J Exp Criminol (2010) 6:355–376
DOI 10.1007/s11292-010-9104-6

Lessons from a successful and failed random assignment testing batterer program innovations

Edward W. Gondolf

Published online: 24 July 2010

Abstract With increasing pressure to conduct experimental evaluations of domestic violence interventions, it is important to weigh further the most challenging aspect of experimental designs: the implementation of random assignment. This paper reviews two attempted experimental evaluations of counseling programs for domestic violence offenders, and formulates implications for conducting and interpreting program evaluations. The two case studies offer an instructive comparison of a maximally implemented experiment and a failed one at the same setting. In the first study, the random assignment was introduced within the counseling program and with implicit leverage of court sanctions for non-compliance to the assignment. In the second, random assignment was disrupted by unforeseeable events and inter-agency breakdowns in the complex referral system. Interestingly, implementation issues in both studies raised divergent interpretations from researchers and practitioners. They appear to imply a need for more disclosure of implementation problems in experimental evaluations and for more caution about over interpreting the existing experimental evaluations in the field.

Keywords Domestic violence · Counseling program · Random assignment · Experimental evaluation · Implementation problems

Introduction

"Bronzed" by implementation?

Experimental evaluations have assumed a decisive role in determining the direction of batterer programs. Based on the evidence drawn primarily from five experimental evaluations, a host of articles assert that batterer programs "don't work" or at least

E. W. Gondolf (✉)
Mid-Atlantic Addiction Research and Training Institute, Indiana University of Pennsylvania, Indiana, PA 15705, USA
e-mail: egondolf@iup.edu

"don't work very well" (e.g., Corvo et al. 2008; Labriola 2005). Most go on to recommend replacing domestic violence intervention with couples counseling through mental health clinics (Corvo et al. 2008), reverting to intensified probation or community service in place of batterer programs (Davis et al. 1998), or adding innovative psychological approaches to batterer programming (e.g., Levesque et al. 2008).[1] These interpretations rest largely on the positivist position that experimental designs are the "gold standard" (Feder and Boruch 2000; Sherman 2009; Weisburd 2003). Random assignment of subjects to treatment and control groups is the best way to isolate the effect of treatment compared to "no treatment" or some alternative treatment. Of course, treatment fidelity and provider competence are equally important in order to ensure validity.

While researchers promote efforts to negotiate the array of challenges facing criminal justice experiments (e.g., Cissner and Farole 2009), even the experts acknowledge that experimental designs too often end up the "bronze standard" rather than the gold: "It is difficult to do randomized experiments well.... Textbook requirements are rarely met" (Berk 2005, p. 20). According to both domestic violence researchers (Dobash and Dobash 2000; Mears 2003; Gondolf 2001; Saunders 2008) as well as evaluators in the criminal justice field (Goldkamp 2008; Berk 2005; Weisburd et al. 2003), there is wide-spread concern in general about sample override and attrition, the effect of subgroups, sufficient treatment monitoring, compensating treatments in the control group, intention-to-treat versus dosage effects, simplistic outcome measures, and low or biased follow-up response. The fundamental contributor to this "bronzing" is the difficulty of achieving a random assignment amidst subjects who opt out of a study, practitioners who override assignment, and agencies that directly oppose or undermine assignment (Berk 2005; Dobash and Dobash 2000; Goldkamp 2008; Hollin 2008).[2]

Several reviews of published articles on experimental evaluations, however, conclude that discussions of random assignment problems, and other implementation issues for that matter, are "insufficient" or "inadequate," (Durlak and DuPre 2008; Kober et al. 2006; Han et al. 2009), yet these sorts of problems are shown to have a substantial impact on outcomes (Durlak and DuPre 2008). The articles reporting on the experimental evaluations of batterer programs similarly fall short on the backdrop of implementation. One might argue that it is necessary to report only the factors needed to judge the success of implementation and let the readers draw their own conclusions. The question remains, however, what are the criteria for identifying and including the complications that others may find sufficient and informative?

Consideration of at least a few batterer program evaluations suggests that there is often much more of a backstory to implementation than meets the eye (Feder et al. 2000; Visher et al. 2008a, b). In this article, we add an account from two of our attempted experimental designs to further elaborate the implementation challenges of

[1] Batterer programs around the country have reported a reduction in funding, a reduction in referrals, and revisions in program standards as a result of the "don't work" interpretation of the existing experimental revaluations of batterer programs.

[2] In the experimental evaluation of a batterer program in south Florida, nearly 30% of the domestic violence offenders eligible for random assignment were excluded from the study because the judge or prosecutor objected to the assignment (Feder and Dugan 2002).

especially random assignment, offer some comparative information, and reinforce the emerging lessons for future research. The case examples, we believe, also raise cautions about the overinterpretation of the current experimental evaluations and the "don't work" inference being drawn from them.

Previous implementation studies

A recent systematic review, including five meta-analyses, examined the impact of design implementation on 500 quantitative evaluations of adolescent "health programs" (e.g., mental health, drug use, violence, bullying) (Durlak and DuPre 2008). It identified a "powerful impact" of implementation on outcomes. Specifically, the effect size is substantially higher "when programs are carefully implemented free of serious implementation problems" (p. 340). Even though their results confirm the importance of using implementation information to interpret evaluation outcomes, the authors' analysis and previous studies show that few research reports sufficiently discuss the design implementation. The authors also identify the factors most associated with the variation in implementation: community factors, provider characteristics, treatment stability, organizational capacity, training and technical assistance, and the relationship among such factors. Other previous reviews have also identified shared decision-making, leadership, administrator support, and coordinated infrastructure.

The authors acknowledge, however, that these factors are likely to differ for other types of intervention and recommend on-going attention to, and study of, implementation in other fields. Further examination of experimental evaluations of batterer programs seems warranted in this regard. The domestic violence intervention admittedly includes a complexity of contextual factors, victim access and safety, collaboration issues, and criminal justice priorities that make evaluation research challenging to begin with (Mears 2003).

A 2007 meta-analysis from the Cochrane Collaboration[3] did apply a set of six implementation criteria to six batterer program evaluations that met its initial standard of a randomized experiment. It first of all determined that "The methodological quality of the included studies was generally low" (Smedslund et al. 2007, p. 17). The overall findings of the meta-analysis suggest that the "low quality" limits the inferences one can draw from the evaluations. Using a relative risk measure and confidence intervals, the authors of the Cochran report surmised: "The research evidence is insufficient to draw conclusions about the effectiveness of cognitive behavioral interventions for spouse abusers... We simply do not know whether the interventions help, whether they have no effect, or whether they are harmful." (p. 18). An earlier analysis funded by the Centers for Disease Control also used quality ratings with a broader inclusion of 50 intervention and prevention programs. It reached a similar conclusion (Morrison et al. 2003, p. 4): "The diversity of data, coupled with the relatively small number of studies that met the inclusion criteria for the evidence-based review, precluded a rigorous, quantitative synthesis of the findings."

[3] The Cochrane Collaboration compiles reviews of bio-medical experimental evaluations to guide especially medical practice (see www.cocrhanelibrary.com).

Batterer program evaluations

Two rather extensive implementation studies of batterer program evaluations reveal the complications that can occur in implementing a randomized experimental evaluation. The problems that arise go well beyond those generally acknowledged in published research reports and suggest that a broader backstory may accompany such research studies. They point not only to additional qualifications of the experimental findings but also to some valuable lessons for further evaluation efforts. At a minimum, they show the need to anticipate the disruption of "uncontrollable events" and develop contingency plans in advance. As the previous reviews indicate, the resistance to randomized assignment and inter-agency barriers stand out as major obstacles to experimental quality.

Feder et al. (2000) reviewed the challenges they faced in conducting the Broward, Florida, experiment of a batterer intervention program, and also a Portland, Oregon, experiment of an arrest and prosecution partnership in domestic violence cases. The Broward experiment exposed familiar problems of resistance from agencies, turnover in key staff, and shifts in agency support despite the initial agreements (Durlak and DuPre 2008). All these problems were compounded by heightening opposition from the prosecutors who objected to the random assignment and used court action to block the study. This opposition increased the resistance from other parties involved in the study, created a hostile environment in the courtroom, lowered the morale of the research staff, and resulted in staff turnover. As a result, the researchers had to do recruiting and training of new staff, convene regular meetings to address morale, and re-educate agencies to the merits of the experiment. The experiment was still completed but not without questions about the effects of the "hostile environment" and procedural adjustments on the results (for discussions of the "questions", see Dobash and Dobash 2000; Gondolf 2001; Saunders 2008).

The authors reinforce the need to educate the participating agencies about the merits of experiments, ensure that agreements extend to frontline staff and all affected agencies, and address resistance from the outside community. In sum, they raise the need for a major public relations effort to counter suspicion and resistance to experiments in general.

Some of these challenges might be eased with well-developed researcher–practitioner relationships prior to initiating an experiment and a "research readiness" at the research site. This sort of "ready" setting is more difficult to find or establish thus far in the domestic violence field. When it is, the site is not likely to be particularly representative of the majority batterer program settings that operate under very different conditions. Even with all of this, "some things are beyond one's control" (Feder et al. 2000; p. 391).

Another extensive implementation study is the process evaluation of the Judicial Oversight Demonstration (Visher et al. 2006, 2008a, b). Visher et al. (2008a, b) conducted a multisite quasi-experimental evaluation of a system-wide intervention, and not specifically of batterer program effectiveness. However, their study does offer further insight into the challenge of implementing evaluation research in the domestic violence field. The objective of the study was to determine the effectiveness of what could be considered "an enhanced coordinated response" to domestic violence cases. Cross-training, inter-agency planning, additional personnel, revised procedures, and managed collaboration were initiated to bring more immediate and needed services to victims and greater accountability to perpetrators.

A comparison of enhanced sites to those without the enhancements, and also a pre-post comparison at one of the three enhanced sites, produced mixed outcomes and a "don't work" interpretation by some (Peterson 2008). The researchers acknowledge that the study was not designed to assess the effectiveness of individual "component services" in what was considered a system intervention. Nonetheless, they remark; "The results suggest, like those of other studies, that referral to batterer intervention programs does not have a powerful effect in reducing intimate partner violence" (Visher et al. 2008a, p. 520).

The extensive implementation reports discuss, however, not so much resistances to the study, but more the inter-organizational barriers (Visher et al. 2006, 2008a, b). Most notable, and perhaps least surprising, were the "competing priorities" between the domestic violence services and the criminal justice system.[4] The differences were compounded by a lack of knowledge about other agencies' operations and of data sharing across agencies. One of the biggest impacts of the project may have contributed to its mixed outcomes. The caseload across the agencies escalated and, along with it, the workload.

We are left, therefore, with uncertainties about whether the mixed outcomes are attributable to an overloaded, insufficiently staffed, and less than optimal system, or to the collaborative intervention itself. Interestingly, practitioners at the sites remain very enthusiastic about the collaboration and its outcome, and feel a researcher–practitioner gap in the implementation may have contributed as well to the results and their current interpretation (see BISCMI 2009).[5]

The backstory information on batterer program evaluations is otherwise relatively limited, but the Cochrane report found some fundamental implementation issues in its review of study reports and follow-up phone calls to the researchers. Researcher presentations, practitioner feedback, and critical views have drawn out some additional challenges faced in these evaluations (e.g., BISCMI 2009; BWJP 2008). Many of these have been summarized in critical narrative reviews of the research (Gondolf 2001; Murphy and Ting 2010; Saunders 2008). Admittedly, an extensive or complete set of implementation studies is not available, and only speculation remains on the extent and impact of potential backstories.

In the initial Brooklyn experiment, there were problems with one of the programs changing its format during the course of the study, and a low response rate that led to the use of investigators pursuing victims for an interview (Davis et al. 1998). The more recent Bronx study had to deal with judges who were inconsistent in their monitoring of non-compliance, and a victim response rate of 25% with an already small sample size (Labriola et al. 2008). The most successful random assignment was in the San Diego Navy experiment (Dunford 2000). There was, however, poor compliance to the couples option which led to collapsing that option with another group option in the analyses. Questions also remain about the impact of the military

[4] It is difficult to generalize across the three sites of the Judicial Oversight Demonstration project since the settings and implementation varied. Insights can be drawn by comparing the conditions and outcomes at the sites. For instance one site did use an innovative and successful data sharing method between criminal justice and batterer intervention programs.

[5] This specific impression is based on emails from batterer program directors involved in the study, conference presentations by judges participating in the study, and comments from practitioners at the advisory meetings for the study.

command structure, the officers' supervision, and the potential sanctions on the subjects' behavior.

The five most recent experimental evaluations, including the Florida study with a published implementation review, suggest that the implementation issues of concern are associated with the prominent evaluations, as previous reviews suggest. The two published implementation reports hint that the extent of such issues could be more far reaching and disruptive than implied in the findings articles. More importantly, they also imply that there are instructive lessons to drawn from the implementation challenges—lessons that might improve future evaluations, as well as shed light on interpretations of the existing ones.

Method

We compare two additional case studies of experimental batterer program evaluations of a different sort (Gondolf 2005, 2009b). They are both tests of program innovations rather than the previous experiments of program effectiveness based on a "non-treatment" control group. The lack of a non-treatment control group was in part justified by our finding of a moderate effect at the program sites, using an instrumental variable analysis (Gondolf and Jones 2001), and subsequent propensity score analyses (Jones et al. 2004).[6]

The first experiment tested culturally-focused counseling for African-American men against conventional cognitive-behavioral batterer counseling in an all African-American group and against the conventional approach with a racial-mixed group (Gondolf 2007). The second experiment was to examine supplemental mental health treatment for batterer program participants who screened positive for mental health problems (Gondolf 2009a). The comparison control group was to be men who screened positive but attended only the batterer program.

The two studies, moreover, offer a comparison of a successful implementation of random assignment and the experimental design overall, and a study in which the random assignment was aborted and the design substantially revised. Yet both attempted experiments were at the same site with nearly the same program and research staff. A series of previous studies at the site helped to reduce the resistance to research and disregard for agreements, experienced in the case studies mentioned above. We had, in fact, previously conducted an experimental evaluation of a parenting class coupled with batterer program sessions at the site (Gondolf 1997), as well as a multisite evaluation of batterer intervention systems (Gondolf 2002).

The speculations about research readiness or agency resistance were eased substantially at this research site. What then might account for the successful versus failed implementation? Is it merely a greater degree of some of the barriers cited in the previous implementation accounts, or uncontrollable events that impinged on the aborted experiment? We are also interested in any recommendations suggested in the answer to these questions. Are there further strategies or procedures that might help improve the implementation of batterer program evaluations in the future?

[6] Even if the existing batterer program was ineffective, the recommended innovations would ideally improve the program outcomes from "not effective" to an effect over the program by itself.

Our method is basically a descriptive case study approach. It is admittedly a largely impressionistic overview of the implementation of two attempted experiments. It focuses on design elements, site circumstances, and research protocol that may have affected particularly the random assignment of subjects to experimental and comparison groups. We focus especially on the categories of barriers and problems identified in the previous implementation accounts, namely, resistance to experiments, unfulfilled agreements, staff turnover, and agency priorities. Our information is drawn from process evaluations conducted during both attempted experiments (Gondolf 2005, 2009b).

The sources for these were direct observation of the random assignment procedures; monitoring of subject assignment compliance in terms of characteristics, overrides, and compliance; development meetings with agency representatives; training sessions with frontline staff; periodic meetings with research, program, and agency staff; and informal conversations and formalized debriefing interviews with agency representatives and staff. Fieldnotes from all these sources were compiled, reviewed for themes and topics, and summarized in the final reports and in an article about the aborted experiment. Drafts of the summaries were reviewed with program staff, research assistants, and consultants involved in the studies, and their feedback was used to verify and revise the interpretations.

Case studies

Culturally-focused batterer counseling

Research design

Our first experiment examined the effect of specialized batterer counseling for African-American men, as mentioned. Numerous clinicians and researchers have recommended culturally-sensitive approaches for the substantial portion of African-American men referred to batterer programs. However, there was previously only a small preliminary study examining the effectiveness of alternative approaches for African-American men sent to batterer programs. We conducted an experimental clinical trial of culturally-focused counseling with African-American men who were mandated to batterer counseling by the domestic violence court in Pittsburgh, Pennsylvania.

A sample of 503 men was randomly assigned to one of three monitored counseling options: culturally-focused counseling, African-American-only conventional counseling, and racially-mixed conventional counseling. No significant difference was found in the reassault rate reported by the men's female partners over a 12-month follow-up period (23% overall). Despite the straightforward implementation and clear cut findings, several questions emerged about the internal and external validity of the study.

Successful random assignment

There are some obvious features that distinguish the successful random assignment in the culturally-focused counseling experiment. The random assignment was made

at the point of batterer program intake, rather than in the courtroom during the court referral or sentencing process. The batterer program referrals were, therefore, already established and assignments were made at the prerogative of the program. The batterer program ultimately had authority over the assignment to its several program groups, and could assign men in the process to a "special group."[7]

Not attending the assigned group was considered non-compliance to the program, excepting special circumstances such as a conflict with one's job schedule, or some verifiable conflict with the group leader or another participant. If a man did not consent to the study or wished to withdraw at a later time, he was still required to attend the assigned group.[8] The program's court liaison reported non-compliance at periodic case reviews (every 30 days), and that led to increased sanctions in terms of more batterer program sessions, fines, further prosecution, or jail-time.

Another factor in the successful random assignment was treatment not being withheld in a control group, as it was in previous experimental program studies. Rather, there was an enhanced treatment that was tailored to the needs of African-American men. Interestingly, we did receive objections from a few men who felt being assigned to an all-African-American group was "racist." They claimed that domestic violence was a human problem, not just a special black problem, and that they should get the same treatment as other men. The premise of the segregated groups was also questioned by a few African-American men who lived and worked in racially mixed neighborhoods. Nearly 20% of the African-American sample were, moreover, involved with Caucasian or Hispanic partners.

In other words, even though there was high "coerced" compliance to the random assignment, there was some assignment resistance that may have affected the outcome. To address the larger influence of cultural identification, we administered the Racial Attitudes and Identity Scale (RAIS) and introduced its results as a control in our analyses along with the race of men's partners. There was no apparent effect of these factors on the reassault outcome.

Disagreement over treatment integrity

The treatment integrity was assessed in several ways, including direct observations from the training consultant overseeing the specialized counseling. According to rating sheets completed by the training consultant and those from a subsample of program participants ($n=100$), the counseling options were distinct in the intended directions. For instance, there was more discussion with the participants, and more cultural topics raised in the culturally-focused group. Despite the relatively positive ratings, the training consultant felt that the culturally-focused counselor was not conducting the sessions to the necessary ideal.

In essence, the trainer questioned the treatment integrity and also the provider competence—specifically, the counselor's ability to conduct the groups. Initially a former part-time staff had been selected to lead the culturally-focused group, but that

[7] At one point we considered assignment to a voluntary open-ended discussion group to serve as a "non-treatment" control group.

[8] Refusals and transfers amounted to less than 10% and were eliminated from the study. A test of possible bias in the characteristics of the deletions was negative.

person left the agency for "personal reasons" during our startup phase. The program director and research staff felt his departure was for the better, but the trainer felt otherwise based on his interaction with and observation of that person. In his assessment, the replacement did not have sufficient group skills or experience. The program director, however, saw that new person as a "usual" hire and fairly representative of the staff entry-skill level. The research answer to the difference of opinion would be to have had multiple counselors in at least a few experimental groups, but the funding ceiling precluded that and the workplan limited the time for further staff recruitment and training.

The trainer also believed that conflicts with the female program director had a negative effect on the culturally-focused counseling. According to the trainer, the director appeared less than supportive of, and even antagonistic towards, the culturally-focused group and the way it was conducted. The culturally-focused counselor complained that she occasionally disrupted his group with advice or demands about the procedures. He thought that he should answer to the training consultant rather than the other program staff. That in turn added to the director's concerns that the specialized counseling was violating basic program procedures regarding program attendance and payments. The trainer suspected racism underlying the director's behavior and comments, and the director felt sexism was behind the trainer's response to her authority. As a result, the culturally focused counseling was more of an appendage to the existing program rather than integrated into it.

I did meet several times with the culturally-focused trainer and program director, individually and together, to discuss their differences. The tensions appeared to subside in the periodic meetings with the trainer and on-going phone chats, and I assumed that the meetings resolved the differences for the most part. The extent of the trainer's disagreements and frustrations resurfaced, however, with a draft of the final report. The trainer voiced a long list of objections that went well beyond the implementation of the experiment. I attempted to represent those concerns in the final report, the journal article, and a conference presentation. I also encouraged the trainer to write a response for the report and offer a presentation at conferences but he chose not to do so.

The overall point is that disagreements inevitably arise among different interests and perspectives, and they need to be anticipated and addressed systematically. I did review the issues with an external advisory committee of researchers and practitioners, as well as present my sense of them in reports to the funding agency and in journals. The trainer continues to feel, however, that his perspective, experience, and observations were not fully heeded and are not fully "told."[9] The obvious steps of discussion with the involved parties and an advisory committee are obviously not always sufficient or suitable. They may be inadvertently influenced by

[9] Part of the problem may be that the intensity and extent of the trainer's objections weren't fully aired or realized until his review of the final report draft, and were not fully evident in his prior observational reports and other requested writing. The trainer was very busy with his own projects and visited the site from a long-distance on a periodic basis. An on-site presence or availability was not possible, but might have aided in such a situation. The researcher and staff, on the other hand, had their own distractions that may as well diffused the communication. The researcher had the responsibilities of overseeing the boarder implementation, program staff collaboration, follow-up interviewing, research staff and advisory committee, and budget and finances. As a result, he had a different impression of the extent of the problems felt by the trainer.

the researcher's own biases or perspective. Ultimately, an independent committee or ombudsman might be established to monitor conflicts over implementation and to hear appeals from trainers and program staff.

Reviewer differences

Interestingly, the reviewers of our reports and articles were split over our presentation of the implementation issues and resulting qualifications. One reviewer of the final report praised the implementation of the experiment, especially the relatively successful random assignment and apparent treatment integrity. He remarked on the high internal validity that these implied.[10] Another reviewer, however, raised concerns about the external validity of the experiment. He questioned the agency's structural support for the culturally-focused counseling. He explained further that a community-based setting rather than a court-linked program was needed to properly test the culturally-focused approach. He also questioned the "coercive" assignment of African-American men to the culturally-focused counseling rather than letting them self-select into such an option.

The research team itself was very aware that the findings contradicted much of the clinical wisdom in the field and worried that a bottom-line directive might be drawn from them. To blunt this tendency, we sought publication in a journal that includes policy responses to research articles. The response articles raise a series of cautions about a bottom-line interpretation. Potter (2007), for instance, argues against "the regressive nothing works" summation that some might draw from the experiment's findings: "A concern of many regarding the findings is the potential for such results to be used to move away from rehabilitative efforts, such as batterer intervention counseling, and continue the focus on 'get tough' or strict and conservative criminal justice policies, which have proved to be largely ineffective" (Potter 2007, p. 371). This concern is especially the case for African-American men who tend to receive more severe sentencing and are disproportionately represented in jail as it is.[11]

Our findings do strongly suggest that adding a culturally-focused group to an existing court-linked program does not necessarily improve outcomes, and broader cultural competency within an agency and community-based support may be warranted. The issues in implementing the experiment, and response to it, bear this out. Basically, the concern is that extra-therapeutic factors ranging from arrest practices, to risk assessment, to collective efficacy (the community's ability to hold offenders accountable) are likely to confound the outcomes.

In our conclusion, we recommended self-selection into a culturally-focused group, rather than imposing one on African-American men. We recognized, too, that the culturally-focused counseling did no worse than the other options, and is likely

[10] This reviewer noted, however, a failure to establish the association of specific cultural factors to outcomes in the research on the topic thus far. This omission raises the question whether the influential cultural aspects were addressed in the counseling and addressed sufficiently.

[11] The reaction essay extends the inferences by pointing out that "race is frequently a proxy for neighborhood" (Potter 2007, p. 370). The ecological context of the subjects needs to be factored into future analyses to make better sense of effectiveness: even the culturally-focused counseling may have simply remained too individualistic.

to improve at the site with some time and experience. Moreover, we observed that the broader discussions in the culturally-focused group raised concerns about unemployment, educational needs, parenting problems, community violence, and alcohol and drug use; all emerged in the broader group discussions. In a follow-up study, we introduced intake screening and systematic referral for such issues in an effort to extend the impact of batterer counseling (Gondolf 2008a, b).

Two major lessons stand out overall. One, despite the successful random assignment and apparent treatment integrity, question—and even objection—lingers over the experimental condition. The all-important trainer believes that the experimental condition did not meet the ideal, and that tensions within the program may have sabotaged his efforts. Two, report reviewers and article respondents suggest that the experiment is naïve to the larger social context. The African-American experience with the criminal justice system, and neighborhoods that are under resourced, need to be considered in the interpretation of the findings.

In sum, there may be much more to an experiment than makes its way into a report. What appears on the surface as "rigorous" or strong may be weak-kneed underneath. Our answer to these issues was to lay them out as qualifications and speculate on their implications. The ideal, at least for us, is to have the findings promote broader discussion and expose deeper questions—and in the process avoid a superficial interpretation of the experiment's results.

Supplemental mental health treatment study

Research design and revision

Despite the challenges in the previous experiment, we were emboldened by the successful random assignment and track-record at the research site to embark on another more ambitious experiment. In our previous research on predicting batterer program outcomes (Jones et al. 2004), mental health problems (based on results of the MCMI-III) were strongly associated with program dropout and re-assault during follow-up. To test the effect of assessment and referral on batterer program outcomes, we developed an experimental evaluation of what we termed "supplemental mental health treatment." This treatment was to be in addition to the required batterer program attendance under a court mandate for a domestic violence offense.[12]

The experiment was to proceed as follows: men referred to the batterer program would complete the Brief Symptom Inventory (BSI) at the centralized intake session, along with a background questionnaire and other introductory information. At the

[12] This "supplemental" approach was used since it reflects the assessment and referral recommendations in the majority of state standards for batterer programs. As we discuss in the study reports (e.g., Gondolf 2009a), other approaches may be more effective. For instance, individualized counseling for some men within the batterers program or integrated behavioral health treatment within the mental health clinic are other options to consider. They both offer the advantage of a more streamlined approach with less chance of non-compliance. The referral approach in our study reflects the assessment and referral recommendation in state standards from batterer programs and the inter-agency ideal of a coordinated community response.

following week's orientation meeting, the men were individually assigned to an on-going counseling group (based on their schedule and residence location). They were also advised to make an appointment for a psychological evaluation at the one of two mental health clinic locations, if they had scored positive on the BSI.

The experimental group was to consist of a random sample of the "positive" men who were informed that the evaluation and any prescribed treatment was part of the court requirement; failure to obtain an evaluation and treatment would constitute non-compliance to the batterer program and result in further court sanctions. The control group would be the other men who screened positive but were told they could seek an evaluation and treatment voluntarily.[13] Considering the lack of compliance to previous voluntary referrals, it was unlikely that these men would seek a clinic evaluation and thus could serve as a "non-treatment" group.

There was, of course, an elaborate referral system behind all of this. In order to develop the referral procedures, administrators from all of the collaborating agencies and research leaders met regularly over a six-month period, communicated through emails, and discussed details over the phone. Forms were produced for the intake staff, the subjects, and the clinic, along with a protocol for verifying compliance and reporting non-compliance to the court. The subjects, following a simple checklist of instructions, were to make an appointment and appear for an evaluation. They would present an evaluation verification form to the clinic evaluator, who in turn would fax the completed form to the batterer program office. The office staff would enter the information into the computer database, and report the compliance status to the court liaisons for the periodic court review of the case. The court would issue additional sanctions to those men who failed to comply either to the batterer program attendance or to "required" mental health treatment.

Despite what we thought was a masterful referral plan, the experimental design was soon to fall apart. We encountered a variety of inter-agency breakdowns and intra-agency problems. Not only was the random assignment aborted, but referral compliance was minimal. We considered closing the project, but reverted instead to a quasi-experimental design based on the phases of referral implementation: a phase of voluntary referral ($n=182$) to serve as a quasi-control group, followed by a phase of mandated referral serving as the "treatment" ($n=148$). As it turned out, a transitional phase ($n=149$) also emerged as we negotiated a number of interagency breakdowns in the mandated referral. Some men were not being properly notified about the supplemental mental health at intake, the sending of the verification forms was inconsistent, the batterer program office forgot to notify the court liaisons of referral non-compliance, and some of the judges chose to accept batterer program completion as sufficient.

In the midst of negotiating these problems, we installed a system coordinator and case manager who monitored each mandated subject and in the process ensured the procedures were followed. The one area where inconsistencies remained was in the court. Some judges and prosecutors continued to overlook the referral non-compliance for practical reasons. They wanted to move cases along, and accepted

[13] The voluntary referral was recommended by our Institutional Review Board, so as not to withhold the results of the BSI test and referral possibilities from them. Those who did voluntarily comply would be deleted from the control group.

some men's claims that the mental health referral was too much of a burden or not needed.[14] With the system coordinator and case manager in place, compliance to the mandated referral increased from less than 5% of the voluntary referrals to 32% of the mandated referrals (or approximately 55% of the men who completed the program). Under the mandate, over 60% had made an appointment for an evaluation, but only 19% received treatment (58% of those evaluated). Interestingly, most of those not receiving treatment were diagnosed as "adjustment disorders" not warranting further treatment; those receiving treatment were primarily men with depressive disorders.

Unforeseen disruption

What when wrong? Unlike resistance faced in previous experiments, there was ready agreement from the batterer program, the domestic violence court, the major mental health hospital, the prosecutors' office, and the battered women's center. The administrator of the mental health clinic that received referrals had worked on projects for battered women and was very eager to assist us. The batterer program director saw the project as not only a way to possibly improve outcomes, but also as leverage for soliciting additional program funds. A local foundation did in fact oblige with money for operations and also for treatment costs. Research colleagues, at the teaching hospital affiliated with the clinics, were also on board as consultants. As with our relatively successful culturally-focused experiment, resistance may have been eased again because treatment was being added, so to speak, rather than withheld, and because of the interest in the issue and the opportunity for inter-agency collaboration.

The undoing of the random assignment was largely due to "uncontrollable events." A series of unforeseen disruptions besieged the intended experiment from the outset. Shortly before the research was scheduled to begin, the human subjects committee at the clinic's affiliated hospital withdrew its approval. The committee decided that the established consent procedures for interviewing the subjects' partners were inadequate. We had prior "human subjects" approval from our university for this study and approval for the same procedures at several other sites in the past (Gondolf 2000a). Our university's committee disagreed with the dissenting hospital committee, and we were left to reconcile the difference.

After that delay, the prosecutor announced that his office could no longer support the experiment due to a high-profile murder that involved a batterer program dropout who stalked and shot his wife while she was in church. He did not want the appearance of mental health treatment being withheld from anyone who knowingly needed it (i.e., the control group). Yet he did find the phased introduction of mental health referral to be acceptable. But even if we were to negotiate a way to proceed with random assignment, it would have been impractical for another reason. The batterer program executive director was exposed as embezzling program funds and committing medical insurance fraud. He was brought under criminal investigation

[14] Men who dropped out of the batterer program also tended not to comply with the mental health referral and were consistently sanctioned.

and fired from his program position. He had in the process also taken some of the operational grant, so we were left trying to find replacement funds for that.

Additionally, the administrator of the clinic left for a few months due to illness, and her temporary replacement was not sufficiently versed nor committed to the project. She was the one taking calls for appointments and coordinating the referrals; without her, many of the referrals fell through the cracks. We also faced turnover in the prosecutor's office. The prosecutor in charge of the domestic violence court was promoted and his replacement was unfamiliar with our project. Last but not least, the court administrator who had assisted us for so many years was dismissed under a city-wide reorganization of the courts.

This laundry list of staff turnover and organizational complications raises the question of whether we had been lucky in our past studies, or just avoided such disruption through a narrow focus. We did lose the initial and preferred counselor in the culturally-focused counseling experiment, and encountered turnover of staff in the subsequent case-management project (Gondolf 2008a). In the mental health treatment study, part of the problem was the fluidity of staff in human services and criminal justice system in general but also the scope of the project and agencies involved in it. We did manage to adjust and continue with the quasi-experimental as a back-up, but not without our frequent intrusion into the setting for retraining, promptings, and negotiations. With the installment of the system coordinator and case manager, we became a "player" in the research setting and were restructuring it to some extent.

Organizational issues

Another set of problems was intra-organizational in nature. There were, for instance, several "fixable" logistical issues in the evaluation of mental health treatment, many of which have already been mentioned. Despite training and lucid protocols, the intake staff occasionally failed to clearly notify the subjects of the mandated referral. They were distracted by participant questions about the program, scheduling conflicts with some of the group assignments, or the volume of men that had to be processed at a given intake session. Intake sessions were occasionally cancelled because their meeting rooms were occupied by other organizations. Some men missed the orientation session where they were to be notified about mental health referral.

Part of the problem was simply system overload that weakened attention to our project demands. Calls to the clinic for a mental health evaluation were not always returned due to the heavy patient load. Irrespective of our referrals, the clinic received over 100 calls a day for appointments and consultation, and clinicians averaged 15 clients a day in individual sessions. The batterer program itself had 80–100 intakes a month and over 20 on-going group sessions at seven different sites around the city.

Another set of familiar problems was the inter-organizational ones that tend to surface in criminal justice collaborations. The agencies involved in our study differed in orientation and priorities. Consequently, there was sometimes a clash of purposes, assumptions, or expectations, but without outright competition or conflict. Probably the most obvious, in this regard, was the difference in the punitive orientation of the court and the accommodating approach of the clinic. The court expected and required the referred men to obtain a mental health evaluation and

assumed that coercion from possible sanctions was sufficient to make them complete that task.

The clinicians, on the other hand, tended to rely on their clients' wanting help or treatment, according to the clinic administrator. They are accustomed to clients who are motivated to present and discuss their mental health problems and needs. The clinicians, moreover, were reportedly reluctant to be involved in court-mandated cases because of the time, persuasion, and documentation they require. These sorts of differences did not directly affect the aim of random assignment, but they likely added to the lack of attention to the referral protocols and lack of urgency to revise and fix breakdowns.

Beyond "don't work"

As reviewers of the reports rightly point out, the failure to implement the intended experiment makes it difficult to assess the effectiveness of supplemental mental health treatment. The low number of men who did comply with the referral had more positive outcomes, but those outcomes are obviously confounded by subject motivation, diagnoses, and other characteristics. The major finding was the extent of those warranting evaluation, and the lack of compliance to clinic referral for a clinic evaluation.

Nonetheless, our practitioner consultants and agency administrators remained very supportive of the referral plan. One reviewer of the final report emphasized, the main contribution of the study is "the real life limitations and problems" associated with program innovation. The difficulties in maintaining collaboration are particularly instructive given the on-going attention to "community coordinated responses." They also raise further questions about interpreting the successful experiments of batterer program effects. It seems important to monitor the larger system performance and its possible impacts on the studied "treatment." For example, the compliance to the mental health referrals was lowered in part because of inconsistent sanctions from the courts.

The research, as it was, seemed to point beyond a "don't work" bottom-line. For instance, our consultants, agency administrators, and report reviewers all pointed to one recommendation in particular: the referral system needed to be streamlined. More agencies mean more breakdowns, was the adage. Our journal articles on the topic, therefore, cited examples of integrated services within the batterer program and specialized clinic units for domestic violence which have reduced referral complications. Overall, our supplemental mental health study dramatized, at least to this researcher, the limits of imposing an experiment on the chaotic world that surrounds it. It lends credence to the oft times critique that field experiments remain, for the most part, an artificial construct on a complex phenomena (Hope 2009).

Discussion

Summary

The ultimate purpose of this paper is not to dismiss experimental evaluations, but rather to raise cautions about over interpreting the existing ones of batterer programs.

In the two case studies discussed here, we also attempt to draw lessons that might help improve the interpretation and application of the existing batterer program evaluations, and also raise recommendations for enhancing evaluation efforts. The findings also reinforce the encouragement to more fully report the implementation issues that surface during experimental evaluations, especially in programs that involve the criminal justice system and multiple agencies (Durlak and DuPre 2008; Smedslund et al. 2007). In a sense, the backstory examples point to a popular theme throughout public decision making—namely, the need for more transparency and disclosure.

As suggested in the previous implementation reviews, as well as this article's case studies, the unexpected challenges that face experimental designs do not get fully discussed, assessed, and factored into the interpretations of the findings. This may be related, in part, to a lack of clear guidelines for reporting and the external demand for bottom-line findings. Practitioners and policy makers want to know whether or not a program "works." The complicated and extensive details of a study are therefore of less interest. If the few available backstories are any indication of what other experimental evaluations have encountered then more of those details are warranted—and with them more qualifications of the results and cautions about their implications.

In order to more systematically examine the extent and nature of the implementation problems across the available experiments, one reviewer recommended CONSORT Tables (Moher et al. 2001). These would go well beyond the shorter tabulation of issues in the Cochrane study mentioned at the outset (Smedslund et al. 2007). CONSORT Tables are based on the Consolidated Standards for Reporting Trails (CONSTORT) introduced in 1996 to improve reporting of experimental clinical trials particularly in the medical field. Such tables would enable a clearer summary of the strengths and weaknesses of the existing studies across 22 implementation issues that include randomization, intention to treat, effect size, conflict of interest, withdrawal and dropouts, and adverse events. They not only identify whether the factors were addressed but also the quality of the overall reporting. Moreover, evaluation researchers in other fields have put forward a variety of instruments to gauge and measure such things and indicate the extent of implementation factors that may influence outcomes (e.g., Fletcher et al. 2009; Melnick et al. 2009). Such instruments can be used as a means of monitoring, for example, staff support and agency collaboration, as well assessing their impact upon completion of the study.

Major lessons

Our two attempted experiments help to confirm the challenges identified in previous implementation of batterer program evaluations (Feder et al. 2000; Visher et al. 2008a, b). The most obvious of these has to do with random assignment. The successful assignment in our experiment of culturally-focused counseling was in a large part due to making the assignment at the point of program intake—outside the court—and with program authority over the assignments. This procedure was enabled in both studies by the experimental condition being an enhancement to the basic treatment rather than a withholding of it. The extensive research previously

conducted at the site also helped to established relationships and familiarity with research demands—a level of research "readiness" was in place.

The failure of the random assignment in the study of supplemental mental health treatment was due primarily to a series of unforeseen disruptions, what previous researchers referred to as "uncontrollable events" (Feder et al. 2000). It also suggests that the disruptions can sometimes expose contextual issues that blunt the best laid plans of researchers and practitioners. How do we contain or avoid such issues not only during an experiment, but also before and after one as well? After both our studies, practitioners went their own way irrespective of the findings, as happens in a substantial portion of such projects (Rogers 2003). System stability, financial resources, and sufficient personnel were not available to continue or further develop the "innovation" projects.

The intra- and inter-organizational issues are another matter that probably warrant study in themselves. This is especially the case with batterer programs that are largely community-based projects with variations in professionalization and administrative structure. Some of the resistance and tension that emerged in the first study may have been a by-product of the "organizational culture." The limited accommodation of the experimental option was attributable in part to the batterer program's linkage to the court, established procedures and assumptions, and lines of authority and respect, as well as the impact of different personalities and expectations. The underpaid frontline staff also perceived an extra burden that an experiment brings—more paper work, training, and oversight.

The interagency breakdowns in the second study appear to have less to do with overt resistance than the inertia and reality of agency operations. Administrative turnover, other priorities, heavy caseloads, and staff prerogatives played a part in the non-compliance to the mental health referrals. The lesson here might be not only the need for streamlined and integrated services, but also for structural reforms that bridge the organizational gaps. The addition of the system coordinator and case manager, for instance, increased compliance markedly. This sort of effort at reform at least increased service delivery in the Judicial Oversight Demonstration project (Visher et al. 2008a, b).

The most fundamental problems throughout are those associated with random assignment amidst the complex 'real world' of the criminal justice system. Experienced experimental researchers suggest some specific strategies to deal with what they acknowledge as "inevitable challenges" in this area (e.g., Berk 2005; Davis and Taylor 2001). Much of it boils down to flexibility and contingency plans. It appears in the available case studies that experience, creativity, negotiation, and cleverness are also essential ingredients. One has to consider not only the integrity of the experiment but also the politics of the situation, the opinions of program staff, the responsiveness of the subjects, and the funding and timeframe of the study.

One design alternative we posed in our study of supplemental mental health treatment was a kind of block random assignment. That was to alternate assigning the program enrollees for one month to the experimental group and the next month to the control group over the course of a year. This would have been less disruptive and confusing to the court, the program, and the subjects themselves, than randomly assigning all the men at a particular intake session. It would have made it easier to supervise and monitor the assignment, and thus make it less vulnerable to problems.

Research recommendations

The experimental program evaluations in the domestic violence field may not only be a matter of communicating the utility of experimental designs but also of advancing the knowledge on how to implement them in the unique circumstances of domestic violence cases. From the brief examples we have so far, it is clear that each field experiment faces complications of its own, including a variety of uncontrollable events. One obvious area of concern is establishing and maintaining agreement among all affected parties, along with building community support (Feder et al. 2000). It may be helpful to derive protocols to address some of the familiar complications, and contingency plans for an experiment disrupted and undone by unforeseen circumstances.

One lesson from our studies is to reaffirm agreements and understandings periodically in writing and in person, and retrain and revitalize frontline staff on a regular basis. Also, an on-going monitoring system of the implementation may help in identifying problems and remedying them. For instance, our review of non-compliance to random assignment helped to identify reasons for the shortcomings and move to correct them. The "system coordinator" collected the information and negotiated with representatives of the participating agencies to solve problems, as well as called subjects to help negotiate obstacles to their compliance in the mental health study. Frequent progress reports to the agency representatives and frontline staff may, furthermore, encourage support, answer doubts, and thwart rumors.

As far as dealing with the unforeseeable, the fallback position is some form of quasi-experimental design as we attempted in our mental health study. Admittedly, the advantages of an experimental design are sacrificed, but matching devices for comparison samples and computer modeling, with expanded sample size, are ways to establish further controls (Angrist 2005). But there may, as well, be deeper lessons raised by disruptions that are also worth chronicling. In our mental health study, for instance, the influence of the prosecutor was very much confirmed by his response to a high profile murder, and batterer program stability was evident in its continuation following the dismissal of the director. Our phased implementation also exposed the strength and weakness of our referral system. A formative or process evaluation in this case showed not only what does not work but also how to make it work (Gondolf 2009b).

A more immediate need, however, may be more extensive reporting of implementation issues and problems, especially since they appear so pervasive among the batterer program evaluations. As we suggested at the outset, this practice would help to build a base of information about implementation that could help in anticipating and avoiding pitfalls. It would also be useful in formulating implementation protocols and contingency plans. Perhaps even more importantly, the implementation disclosures would help to gauge the integrity of the design and extraneous influences on the results. Such information would also help the reader to interpret the results and to further appreciate the influential context and complexities of domestic violence intervention (Hanson et al. 2009; McCarty 2009; Marlowe et al. 2006; Mears 2003). It might also increase the understanding of what it takes to conduct program evaluations of batterer programs and program innovations.

Revising existing interpretations

Our concluding impression is a more cautious one than that of the researchers of other batterer program evaluations. We obviously see benefits and contributions of the existing experimental evaluations and recognize the need for future ones. More needs to be done to replicate experiments at a wider variety of existing and evolving programs. Many batterer program staff argue that the few existing evaluations simply do not apply to them, especially given the diversity of program approaches, the difference in settings, advancing assessment of offenders, linkages to the court and other services, and variations in offender demographics.[15] They often point to the replications of the Minneapolis Police Domestic Violence experiment that produced varied results at different sites (Sherman 1992).

However, we also are concerned about the implementation impacting experiment results. Despite the exceptional implementation of our culturally-focused study, the results are debated because of staffing issues, internal conflicts, and external linkages. Similar questions could probably be raised about the implementation of other batterer program evaluations based on the current implementation studies. If so, what does that do to the apparent "verdict" that batterer programs don't work, and the consequences that that is bringing to batterer programs? Our experience would say that the bottom-line claims warrant some revision (see Sherman and Strang 2004; Weisburd et al. 2003).

Moreover, the difficulties of implementing experiments in this field, no doubt, account in part for their fewness. They also suggest the need for other research designs and approaches to supplement and confirm experiments. There appears to be a need for studies that show what program features are working and with whom—in other words, studies that produce a differential outcome. Propensity score analysis is an increasingly popular approach in this regard. It statistically simulates balanced comparison groups within non-experimental data (e.g., Jones et al. 2004) to compute outcomes for subgroups of subjects in contrast to the average effect of the experimental outcomes. Also, instrumental variable analysis is being used to control for contextual factors and address the policy concern about the effect of actually receiving treatment ("intention to treat" examined by the experiments is confounded by high levels of the non-compliance to the treatment; e.g., Jones and Gondolf 2001).

Both these approaches require larger sample sizes and more background variables than are typical in experiments, but they preclude the implementation problems associated with experiments through their reliance on non-experimental data. The "low quality" of the existing batterer program experiments identified in the Cochrane meta-analysis (Smedslund 2004), and the additional implementation challenges exposed here, reinforce the utility of statistical modeling in criminal justice evaluations (Angrist 2005; Loughran and Mulvey 2010). They also add weight to

[15] The influence of the organizational context might also be examined separately considering its apparent impact on intervention outcomes, as suggested in the current sexual assault meta-analysis (Hanson et al. 2009), our study of court mandated review (Gondolf 2000b), and other studies in the criminal justice field (e.g., McCarty 2009). To accomplish this, a fuller disclosure and examination of experimental implementation is obviously needed.

the competing findings from propensity score and instrumental analyses conducted thus far on batterer program outcomes (Jones et al. 2004; Jones and Gondolf 2001).

Clinical trials with several treatment options at multiple sites, like those conducted with alcohol and depression treatments, are still the ideal (Elkin et al. 1989; Project MATCH Research Group 1997). The National Institute on Drug Abuse Clinical Trial Network (CTN) has, for instance, helped to facilitate these sorts of trails at regional research-ready sites. It also provides problem-solving for implementation as well as the infrastructure to support experiments. The cost and requirements of such an approach have limited its use in the criminal justice field thus far. Multi-site clinical trials would, however, advance the experimental research of batterer programming and intervention, and enable an examination of the impact of context and diverse populations on the outcomes.

References

Angrist, J. (2005). Instrumental variables methods in experimental criminological research: What, why, and how? *Journal of Experimental Criminology, 1*, 23–44.

Berk, R. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1*, 416–433.

BISCMI. (2009). *Coordinated community responses: Are they effective? The reality, research, and results behind the Judicial Oversight Project. Conference announcement and program*. Ann Arbor, MI (www.biscmi.org/jod/BISC-MI_2009_National_Conference_Program.pdf).

BWJP. (2008). Training opportunity: Audio conference on violence against women. *Research/Practitioner discourse/ Violence Against Women*, *14*, 732-733 (vaw.sagepub.com/cgi/reprint/14/6/732.pdf).

Cissner, A., & Farole, D. (2009). *Avoiding Failures of Implementation: Lessons from Process Evaluation*. Washington: Bureau of Justice Assistance.

Corvo, K., Dutton, D., & Chen, W. (2008). Toward evidence-based practice with domestic violence perpetrators. *Journal of Aggression, Maltreatment, and Trauma, 16*, 111–130.

Davis, R., & Taylor, R. (2001). Does batterer treatment reduce violence? A synthesis of the literature. *Women & Language, 24*, 69–93.

Davis, R., Taylor, B., & Maxwell, C. (1998). *Does batterer treatment reduce violence? A randomized experiment in Brooklyn*. Final report to the National Institute of Justice, Washington, DC.

Dobash, R. E., & Dobash, R. P. (2000). Evaluating criminal justice interventions for domestic violence. *Crime and Delinquency, 24*, 252–270.

Dunford, F. (2000). The San Diego Navy Experiment: An assessment of interventions for men who assault their wives. *Journal of Consulting and Clinical Psychology, 68*, 468–476.

Durlak, J., & DuPre, E. (2008). Implementation Matters: A review of research on the influence of implementation on program outcomes and the factors affective implementation. *American Journal of Community Psychology, 41*, 327–350.

Elkin, I., Shea, M. T., & Watkins, J. T. (1989). The National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46*, 971–982.

Feder, L., & Boruch, R. (2000). The need for experiments in criminal justice settings. *Crime and Delinquency, 46*, 291–294.

Feder, L., & Dugan, L. (2002). A test of the efficacy of court-mandated counseling for domestic violence offenders. *Justice Quarterly, 19*, 343–376.

Feder, L., Jolin, A., & Feyerherm, W. (2000). Lessons from two randomized experiments in criminal justice settings. *Crime and Delinquency, 46*, 380–400.

Fletcher, B., Lehman, W., Wexler, H., Melnick, G., Taxman, F., & Young, D. (2009). Measuring collaboration and integration activities in criminal justice and substance abuse treatment agencies. *Drug and Alcohol Dependence, 1003*(Supplement 1), 54–64.

Goldkamp, J. (2008). Missing the target and missing the point: "successful random assignment but misleading results. *Journal of Experimental Criminology, 4*, 83–115.

Gondolf, E. (1997). *An Experimental Evaluation of Child Abuse Prevention Classes for Court-ordered Woman Batterers. Final report submitted to Child Trust Fund of Pennsylvania*. Harrisburg: Pennsylvania Department of Welfare.

Gondolf, E. (2000a). Human subject issues in batterer program evaluation. *Journal of Aggression, Maltreatment, and Trauma, 4*, 273–297.

Gondolf, E. (2000b). Mandatory court review and batterer program compliance. *Journal of Interpersonal Violence, 15*, 437–438.

Gondolf, E. (2001). Limitations of experimental evaluations of batterer programs. *Trauma, Violence & Abuse, 2*, 79–88.

Gondolf, E. (2002). *Batterer Intervention Systems: Issues, Outcomes, and Recommendations*. Thousand Oaks: Sage Publications.

Gondolf, E. (2005). *Culturally-focused batterer counseling for African-American Men: A clinical trial of effectiveness*. Final report submitted to the National Institute of Justice, Washington, DC.

Gondolf, E. (2007). Culturally-focused batterer counseling for African American men: A clinical trial of re-assault and re-arrest outcomes. *Criminology and Public Policy, 6*, 341–366.

Gondolf, E. (2008a). Implementation of case management for batterer program participants. *Violence Against Women, 14*, 208–225.

Gondolf, E. (2008b). Outcomes of case management for African American men in batterer counseling. *Journal of Family Violence, 23*, 173–181.

Gondolf, E. (2009a). Outcomes from referring batterer program participants to mental health treatment. *Journal of Family Violence, 24*, 577–588.

Gondolf, E. (2009b). Implementing mental health treatment for batter program participants: Interagency breakdowns and underlying issues. *Violence Against Women, 15*, 638–655.

Gondolf, E., & Jones, A. (2001). The program effect of batterer programs in three cities. *Violence and Victims, 16*, 693–704.

Han, C., Kwak, K., Marks, D., Pae, C., Wu, L., Bhatia, K., et al. (2009). The impact of the CONSORT statement on reporting of randomized clinical trials in psychiatry. *Contemporary Clinical Trials, 30*, 116–112.

Hanson, K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior, 36*, 865–891.

Hollin, C. (2008). Evaluating offending behaviour programmes: Does only randomization glister? *Criminology and Criminal Justice, 8*, 89–106.

Hope, T. (2009). The illusion of control: A response to Professor Sherman. *Criminology and Criminal Justice, 9*, 125–134.

Jones, A., & Gondolf, E. (2001). Time-varying risk factors for reassault by batterer program participants. *Journalof Family Violence, 16,* 345–359.

Jones, A., D'Agostino, R., Gondolf, E., & Heckert, A. (2004). Assessing the effect of batterer program completion on reassault using propensity scores. *Journal of Interpersonal Violence, 19*, 1002–1020.

Kober, T., Trelle, S., & Engert, A. (2006). Reporting of randomized controlled trials in Hodgkin Lymphoma in biomedical journals. *Journal of the National Cancer Institute, 98*, 620–625.

Labriola, M. (2005). *Testing the impacts of court monitoring and batterer intervention programs*. Paper presented at the Annual Meeting of the American Society for Criminology, Toronto, Canada, November, 17.

Labriola, M., Rempel, M., & Davis, R. (2008). Do batterer programs reduce recidivism? Results from a randomized trial in the Bronx. *Justice Quarterly, 25*, 252–282.

Levesque, D. A., Driskell, M. M., & Prochaska, J. M. (2008). Acceptability of a stage-matched expert system intervention for domestic violence offenders. *Violence and Victims, 23*, 432–445.

Loughran, T., & Mulvey, E. (2010). Estimating treatment effects: Matching Quantification to the question. In A. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 163–181). New York: Springer.

Marlowe, D. B., Festinger, D. S., Lee, P. A., Dugosh, K. L., & Benasutti, K. M. (2006). Matching judicial supervision to clients' risk status in drug court. *Crime and Delinquency, 52*, 52–76.

McCarty, D. (2009). Understanding the importance of organization and system variables on addiction treatment services within criminal justice settings. *Drug and Alcohol Dependence, 103*, 91–93.

Mears, D. (2003). Research and interventions to reduce domestic violence revictimization. *Trauma, Violence & Abuse, 4*, 127–147.

Melnick, G., Ulaszek, W., Lin, H., & Wexler, H. (2009). When goals diverge: Staff consensus and the organizational climate. *Drug and Alcohol Dependence, 103*(Supplement 1), 17–22.

Moher, D., Schulz, K., & Altman, D. (2001). The CONSORT Statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association, 285*, 1987–1981.

Morrison, S., Lindquist, C., Hawkins, S., O'Neil, J., Nesius, J., Mathew, A. (2003). *Evidence-based review of batterer intervention and prevention programs*. Final report to the Centers for Disease Control and Prevention (CDC), Atlanta, GA.

Murphy, C., & Ting, L. (2010). Interventions for perpetrators of intimate partner violence: A review of efficacy research and recent trends. *Partner Abuse, 1*, 26–44.

Peterson, R. (2008). Reducing intimae partner violence: Moving beyond criminal justice interventions. *Criminology and Public Policy, 7*, 537–545.

Potter, H. (2007). Reaction Essay: The need for a multi-faceted response to intimate partner abuse perpetrated by African-Americans. *Criminology and Public Policy, 6*, 367–376.

Project MATCH Research Group. (1997). Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcome. *Journal of Studies on Alcohol, 58*, 7–29.

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Saunders, D. (2008). Group interventions for men who batter: A summary of program descriptions and research. *Violence and Victims, 23*, 156–172.

Sherman, L. (1992). Policing domestic violence: Experiments and dilemmas.

Sherman, L. (2009). 'Evidence and Liberty: The Promise of Experimental Criminology. *Criminology and Criminal Justice, 9*, 5–28.

Sherman, L., & Strang, L. (2004). Verdicts or interventions?: Interpreting results from randomized controlled experiments in criminology. *The American Behavioral Scientist, 47*, 575–607.

Smedslund, J. (2004). *Dialogues about a new psychology.* Chagrin Falls, OH: Taos Institute

Smedslund, G., Dalsbo, T., Steiro, A., Winsvold, A., Clench-Aas, J. (2007). *Cognitive behavioural therapy for men who physically abuse their female partner. The Cochrane Databasae of Systematic Reviews, Issue 4, Article No. CD006048*. (available at www.cochranelibrary.com).

Visher, C., Newmark, L., & Harrell, A. (2006). *Final report on the evaluation of the Judicial Oversight Demonstration (volume 2): Findings and lessons on implementation*. Washington: Urban Institute.

Visher, C., Harrell, A., & Yahner, J. (2008a). Reducing intimate partner violence: An evaluation of a comprehensive justice system-community collaboration. *Criminology and Public Policy, 7*, 495–523.

Visher, C., Newmark, L., & Harrell, A. (2008b). *The Evaluation of the Judicial Oversight Demonstration: Findings and Lessons on Implementation*. (National Institute of Justice: Research for Practice), Washington: National Institute of Justice, U.S. Department of Justice (http://www.ojp.usdoj.gov/nij/pubs-sum/219077.htm).

Weisburd, D. (2003). Ethical practice and evaluation of interventions in crime and justice: The moral imperative for randomized trials. *Evaluation Review, 27*, 336–354.

Weisburd, D., Lum, C., & Yang, S. (2003). When can we conclude that treatments or programs "don't work"? *The Annals of the American Academy of Political and Social Science, 587*, 31–58.

Edward W. Gondolf EdD, MPH, is research director for the Mid-Atlantic Addiction Research and Training Institute (MARTI) and professor of sociology at Indiana University of Pennsylvania. He conducts grant-funded research on the response of the courts, mental health practitioners, alcohol treatment clinicians, and batterer treatment programs to domestic violence. He is the author of numerous articles and books on these topics, including *Assessing Women Battering in Mental Health Services*; *Batterer Intervention Systems: Issues, Outcomes, and Recommendations*; and *The Future of Batterer Programs Amidst "Evidence-Based Practice."*