment Center, Mauston, WI. thorndm@dhfs.state.wi.us

Pamela M. Yates, Ph.D. is a psychologist with the Correctional Service Canada and specializes in the treatment of sexual offenders. YatesPM@csc-scc.gc.ca