Sexual Offender Treatment Outcome Research: CODC Guidelines for Evaluation

Part 1: Introduction and Overview

2007-02

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.: PS4-38/1-2007E-PDF
ISBN No.: 978-0-662-45553-0

Preface

This document is intended for those seriously interested in research on the effectiveness of treatment for sexual offenders. As such, it is addressed to three related readerships: a) reviewers wishing to critically examine existing sexual offender treatment outcome studies, including journal editors, scientific reviewers and meta-analysts; b) program evaluators wishing to determine the level of benefit of a particular treatment program; and c) researchers designing new studies to evaluate the effectiveness of treatment for sexual and other types of offenders.

After reviewing existing scales for rating study quality, the committee decided that a new rating scale was required. The previous scales were poorly suited to the types of designs commonly used in sexual offender research. As well, most of the previous scales are heterogeneous, containing items related to reporting quality, ethical issues, and data interpretation rather than bias or internal validity.

This document focuses on sexual offenders, but much of the discussion is also relevant to the evaluation of other extended, complex behavioural interventions where failure results in harm to others, where failure is not observed in treatment, and where failure may not even be noticed until years afterwards (e.g., domestic violence, impaired driving).

This document is a group effort. The Collaborative Outcome Data Committee was formed with the goal of advancing outcome research on sexual offenders. Members of the committee were selected based on their expertise in sexual offender research evaluation, and their ability to bring distinct perspectives. Individual members have voiced divergent opinions concerning the effectiveness of treatment, but the aim of this project was to establish common ground. Complete agreement was neither expected nor desired; instead, we hoped to articulate the common assumptions concerning the characteristics of credible and less credible studies. Specifically, we present a scheme for rating the quality of sexual offender treatment outcome studies that is plausible, reliable (independent raters agree on how to classify studies) and widely accepted by leaders in the field.

Since 1980, there have been more than 20 reviews of the effectiveness of treatment for sexual offenders. Although the treatment groups, on average, show lower recidivism rates than the comparison groups, all reviews have noted problems with the available studies, thereby limiting any strong conclusions. If the Committee's aspirations are fulfilled, future researchers will be able to present conclusions with increasing confidence and precision, and future reviewers will be better able to evaluate treatment outcome research.

Table of contents

Introduction

Does sexual offender treatment work? Although considerable research has addressed this question, experts continue to debate the effectiveness of interventions intended to reduce the recidivism risk of sexual offenders.

Furby, Weinrott and Blackshaw (1989) in an early, influential review, concluded that there was no evidence that treatment reduced recidivism rates. Their findings contrasted with a contemporaneous review published by the Solicitor General Canada (Sex Offender Treatment Review Working Group, 1990), which concluded that "treatment can be effective in reducing recidivism from about 25% to 10-15%" (p. 19).

The basic positions have changed little in the last 15 years. Hall (1995) conducted a meta-analysis of 12 studies published after Furby et al.'s (1989) review, and found overall positive treatment effects for cognitive-behavioural treatment and hormonal treatments. Hall's (1995) review, in turn, was criticized because it included studies that were insufficiently rigorous (Harris, Rice & Quinsey, 1998). Gallagher, Wilson, Hirschfield, Coggeshall and Mackenzie (1999) conducted an updated meta-analysis of 22 studies, in which they attempted to include only the best available studies. Nevertheless, they included some of the "flawed" studies criticized by Harris et al. (1998), as well as preliminary versions of several studies whose results changed upon further analysis.

The Collaborative Outcome Data Committee was formed in 1997 with the goals of organizing the existing sexual offender outcome studies and promoting high quality evaluations. The committee's first report – a meta-analysis of 43 studies (Hanson et al., 2002) concluded that there was a small positive effect for current treatments, but that firm conclusions awaited more and better research. The findings of the Collaborative Outcome Data Committee were subsequently replicated in a larger review by Lösel and Schmucker (2005), who also found a significant treatment effect for cognitive-behavioural treatment.

Not surprisingly, these meta-analyses have attracted criticism. Rice and Harris' (2003) response to the Hanson et al. (2002) report was that observed results could most easily be explained by potential biases in subject assignment to treatment and comparison groups. In the "best" studies identified by Rice and Harris (2003), there was no overall treatment effect. Similarly, a review of nine sexual offender treatment outcome studies conducted for the Cochrane Collaboration found no effect for treatment (Kenworthy, Adams, Bilby, Brooks-Gordon & Fenton, 2004; Brooks-Gordon, Bilby & Wells, 2006). The Kenworthy et al. (2004) study is noteworthy because it only included studies meeting criteria that are well-established among medical researchers (i.e., random assignment).

The problem facing the field of sex offender research is that the best studies identified by Rice and Harris (2003), by Kenworthy et al. (2004) and by Hanson et al. (2002) were all different. It was not that one group of researchers was more lenient or more restrictive than another

concerning study quality; the problem is that most of the studies rated as credible by one group were considered inherently biased by the other groups. For example, Kenworthy et al. (2004) included studies in which the outcome criteria involved self-reported changes on psychological characteristics, whereas Rice and Harris (2003) and Hanson et al. (2002) excluded such studies on the grounds that intermediate measures lack sufficient validity to make strong conclusions. Only one study was included among the "best" studies in all three reviews: California's Sex Offender Treatment and Evaluation Project (SOTEP; Marques, Wiederanders, Day, Nelson & van Ommeren, 2005). The SOTEP study is unique in that it used a strong research design (random assignment) to evaluate a credible (i.e., cognitive-behavioural) treatment program for adult sexual offenders.

Definition of study quality

In order to rate study quality, it is necessary to have a definition of what is being rated. Consistent with the recommendations of the Potsdam Panel (Cook, Sackett & Spitzer, 1995), we consider good studies to be those that minimize bias. In the ideal study the effect size calculated from the study would be wholly attributable to differences in treatment (plus random error). Bias is the major criterion for judging study quality, but, it is also worth considering the *confidence* that can be placed in the finding. A random assignment study, for example, would not be expected to produce systematic differences between groups; nevertheless, increased confidence can be placed in the results when the researchers explore various potential threats to validity and are able to demonstrate that the study was implemented as intended. Consequently, a high quality study is one in which the judgement of *minimal bias* can be made with *high confidence*.

Assumptions guiding the rating scheme

The Collaborative Outcome Data Committee's Guidelines for the Evaluation of Sexual Offender Treatment Outcome Studies (CODC Guidelines) were based on the following assumptions.

A) It is possible to rate study quality

One initial assumption is that studies vary in the extent to which they can inform research questions, and that the better studies should be given more weight than lesser studies. This assumption is not universally shared. Greenland (1994a, 1994b) argued that rating study quality introduces subjective bias and has little relationship to outcome (Greenland & O'Rourke, 2001). It is difficult to create an internally consistent (single factor) measure of study quality and even harder to infer such a dimension from published reports. Instead of including global measures of study quality, Greenland recommends examining the effects of the quality items (the score components; Greenland, 1994a; Greenland & O'Rourke, 2001). For example, meta-analytic reviews could test whether studies that use long follow-up periods find different results than studies using short follow-up periods.

Greenland's position needs serious consideration given that the results of meta-analyses can differ based on different study quality rating schemes (Juni, Witschi, Bloch & Egger, 1999). Nevertheless, researchers and reviewers must make some judgement concerning study quality, even if it is the dichotomous decision about whether a particular study should be considered evidence or not. Given a large number of studies with uncorrelated variation on study attributes, it may be possible to empirically model the effects of study attributes on outcome. When there is a relatively small number of studies with correlated features, the statistical modelling suggested by Greenland is unlikely to be informative.

In agreement with all editors of scientific journals, we believe that reviewers and researchers can and should make judgements concerning study quality. Guidelines for rating study quality are not only helpful for evaluating existing research, but they can motivate researchers to conduct new studies that are as informative as possible.

B) Knowledge is cumulative

One debate within the research community is whether questions can be best answered through a single, definitive study or through the accumulation of results from many lesser studies. In medicine, the single definitive study is often a multi-site, randomized clinical trial involving thousands of patients. Although the results of such studies can be convincing, they are slow and costly enterprises, which are only mounted once there is a reasonable expectation of success based on earlier, lesser studies. One irony is that a cumulative meta-analysis of the earlier, smaller studies often provides the same answer as the definitive study, prompting debate about the necessity of such large scale clinical trials (Lau, Schmid & Chalmers, 1995). Both are needed: given disagreement, the large clinical trial is considered more convincing than the summary of diverse, smaller studies (LeLorier, Grégoire, Benhaddad, Lapierre & Derderian, 1997).

In the field of sexual offender treatment, it is unlikely that there will ever be a "definitive" study, however desirable that would be. The complexity of the interventions and the long delays needed before knowing the ultimate outcome (i.e., recidivism) present significant technical obstacles, even if there was the social and political will for generous investment in sex offender research. Furthermore, the heterogeneity of the sexual offender population precludes the answers from being found in any single study. Consequently, the future of sexual offender outcome research will involve the accumulation of evidence given by small studies.

C) Multiple methods are needed

Research is a problem solving activity, and there is no single method for determining the truth. Nevertheless, the last century has seen the acceptance of certain standard solutions to common research problems. In particular, random assignment has been recognized as the gold standard for minimizing pre-existing differences between the

treatment and comparison groups. Random assignment does not eliminate the differences, but, if well executed, creates the expectation that the influence of these differences will average out to zero. Random assignment studies have been criticized because they can result in withholding treatment from a potentially dangerous clientele; however, random assignment may be the most ethical approach to assigning treatment when the demand for treatment exceeds the resources available.

For complex social interventions, random assignment studies face significant challenges, both conceptually and practically. Consequently, researchers have developed a range of alternative designs for evaluating social problems (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002). These designs are often referred to as "quasi-experiments" because the researcher does not have full control over who receives the intervention. It is widely accepted that quasi-experimental designs can make important contributions to knowledge, but that special care is required in their design, implementation and interpretation.

Random assignment studies clearly have merit, and researchers should search for opportunities to implement such designs. They are not, however, the only source of information; different approaches are needed to address different research questions. Within the field of general psychological treatment, the limitations and unintended consequences of randomized clinical trials are becoming increasingly recognized (Haaga, 2004; Westen, Novotny & Thompson-Brenner, 2004). It is our opinion that knowledge of the effectiveness of sexual offender treatment will be based on diverse research methods. Although none of these designs can be conclusive in themselves, the cumulative contribution of different studies will increasingly restrict the range of plausible interpretations.

D) Program evaluation can and should contribute to cumulative knowledge

Most studies of sexual offender treatment are program evaluations, not scientific experiments. In scientific experiments, the research is designed to address questions of scientific interest. The results and implications of the experiment are important; what happened in the experiment itself is simply a means to the end of advancing knowledge. In contrast, program evaluations are concerned with the workings of a specific program. Administrators want to know if this specific program works (not "programs in general"), and funding decisions often hinge on the results of such evaluations.

It is our position that well-designed program evaluations can contribute to cumulative knowledge. Even when the program was not designed as a research study, it is possible for evaluators to collect information that informs questions concerning the efficacy of both "this specific program" and programs "like this one". Furthermore, when the effectiveness of programs for sexual offenders is debatable, administrators sponsoring treatment programs have a responsibility to evaluate their programs and to contribute to

cumulative knowledge of what works for sexual offenders. Consequently, the CODC Guidelines focus considerable attention on how to maximize the contribution of program evaluations to cumulative knowledge.

Review of methods for rating study quality

Although formal assessments of study quality are relatively new, a large number of scales and checklists have been developed within the medical field to assess the quality of randomized and clinical trials (see Juni et al., 1999). Study quality assessments have been used in systematic reviews and meta-analyses; as well, medical practitioners have been encouraged to use study quality guidelines to critically evaluate research studies in order to improve the treatment of their patients.

Moher et al. (1995) identified 25 scales and nine checklists, noting considerable heterogeneity among them. The number of items ranged from 3 to 34, with variable methodological features and reporting characteristics assessed. Most of the scales had at least one item evaluating patient assignment, masking procedures, and statistical analyses. In addition, many had items evaluating the quality of the reporting, ethical issues (e.g., obtaining consent), and the interpretation of the results.

These scales, however, do not measure a common definition of "quality". Juni et al. (1999) found that the choice of quality assessment scale affected the results of the meta-analysis. They scored 17 studies comparing low-molecular-weight heparin with standard heparin for the prevention of postoperative thrombosis using 25 different quality assessment scales. These scales did not consistently identify the same studies as "high quality", and the best studies identified by these different scales yielded different results. They concluded that the quality rating scales are heterogeneous and that many of the items concern reporting quality, ethical issues, and data interpretation rather than bias or internal validity.

In criminology, one of the most influential rating scales is the Maryland scale (Sherman et al., 1997). Originally developed to help identify promising crime prevention programs (see also Aos, Phipps, Barnoski & Liebe, 1999), it was the quality rating scale used by Lösel and Schmucker (2005) in their meta-analysis of sexual offender treatment outcome studies.

Raters using the Maryland scale consider seven elements of "methodological rigour" prior to forming an overall rating. The seven elements are the following:

1) sample size

2) type of comparison groups

3) use of control variables to account for initial group differences

4) appropriateness of the variables assessed

5) attrition

6) length of follow-up

7) whether or not the study used statistical tests

The final rating ranged from 1 to 5, with level 5 being the most rigorous.

The overall rating is based on the following three general concerns:

1) the study's ability to control extraneous variables;

2) the expected amount of measurement error; and

3) statistical power.

The Maryland scale has the advantage of being widely applicable to a broad selection of criminal justice intervention studies. However, like most of the rating scales used in the medical field, the Maryland scale lacks a coherent definition of quality, combining concerns about statistical power and bias.

The Maryland scale assumes that the reviewers are interested in the conclusions of the different studies rather than collecting the data from these studies for secondary analysis. When rating studies for meta-analysis, concerns about statistical power or measurement error fade in comparison to concerns about bias.

A rating scheme specific to sex offender treatment outcome

Despite the considerable work on developing study quality rating schemes, none of the existing scales are well-suited to measuring the quality of sexual offender treatment outcome studies. There are some common features that are relevant to most research studies, but the important threats to validity vary with the questions being addressed. For example, body mass, diet and exercise would be important variables in a study examining the efficacy of treatment for diabetes, whereas sexual offender researchers would be more interested in variables such as marital status, lifestyle impulsivity, and the equivalence of the recidivism criteria. Evaluating study quality requires knowledge of the problem being studied.

The CODC Guidelines focus on the special concerns associated with the design and implementation of sexual offender treatment outcome research. They were intended for studies that compared a treated group of sexual offenders to a comparison group (or norms), using recidivism as the outcome criterion. Because the outcome criterion of interest occurs many years after the end of treatment, certain designs are difficult to implement (e.g., wait-list control, regression discontinuity) and are not discussed. Instead, the Guidelines focus on the decisions commonly faced in sexual offender outcome studies, such as the choice of control variables, recidivism criteria, and sample size.

Overview of the CODC Guidelines

The Guidelines were based on a review of existing study quality scales (e.g., Cowley, 1995; Downs & Black, 1998; Gibbs, 1989; Miner, Murphy & Yates, 2002; Reisch, Tyson & Mize, 1989; Sherman et al., 1997; Thomas, Ciliska, Dobbins & Micucci, 2004; Wortman, 1994; Zaza et al., 2000) as well as specific concerns that have been raised about sexual offender research (e.g., Rice & Harris, 2003). Much of the content and structure of the CODC Guidelines were derived from an analysis of how CODC members described the strengths and weaknesses of individual research studies. Although we originally envisioned separate criteria for specific designs (e.g., random assignment, cohort), the concerns raised for the different designs were remarkably similar. Consequently, the CODC Guidelines present general criteria for evaluating sexual offender treatment outcome studies, and only occasionally provide distinct questions for specific research designs.

The CODC Guidelines contain 20 items organized into the following seven categories:

1) administrative control of the independent variable;

2) experimenter expectancies;

3) sample size;

4) attrition;

5) equivalence of groups;

6) outcome variables; and

7) correct comparison conducted.

As well, there is one item to be rated only for cross-institutional designs (Sample Size of Institutions), and three checklists to help with the rating of Item 13 (A Priori Equivalence of Groups) for specific types of designs (random assignment, risk band/norm, and cohort studies). A flow chart, adapted from Zaza et al. (2000), is provided to help reviewers categorize studies.

The 20 (21) items concern the extent to which the study's features introduce bias in the estimation of the treatment effect, or influence the confidence that can be placed in the study's findings. If the information is limited, raters are encouraged to seek out additional information and re-rate the item.

The overall judgement of study quality is a form of structured judgement. After rating the individual items, evaluators are asked to form global judgements as to the extent of "bias" inherent in the research design, and the "confidence" that can be placed in the bias rating. The bias ratings are as follows: a) no bias or minimal bias expected; b) some bias expected; and c) considerable bias expected. The overall confidence ratings similarly use a three-point scale: a)

little or no confidence in the results; b) some confidence; and c) confidence in the results as reported. Based on the ratings of confidence and bias, studies are placed in one of four categories:

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

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

3) WEAK
   Some confidence that the studies have no more than a small amount of bias (intermediate rating). These studies have significant flaws, but are of possible relevance to the question of treatment effectiveness. Weak evidence at best.

4) REJECTED
   Low confidence in the results, *or* considerable bias. These studies have 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.

Reliability

Two undergraduate students (3rd year Criminology; 4th year Psychology) were given 5 days training on the use of CODC guidelines[1]. This training primarily involved rating and reviewing eight practice studies with a trainer[2]. The two raters then independently coded 10 studies.

On the Global rating, the coders agreed on nine of the 10 studies (*ICC* = .95). There was 100% agreement on Global Confidence (*Kappa* = 1.0; *ICC* = 1.0), 90% agreement on Global Bias (*ICC* = .69; *Kappa* could not be computed), and 70% agreement on Global Direction of Bias (*Kappa* could not be computed).

The level of agreement for the individual items was also high. For most of the categories, the median level of agreement was 1.0.

[1] Leslie Helmus, Shannon Hodgson.
[2] Guy Bourgon

A second reliability study was conducted using 12 experts in the field of sexual offender research[3]. To examine the reliability of the expert's ratings, 10 hypothetical studies were used. Real studies were not used because the experts would be expected to have already formed opinions about the existing studies, either as authors or reviewers. The experts were not provided with any specific training in the use of the CODC guidelines, although half of them would have known about the guidelines through their membership on the CODC committee.

The degree of agreement among the experts was poor. The experts had moderate levels of agreement for the individual items, but disagreed on the overall ratings. Of the 10 studies, only three had two common ratings (all rejected). Some of the disagreements were due to errors that could have been corrected given training and additional care and attention (failure to notice study features, misinterpretation of the coding rules). Subsequent discussions among the experts, however, revealed principle disagreements concerning the minimum features required for the studies to be considered "good enough".

There was substantial agreement that the individual features identified in the CODC guidelines were important indices of study quality, but the experts had divergent views as to the relative importance of these features in influencing overall study quality.

The main conclusions of the reliability studies are that it is possible to train naïve raters to reliably use the CODC guidelines; simple exposure to the guidelines, however, was insufficient to change strongly held beliefs about the appropriate methodology to use in sexual offender outcome research (for similar findings in medical research, see Schroter et al., 2004). We had initially assumed that the general principles of research design would be intuitively obvious to knowledgeable experts, but this assumption proved to be false.

Many of the experts passionately disagreed on what constitutes a "good" study. Nevertheless, all the experts agreed that the features outlined in the CODC guidelines were important to consider in rating study quality.

Uses of the CODC Guidelines

There are three tasks for which the CODC Guidelines should be helpful. The tasks are the following:

1) reviewing existing studies;

2) evaluating existing programs; and

3) designing new studies of treatment effectiveness.

[3] Guy Bourgon, Andrew Harris, Grant Harris, Niklas Långström, Roxanne Lieb, Ruth Mann, Robert McGrath, William Murphy, Vernon Quinsey, Marnie Rice, David Thornton, Pamela Yates.

Reviewers can use the Guidelines as a part of their selection criteria for narrative or quantitative syntheses of the evidence of the effectiveness of treatment for sexual offenders. The Guidelines should also be helpful to editors of professional journals (and reviewers) as a means of rating study quality and providing direction for improvements. For program developers, the guidelines can suggest features that facilitate future evaluation (e.g., routinely collecting information on the individuals not admitted to the program).

Program evaluators are often given less than ideal conditions under which to determine the effectiveness of treatment. Nevertheless, evaluators can use the Guidelines to make design decisions that maximize information at a minimal cost. For example, limited assessment resources can be focussed on risk-relevant variables and established risk scales, the inclusion criteria can be clearly specified, and the analyses can include all offenders assigned to treatment (intent-to-treat).

When researchers have the opportunity to design new research studies, we recommend that they use strong research designs, including random assignment to treatment and comparison conditions. Furthermore, we recommend that offenders are matched on risk prior to being assigned to treatment. Random assignment studies are politically unpopular and difficult to implement, but the benefits of these studies are such that researchers should advocate for random assignment studies whenever possible. Researchers using random assignment studies, however, should be prepared for breakdown in the randomization procedure. Consequently, we recommend that all participants (treatment and control) are assessed pre-treatment on risk relevant variables, and that researchers are vigilant to problems of treatment integrity, attrition, and cross-over (comparison group receiving equivalent services). In addition, dedicated research studies should use a clearly specified treatment that has a reasonable expectation of being effective (e.g., Andrews & Bonta, 2006, Chapter 10). It would be inappropriate to represent a study to be one of “sexual offender treatment” if the intervention was not considered credible by contemporary standards.

References

Andrews, D. A., & Bonta, J. S. (2006). *The psychology of criminal conduct* (4th edition). Cincinnati, OH: Anderson.

Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (1999). *The comparative costs and benefits of programs to reduce crime: A review of national research findings with implications for Washington State* (Document No. 99-05-1202). Olympia, Washington: Washington State Institute for Public Policy.

Brooks-Gordon, B., Bilby, C., & Wells, H. (2006). A systematic review of psychological interventions for sexual offenders I: Randomised control trials. *Journal of Forensic Psychiatry and Psychology, 17*, 442-466.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Cook, D.J., Sackett, D.L., & Spitzer, W.O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis. *Journal of Clinical Epidemiology, 48,* 167-171.

Cowley, D.E. (1995). Prostheses for primary total hip replacement: A critical appraisal of the literature. *International Journal of Technology Assessment in Health Care, 11*, 770-778.

Downs, S.H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*, 377-384.

Furby, L., Weinrott, M.R., & Blackshaw, L. (1989). Sex offender recidivism: A review. *Psychological Bulletin, 105,* 3-30.

Gallagher, C. A., Wilson, D. B., Hirschfield, P., Coggeshall, M. B., & MacKenzie, D. L. (1999). A quantitative review of the effects of sex offender treatment on sexual reoffending. *Corrections Management Quarterly, 3,* 19-29.

Gibbs, L.E. (1989). Quality of study rating form: An instrument for synthesizing evaluation studies. *Journal of Social Work Education, 25,* 55-67.

Greenland, S. (1994a). Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology, 140,* 290-296.

Greenland, S. (1994b). Quality scores are useless and potentially misleading. Reply to "Re: A critical look at some popular meta-analytic methods." *American Journal of Epidemiology, 140,* 300-301.

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics, 2,* 463-471.

Haaga, D.A.F. (2004). A healthy dose of criticism for randomized trials: Comment on Westen, Novotny, and Thompson-Brenner (2004). *Psychological Bulletin, 130,* 674-676.

Hall, G.C.N. (1995). Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology, 63,* 802-809.

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.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1998). Appraisal and management of risk in sexual aggression: Implications for criminal justice policy. *Psychology, Public Policy, and Law, 4,* 73-115.

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.

Kenworthy, T., Adams, C.E., Brooks-Gordon, B., & Fenton, M. (2004). Psychological interventions for those who have sexually offended or are at risk of offending (Cochrane Review). *Cochrane Library, Issue 3.* Chichester, UK: John Wiley & Sons.

Lau, J., Schmid, C.H., & Chalmers, T.C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology, 48,* 45-57.

LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *The New England Journal of Medicine, 337,* 536-542.

Lösel, F., & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology, 1,* 117-146.

Marques, J.K., Wiederanders, M., Day, D.M., Nelson, C., & van Ommeren, A. (2005). Effects of a relapse prevention program on sexual recidivism: Final results from California's Sex Offender Treatment and Evaluation Project (SOTEP). *Sexual Abuse: A Journal of Research and Treatment, 17,* 79-107.

Miner, M., Murphy, W., & Yates, P.M. (2002). *Research criteria for ATSA Collaborative Data Project: Subcommittee Report*. Unpublished manuscript.

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*, 62-73.

Reisch, J.S., Tyson, J.E., & Mize, S.G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics, 84*, 815-824.

Rice, M. E., & Harris, G. T. (2003). The size and sign of treatment effects in sex offender therapy. *Annals of the New York Academy of Sciences, 989,* 428-440.

Schroter, S., Black, N., Evans, S., Carpenter, J, Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: Randomized controlled trial. *British Medical Journal, 328*, 673-5.

Sex Offender Treatment Review Working Group. (1990). *The management and treatment of sex offenders.* Ottawa: Solicitor General Canada.

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

Sherman, L.W., Gottfredson, D., Mackenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising. A report to the United States Congress.* College Park, Maryland: University of Maryland, Department of Criminology and Criminal Justice.

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.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges, (Eds.), *The handbook of research synthesis* (pp. 97-109). New York: Russell Sage Foundation.

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-sp.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-sp.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 Honorary 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