J Exp Criminol (2011) 7:201–224
DOI 10.1007/s11292-011-9122-z

Missing links: how descriptive validity impacts the policy relevance of randomized controlled trials in criminology

Charlotte E. Gill

Published online: 16 April 2011

Abstract
Objectives To assess quality of reporting of issues that may affect internal and external validity in randomized controlled trials (RCTs) in criminology, and explore the impact of reporting quality (descriptive validity) on the policy relevance of rigorous research.
Methods Reporting indicators based on CONSORT standards from the health sciences are constructed and applied to a sample of 38 RCTs, covering a range of criminal justice interventions, published in journals between 2002 and 2008. A Descriptive Validity Matrix is constructed to visually convey information about reporting quality across a group of studies, based on the reporting indicators, to decision-makers.
Results Criminological RCTs are moderately well-reported. The sample of studies show medium descriptive validity in reporting on elements relevant to internal validity, and high descriptive validity for items relevant to external validity. However, there was considerable variation in the quality of reporting on key issues, especially those related to implementation of the random assignment sequence, deviations from the planned study, and attrition of participants.
Conclusions This study and the Descriptive Validity Matrix provide a useful framework for assessing descriptive validity. Although the indicators developed were not specific to criminology, and the analysis was limited to a small number of studies published in academic journals, this study is an important starting point for continued research and discussion on the relationship between implementation of field experimentation, reporting quality, and policymaking. The ability to report research clearly is as important as choosing the most rigorous research design for enhancing the objectives of evidence-based crime policy.

Keywords CONSORT · Descriptive validity · Evidence-based policy · External validity · Internal validity · Methodological quality · Randomized controlled trials

C. E. Gill (✉)
Center for Evidence-Based Crime Policy, George Mason University,
4400 University Dr., MS 6D3, Fairfax, VA 22030, USA
e-mail: cgill9@gmu.edu

The quality of program evaluations, and its impact on their policy relevance, is a key component of the discourse on 'what works' in crime and justice (e.g., Sherman et al. 1997; Boruch et al. 2000; MacKenzie 2000; Weisburd et al. 2001; Farrington 2003a; Lum and Yang 2005; Sherman et al. 2006). Academics and policymakers alike are increasingly calling for better evidence. This has led to a focus on the randomized controlled trial (RCT) as the "gold standard" research design for maximizing internal validity (e.g., Berk and Rossi 1999: 20-21; Farrington 2003b). Internal validity, or the extent to which causal inferences about program effectiveness may be drawn from a given study, is considered one of the most important of four criteria proposed by Cook and Campbell (1979) for methodological quality (Farrington 2003a; Shadish et al. 2002). Another important criterion, external validity, while not necessarily maximized by a randomized design, is nonetheless crucial to evidence-based public policy because it indicates the extent to which a program's outcomes may be applicable to settings and populations different from those under which it was tested (see also Berk and Rossi 1999: 22).[1]

Despite the extent of the recent interest in methodological rigor and the increased use of RCTs, debate around the quality of evaluation research in criminal justice has not subsided. The RCT is not, in itself, a guarantee of validity. One of the most prominent recent critiques, from the United States General Accounting Office (GAO), indicated that the majority of evaluations managed by the National Institute of Justice over the preceding 10 years, *even those deemed "sufficiently designed,"* were so beset with methodological and implementation problems that it was difficult to "draw meaningful conclusions about the programs' effectiveness" (reported in Lauritsen 2006: 365). There are numerous barriers to successful experimental research in criminal justice populations and settings. Attrition of participants (both pre- and post-random assignment) can occur in any research study involving human subjects, leading to biased results and limited external validity, but may be more pronounced in the frequently "risky" and "less accessible" subjects who come into contact with the criminal justice system (e.g., Goldkamp 2008: 86). The politicized, bureaucratic nature of many criminal justice agencies can create practical and financial obstacles to access and the implementation of research projects. Overall, practitioners in the criminal justice arena are not committed to a tradition of experimental research and practice to the same extent as in other disciplines, such as health (Shepherd 2003). Complex ethical concerns about the potential risks to subjects and the public of denying (or mandating) treatment may hinder the design and implementation of true experiments, particularly when experienced practitioners believe strongly in the effectiveness of an intervention (Weisburd 2003; Farrington and Welsh 2005). With such deep-set structural factors apparently limiting the production of high-validity research, how can criminal justice policymakers decide what really constitutes the 'best' evidence for guiding practice?

One (relatively) simple fix among a host of suggestions put forward by Lipsey et al. (2006: 295) in response to the GAO report is the proposal that evaluators make full results and technical details about trials available to the research and policy communities. Lipsey and colleagues argue that this would facilitate discussion on how to improve evaluation methods and practice. It could also be useful in assisting policymakers and

[1] The third and fourth forms of validity described by Cook and Campbell (1979) are statistical conclusion validity (the relationship between cause and effect, and the ability of the research design to identify it), and construct validity (the adequacy of operational definitions and measurement).

other research consumers to sort the good evaluations from the poor when making judgments about what works. This proposal requires a sharper focus on another type of validity not discussed by Cook and Campbell: *descriptive validity*. Farrington (2003a: 55) defines descriptive validity as "the adequacy of the presentation of key features of an evaluation in a research report… such as the number of participants and the effect size." Farrington places descriptive validity second only to internal validity in terms of importance for assessing the quality of a trial (ibid.: 61).

Descriptive validity has a strong relationship with both internal and external validity (Perry 2010: 333). If an RCT is to be held up as the "gold standard" of internally valid research, transparent reporting is crucial. Sufficient evidence must be provided that the experiment was designed, implemented, and analyzed such that internal validity is truly maximized. If this validity is compromised in any way, a detailed explanation must be provided to allow the research consumer to judge the extent to which the evaluation remains of satisfactory quality to be considered relevant to policy. Furthermore, if the results are to be meaningful to policymakers under any circumstances, study authors must provide details not only of the intended target population for the program, but the characteristics of those who received it (and where and when), so that outcomes may be generalized for implementation on the broader scale.

In terms of policy relevance, descriptive validity is also crucial to the discipline of systematic review and meta-analysis (for example, the work of the Campbell Collaboration[2]), which seeks to distill rigorous evidence on a particular intervention into statements and measures of overall effectiveness for the benefit of decision-makers. Transparent and detailed reporting of studies is necessary not only to ensure comparability between the programs included in a review, and for authors to make judgments about methodological rigor, but also for the calculation of meta-analytic effect sizes. Inconsistent reporting of results across different evaluations of the same intervention may prevent the meta-analyst from calculating comparable effect sizes, thus limiting the pool of studies that can be meaningfully combined.[3] Ultimately, this prevents systematic review authors from making the strong, unequivocal statements of effectiveness policymakers want to hear, which in turn damages the policy relevance of criminal justice research. In this and all the other ways described above, the concept of descriptive validity clearly represents the 'missing link' in the process of translating the best research evidence into policy and practice.

Despite its importance, the issue of descriptive validity is less often discussed in the field of criminology (cf. Lösel and Köferl 1989; Boruch 1997; Farrington 2003a; Petrosino et al. 2006). This contrasts sharply with the health and medical sciences, in which the development of quality standards for the reporting of trials has been widely discussed and advanced over the past 15 years. The Consolidated Standards of Reporting Trials (CONSORT), first set out by the CONSORT group in 1995 and revised in 2001 (Altman et al. 2001), currently consists of a 22-item checklist of trial characteristics to be reported. These standards have been adopted by many leading medical journals, including the *Journal of the American Medical Association,* the *British Medical Journal*, and *The Lancet.* Empirical studies have shown that overall, the use of CONSORT is associated with improvements in reporting quality over time (Moher et al. 2001; Plint et al 2006).

[2] http://www.campbellcollaboration.org/ .

[3] I am grateful to an anonymous peer reviewer for this observation.

CONSORT has also been adopted by the American Psychological Association, which represents a health discipline more familiar to many criminologists (Petrosino et al. 2006).

Although general efforts have been made to improve the reporting of criminological trials (e.g., Farrington et al. 2006), and some have specifically called for the development of a checklist and proposed basic frameworks (e.g., Farrington 2003a; Petrosino et al. 2006; see also Boruch 1997: Ch. 10, for social sciences generally), no consensus on standards similar to that seen in the health sciences has been reached. Recently, several reports of RCTs in the *Journal of Experimental Criminology* have explicitly and voluntarily adhered to CONSORT standards (e.g., Sherman et al. 2005; Watt et al. 2008; Barnes et al. 2010). Perry and Johnson (2008) provide empirical evidence that criminological trial reports only partially adhere to CONSORT. In a review of 17 RCTs on mental health services for juvenile offenders, they found considerable variability in the extent to which certain details were reported. Acknowledging that the narrow focus of their review may have overstated CONSORT compliance in criminology because the interventions they examined were rooted in the health and psychology disciplines, Perry and her colleagues repeated their investigation with a broader range of criminological trials and concluded that overall, descriptive validity is generally low (Perry et al. 2010).

The present study

The aim of the present study is to build upon previous discussion and empirical investigation of reporting quality in criminology, incorporating the relationship of descriptive validity to internal and external validity. The two studies described above (Perry and Johnson 2008; Perry et al. 2010) represent the only comprehensive attempts to apply the CONSORT checklist in its entirety to samples of criminological trials. I extend their analysis by focusing on the extent to which RCTs in criminology contain sufficient information to allow research consumers to judge the internal and external validity of the study. I assess this by developing indicators of the extent to which studies report information relevant to internal and external validity that could be used to create a rating system for policymakers.

It is important to emphasize at the outset that these reporting indicators are not intended to tell research consumers whether or not an experiment *is* internally or externally valid. Thus, I label the indicators 'R-IV' and 'R-EV' to remind the reader that they relate to information *reported* about the two types of validity. For example, an experiment with a high R-IV score may still have produced biased results due to differential attrition of participants, but the score shows that the report authors have provided enough information about the issues that affect internal validity to allow the reader to assess the extent to which the results remain meaningful. Conversely, a trial with a low score could have been perfectly implemented in practice, but the report provides so little information that to place substantial weight on its conclusions would be based on mere assumption. A study that rates highly on R-EV (external validity) provides sufficient descriptions of the setting, participant characteristics, and intervention details that policymakers can decide whether its outcomes could extend to the populations they serve.

Construction of internal and external validity indicators

In the absence of a reporting-standards checklist developed specifically for the field of criminology, I followed the methodology of Perry and Johnson (2008) and used the CONSORT checklist items, broken out into 45 individual elements. The elements pertinent to internal and external validity were selected to construct the indicators (see Fig. 1). Ultimately, the selection of elements was subjective, since there is also no agreed-upon measure of scientific validity or checklist for methodological quality (Shadish et al. 2002: 100; Farrington 2003a: 61-2). However, each element was chosen according to my assessment of whether it provided information relevant to the respective definitions of internal and external validity. Note that some of the selected elements are technically more relevant to Cook and Campbell's (1979) other validity criteria—statistical conclusion validity and construct validity. These concepts are closely related to internal and external validity respectively (Shadish et al. 2002). For example, it is necessary to know details such as the number of participants in the trial to establish the existence of an effect (statistical conclusion validity) before drawing causal inferences (internal validity). A clear description of the intervention is needed to establish that it is a valid representation of the theoretical concept being measured (construct validity) as well as to extrapolate to variations of that concept (external validity).[4]

CONSORT elements relevant to internal validity related to the generation and implementation of the random assignment sequence; the flow of participants through the trial; the length of the follow-up period; the number of participants analyzed; whether analysis was based on intention-to-treat (according to randomized treatment) or per protocol (according to treatment actually received); and whether the authors believed the results to be affected by bias. Threats to or overrides of the random assignment sequence, differential attrition, or outcome analysis based only on those who successfully completed the intervention, all affect the extent to which causation may be inferred.

CONSORT elements relevant to external validity related to the eligibility criteria for participants; setting of the trial; the dates of the recruitment period; baseline characteristics of the participants; and the authors' interpretation of the generalizability of their findings and the extent to which the results fit within the existing evidence-base. Details about these elements set the trial within geographic, cultural, and historical contexts, and allow policymakers to determine the extent to which the studied population and intervention aligns with their planned course of action in their own communities.

Sample selection

The sample of studies to which the reporting indicators are applied was drawn from a total of 28 journals (see Fig. 2), which were hand-searched by the author. Initially, I searched the top 20 criminology and penology journals (according to the 2007 Impact Factor ranking).[5] Since some of these journals were unlikely to publish RCTs

[4] This observation was also pointed out by an anonymous peer reviewer.

[5] ISI Web of Knowledge, 2007 Journal Citation Reports, Social Science Edition. Access provided through University of Pennsylvania Libraries, March 2009.

Individual Element #	CONSORT Item #	Descriptor
INTERNAL VALIDITY (R-IV)		
16	8	Method for generating the random assignment (RA) sequence
17	8	Details of RA restriction (e.g., blocking)
18	9	Method of implementation and concealment of RA sequence
19	10	Who generated the RA sequence
20	10	Who enrolled participants
21	10	Who assigned participants to their groups
12	13	Number of participants in the trial
26	13	Flow of participants through each stage of the trial
27	13	Details of number of participants randomly assigned
28	13	Details of number of participants receiving intended treatment
29	13	Details of number of participants completing study protocol
30	13	Details of number of participants analyzed for primary outcome
31	13	Description of any deviations from planned study protocol
33	14	Dates or timing of follow-up period
36	16	Was analysis based on intention to treat or per protocol
37	16	Number of participants (denominator) in each group used for analysis
43	20	Description of potential bias or confounding in the results
EXTERNAL VALIDITY (R-EV)		
3	3	Eligibility criteria for participants
4	3	Setting/location where data collected
5	4	Details of intervention for each group
6	4	How intervention was administered for each group
32	14	Date of recruitment period
34	15	Baseline demographic characteristics of participants in both groups
35	15	Relevant 'clinical' information about participants at baseline (*interpreted as criminal history information, etc. in this context*)
44	21	Discussion of generalizability (external validity) of findings
45	22	Interpretation of results in context of current evidence

Fig. 1 Reporting indicators of internal and external validity

(e.g., *Theoretical Criminology*), I boosted the sample size with several other criminology (e.g., *Journal of Experimental Criminology*[6]) and general evaluation journals (e.g., *Evaluation Review*). These additional journals were selected based on my familiarity with them and expectation that they might include experiments involving criminal justice settings and/or outcomes. I searched journal issues published between January 2002 and December 2008. This time period was selected in part to maintain a manageable number of trials for analysis by a sole author, but also to reflect a time period in which debate over reporting quality, and potentially also authors' and journal editors' familiarity with the issue, were increasing. The year 2002 was selected as the start year to incorporate a time lag between the 2000 inception of the Campbell Collaboration and the 2001 publication of the most recent CONSORT standards in medicine.

All primary reports of field experiments involving random allocation of human participants to treatment and control groups were examined for inclusion. Although the exclusion of studies with non-human units of analysis (i.e., place-based randomized trials such as 'hot spots' experiments) excludes a

[6] The *Journal of Experimental Criminology* is an obvious starting point in a search for RCTs. However, because the journal was founded more recently (2005), it was too new at the time of this research to be included in the Journal Citation Reports.

Journal Name	2007 Rank	# RCTs Included
Criminology	1	0
Crime & Delinquency	2	3
Criminal Justice & Behavior	3	4
Sexual Abuse: A Journal of Research & Treatment	4	1
Journal of Criminal Law & Criminology	5	0
British Journal of Criminology	6	1
Journal of Research in Crime & Delinquency	7	2
Journal of Quantitative Criminology	8	1
Punishment & Society	9	0
Journal of Interpersonal Violence	10	3
Aggression & Violent Behavior	11	0
Theoretical Criminology	12	0
Psychology, Crime, & Law	13	1
Justice Quarterly	14	4
International Journal of Offender Therapy & Comparative Criminology	15	1
Journal of Forensic Psychiatry & Psychology	16	0
Homicide Studies	17	0
Journal of Criminal Justice	18	1
Canadian Journal of Criminology	19	0
Legal & Criminological Psychology	20	0
Australian & New Zealand Journal of Criminology	N/A	0
Criminology & Public Policy	N/A	5
Journal of Experimental Criminology	N/A	8
Journal of Offender Rehabilitation	N/A	3
Annals of the American Academy of Social & Political Science	N/A	0
American Journal of Sociology	N/A	0
American Sociological Review	N/A	0
Evaluation Review	N/A	0

Fig. 2 Eligible studies by journal

substantial pool of recent experiments in policing, I felt that their inclusion might create a downward bias in reporting quality given my use of CONSORT. The CONSORT checklist was designed for simple two-group comparisons of human subjects, and several of the items that comprise my indicators do not apply to place-based trials in the form in which they appear in CONSORT. Thus, these studies would fail to score highly on the two indicators not because of poor reporting, but because there is currently no checklist specifically designed for the types of experiments many criminologists conduct. Laboratory-based and vignette studies were also excluded to avoid similar bias. These studies operate under more controlled conditions than those conducted in the field, so authors often do not need to report on attrition or implementation issues. Finally, I excluded any articles that did not report 'true' experiments: systematic reviews; follow-up surveys or analyses of subsets of RCT participants in which random assignment was not maintained; reports on multiple experiments, unless each one was fully reported (e.g., Goldkamp and White 2006); and preliminary results of trials, unless the authors purported to describe the experiment in full (e.g., Gottfredson and Exum 2002; Marlowe et al. 2003), and all the participants randomly assigned so far were included in the analysis. Based on these criteria, 38 RCTs were identified for analysis through title and abstract screening, and more thorough reading where necessary.

Coding of studies

The coding protocol developed for this study is reproduced in the Appendix. It was originally designed to gather full information about the reporting of each CONSORT element. I also recorded the publication year; the type of intervention studied (e.g., corrections); the institutional affiliation of the lead author at the time of publication; the field of the lead author's highest degree[7]; and whether or not the authors mentioned CONSORT. Each element was coded 2 if the authors described it in the report, and 0 if they did not. The code 1 was used where the report was partial or unclear; for example, where CONSORT required a description of the interventions for both the treatment and control groups, and the authors only described the treatment group. As this was a small-scale study conducted by this author alone, it is important to stress the caveat that these studies have not been double-coded for reliability.

Analysis plan

The R-IV and R-EV indicators for each study were constructed by taking the mean score (0-2) over each of the relevant CONSORT elements listed in Fig. 2, rounded to 1 decimal place. Thus, each study is assigned an R-IV rating between 0 and 2, and an R-EV rating between 0 and 2, to indicate the extent to which factors affecting internal and external validity were described. Higher scores indicate more comprehensive reports.

In the following section, I present a descriptive analysis of these reporting indicators. I first examine the mean R-IV and R-EV ratings across the full sample. Mean scores are then broken down by subgroups: publication year, to examine whether reporting standards improved over time; intervention type, to examine whether trials of certain types of programs in certain settings lend themselves to better reporting practices[8]; by the lead author's current institution and discipline (if in academia) and field of training, both of which could influence the extent to which authors consider the reporting of certain details in their work. Lum and Yang (2005) examined these last two factors in relation to authors' preferences for choosing experimental or non-experimental methods when designing research studies. They hypothesized that "disciplinary norms" and the field of training may determine academics' inclination toward or confidence in producing RCTs. By extension, I investigate the possibility that these norms affect attention to detail in writing up trial reports.

Finally I explore the relevance of the reporting indicators to policymakers and other research consumers. Each individual study's R-IV and R-EV scores convey information to the reader about the sufficiency of its reporting on matters relevant to internal and external validity. However, decision-makers committed to evidence-based policy usually need to take a range of studies into account when deciding whether

[7] When information about the lead author was not presented within the publication, I obtained it through web searches of publicly-available details (e.g., faculty Web pages, biographies, and CVs).

[8] Perry, Weisburd, and Hewitt (2010) suggest that more rigorous practices may have developed in certain criminal justice domains, perhaps due to funding availability or the backgrounds of researchers who work in those areas.

to implement an intervention or strategy. Graphical conceptualizations, such as matrices, are extremely useful for condensing large amounts of information into domains and patterns that can be easily digested by busy readers. A recent example in the field of criminology is the Evidence-Based Policing Matrix (Lum et al. in press), which visually classifies studies of policing strategies across several crime prevention dimensions. Following this example, I develop a 3×3 Descriptive Validity Matrix, which can be used to plot studies along dimensions of High, Medium, and Low descriptive validity on items relevant to the assessment of internal and external validity.

Results

Tables 1 and 2 show the R-IV and R-EV scores assigned in this study. Table 1 shows the mean scores across the whole sample, broken out by each individual CONSORT element that formed the reporting indicators. Table 2 shows the R-IV and R-EV score for each RCT coded. In the following discussion, I consider scores above 1.3 to indicate 'High' descriptive validity, scores between 0.7 and 1.3 'Medium,' and scores below 0.7 'Low.'

Overall, the mean R-IV and R-EV scores across the sample are fairly promising. These studies show medium descriptive validity in reporting on elements relevant to internal validity (mean R-IV=1.0), and high descriptive validity on items relevant to external validity (mean R-EV=1.5). However, the individual elements indicate a great deal of variation within these indicators. Reporting scores across the individual elements ranged from a low of 0.1 for R-IV item 18, 'method of implementation and concealment of random assignment sequence,' which was fully described in just one study (Zhang and Zhang 2005); to a perfect 2.0 for R-EV item 4, 'setting/location where data collected,' which was discussed in all the reports. Within almost all the individual elements, reporting scores ranged from 0 to 2.

Descriptive validity was high on internal validity items relating to the number of participants, dates and timing of follow up, and numbers randomly assigned and analyzed for the primary outcome. Many authors also attempted to assess potential biases in their conclusions. Less consistently reported were details about the number of participants actually receiving the treatment and completing the study protocol, whether outcomes were based on intention-to-treat or per protocol, who enrolled participants and assigned them to groups, and any deviations from the planned treatment. Low-scoring items related to the construction and implementation of the random assignment sequence and precise information on the flow of participants through each stage of the trial.

Within the external validity indicator, almost all the elements scored highly. Authors provided a good amount of detail about the setting and location of the study, eligibility criteria, dates of recruitment, description of the interventions, and how their findings related to current knowledge. However, fewer authors specifically addressed the issue of generalizability (external validity) in their conclusions.

The mean R-IV and R-EV scores were also broken down by subgroups, reflecting variation in the sample in publication date, topic area, and author affiliation and experience. Figure 3 shows the mean scores by publication year. The line represents the number of studies published in that year, or the denominator on which the means are based. The number of eligible studies published per year ranged from three in

Table 1 Mean R-IV and R-EV scores by element

Element #	Obs.	Mean	Std. dev.	Min	Max
IV-12	38	1.9	0.5	0	2
IV-33	38	1.8	0.6	0	2
IV-27	38	1.7	0.7	0	2
IV-37	38	1.6	0.8	0	2
IV-30	38	1.5	0.9	0	2
IV-43	38	1.5	0.9	0	2
IV-28	38	1.0	1.0	0	2
IV-36	38	1.0	1.0	0	2
IV-21	38	0.9	1.0	0	2
IV-20	38	0.8	1.0	0	2
IV-31	38	0.8	1.0	0	2
IV-29	38	0.8	1.0	0	2
IV-16	38	0.6	0.9	0	2
IV-17	38	0.6	0.9	0	2
IV-26	38	0.3	0.7	0	2
IV-19	38	0.3	0.7	0	2
IV-18	38	0.1	0.4	0	2
Mean R-IV	38	1.0	0.3	0.2	1.5
EV-4	38	2.0	0.0	2	2
EV-5	38	1.7	0.5	1	2
EV-3	38	1.6	0.8	0	2
EV-34	38	1.6	0.6	0	2
EV-35	38	1.5	0.7	0	2
EV-45	38	1.5	0.9	0	2
EV-32	38	1.3	0.9	0	2
EV-6	38	1.3	0.7	0	2
EV-44	38	0.8	1.0	0	2
Mean R-EV	38	1.5	0.4	0.6	2

2003 to eight each in 2002 and 2008. R-IV and R-EV were both at their highest in 2003, although the means are based on just three studies (Armstrong 2003; Gottfredson et al. 2003; Marlowe et al. 2003) that all scored highly on at least one indicator. There appears to be little pattern or variation in descriptive validity over the time period, although scores improved slightly during the later years. R-IV scores were medium in each year, and R-EV was generally high, although it dipped slightly in 2004, which was the lowest-scoring year overall.

Figure 4 shows the results by topic area. Again, there is little variation in scores, with medium R-IV in all categories and high R-EV in most. Reports of RCTs in more 'traditional' criminal justice domains, such as corrections and courts, scored slightly higher than those in psychological therapies and treatments, which is

Table 2 R-IV and R-EV analysis by study

	Study	R-IV	R-EV
★	Zhang and Zhang 2005	1.5	2.0
★	Weisburd, Einat, and Kowalski 2008	1.5	1.8
★	McGarrell and Hipple 2007	1.5	1.7
★	Haapanen and Britton 2002	1.5	1.6
★	Watt et al. 2008	1.5	1.6
★	Lane et al. 2005	1.4	2.0
★	Gottfredson et al. 2003	1.4	1.8
★	Labriola et al. 2008	1.4	1.8
★	Gottfredson et al. 2006	1.4	1.6
▲	Gottfredson and Exum 2002	1.5	1.2
■	Kinlock et al. 2008	1.1	2.0
■	Marlowe et al. 2003	1.1	2.0
■	Gondolf 2007	1.1	1.8
■	MacKenzie et al. 2007	1.1	1.8
■	Armstrong 2003	1.1	1.7
■	Marques et al. 2005	1.1	1.7
■	Van Voorhis et al. 2004	1.0	1.8
■	Little et al. 2004	0.9	1.6
■	White et al. 2006	0.9	1.6
■	Biggam and Power 2002	0.9	1.4
■	Kilmer 2008	0.8	1.8
■	Giblin 2002	0.8	1.6
■	Sacks et al. 2008	0.8	1.6
■	Banks and Gottfredson 2004	0.7	1.4
■	Taxman and Thanner 2006	0.7	1.4
●	Feske 2008	1.2	1.3
●	Stickle et al. 2008	1.2	1.3
●	Goldkamp and White 2006	1.2	1.2
●	Feder and Dugan 2002	1.1	1.3
●	Bottcher and Ezell 2005	0.9	1.3
●	Blais and Bacher 2007	0.8	0.9
●	Vannoy and Hoyt 2004	0.8	0.7
●	Bijleveld 2007	0.7	1.1
●	Sullivan et al. 2002	0.7	1.1
●	Phillips 2004	0.7	0.7
⬣	Armstrong 2002	0.6	1.4
◆	LeSure-Lester 2002	0.4	1.1
▼	Moynahan and Strømgren 2005	0.2	0.6

somewhat surprising given the psychology field's closer alliance with health science and its adoption of CONSORT standards during the time period covered in this study.

In Fig. 5, average R-IV and R-EV scores are broken out by the institutional affiliation of the lead author at the time of publication. The 'Government' category, for example, might include authors affiliated with a state or local Department of Corrections in the U.S. The 'Research Organization' category includes non-profit

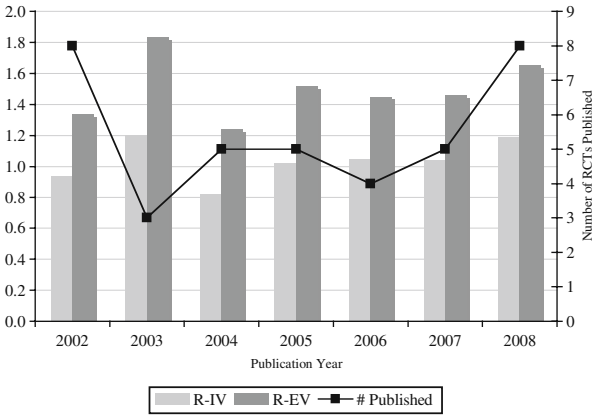
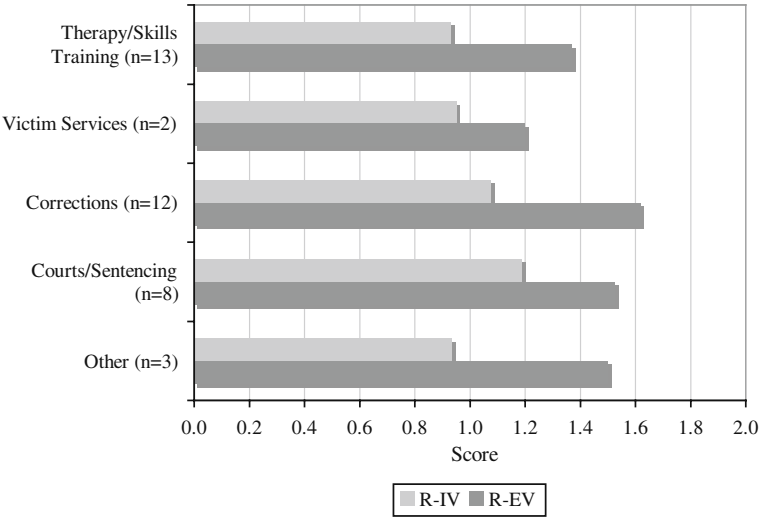


Fig. 3 Reporting indicators by publication year

and for-profit agencies like the RAND Corporation. Authors working in university departments or research centers are classified under 'Academic,' which is broken down into broad disciplinary areas to account for differing norms and practices. Unsurprisingly, authors from medical schools or departments were most successful in conveying details relevant to internal validity and also scored very highly on R-EV. Academic sociologists provided the most detail about external validity, and authors working in research organizations also scored well on that indicator. Criminologists were relatively successful, with medium R-IV and high R-EV scores. Again, authors from psychology backgrounds performed less well, with an R-IV score on the borderline of low and medium. It is important to note that these results may well be biased for affiliations other than criminology and research organizations, due to small numbers of authors falling into the other categories.



Note: "Other" interventions were medical treatments (acupuncture; methadone maintenance) and letters sent to insurance claimants to deter claim padding.

Fig. 4 Reporting indicators by topic area

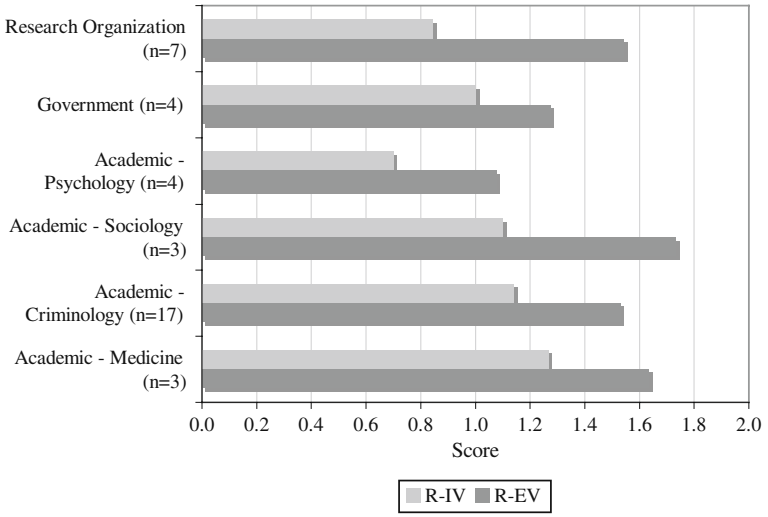
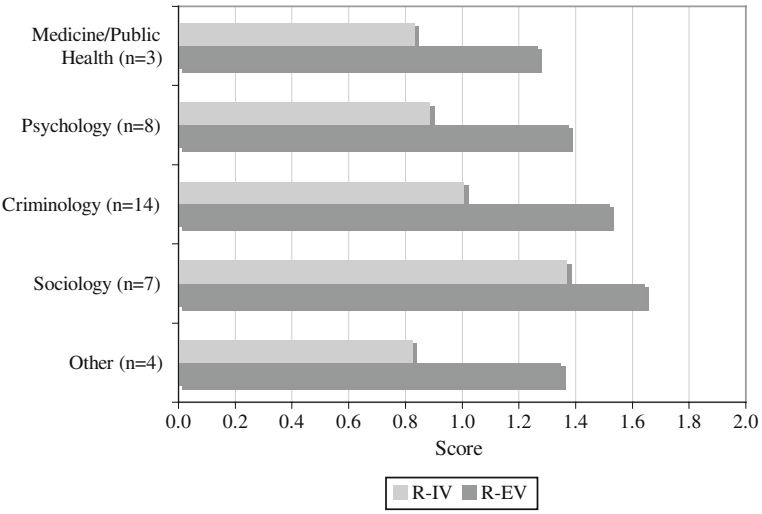


Fig. 5 Reporting indicators by lead author's institutional affiliation

Following Lum and Yang (2005), I also investigated whether the lead author's field of training, as distinct from their current affiliation, affected their reporting practices (Fig. 6). Thirty-seven of the 38 lead authors (97.4%) either held or were pursuing a Ph.D. at the time of publication. I was unable to obtain information about the field in which two authors had received their doctorates. Again, a similar pattern emerges: R-IV is medium and R-EV is high across most fields. However, it is interesting to note that authors who received their Ph.D. in sociology or criminology/criminal justice had the highest scores on both R-IV and R-EV, whereas scores for those trained in medicine or public health were lowest. Again, small numbers of authors trained in the medical field may have skewed these results. However, the relatively high scores for sociologists (some of whom would have specialized in criminology) and criminologists are very promising for the field, indicating some tradition of good reporting practices in these disciplines even in the absence of a dedicated checklist.

One reason for the promising results observed among authors trained or working in criminology and sociology may be that many of the leading proponents of evidence-based policy and experimental practice in these fields have themselves been directly involved in conducting randomized controlled trials and would likely incorporate issues of quality and descriptive validity into their reports. In order to examine this hypothesis further, I coded each study according to whether or not a current Fellow of the Academy of Experimental Criminology (AEC) had been involved, either as lead author or a co-author. AEC Fellows are specifically recognized for their experience and success in conducting randomized controlled trials in criminology. In this sample, ten AEC Fellows (including two past or current AEC presidents) are represented: eight as lead authors (several of whom also co-authored other studies in the sample), and two as co-authors.[9] There is also some

[9] Lynette Feder, John Goldkamp, Denise Gottfredson, Doris MacKenzie, Edmund McGarrell, Jonathan Shepherd, Faye Taxman, Susan Turner, Patricia van Voorhis, and David Weisburd. A list of AEC Fellows is published at http://www.crim.upenn.edu/aec/fellows.htm.



Note: data are missing for 2 studies. "Other" fields were Education (2), Mathematics, and Public Policy.

Fig. 6 Reporting indicators by lead author's field of training

overlap between AEC Fellows and members of the Campbell Collaboration Crime and Justice Group steering committee, although I do not break down the results by Campbell membership as only one member was the lead author of a trial in this sample.[10]

Figure 7 shows a distinct improvement in scores, particularly for reporting of internal validity, when an AEC Fellow is involved in authoring a study, particularly when he or she is a lead author (R-IV, lead author=1.2; co-author=1.1; no involvement=0.9). External validity is high in all groups, but slightly higher for studies involving an AEC fellow.

Having examined the characteristics of the sample as a whole, I now turn to the visual presentation of each study's individual score (from Table 2, above) on the Descriptive Validity Matrix (Fig. 8). The Matrix is divided into 9 squares, representing each combination of Low, Medium, and High reporting quality for R-IV and R-EV. Each study's R-EV score is plotted against its R-IV score on the Matrix to summarize reporting quality across both domains. Studies that fall closest to the top right-hand corner of the Matrix scored highest on both R-IV and R-EV. Each study is labeled with a differently shaped symbol depending on which dimension of the Matrix it falls into. This symbol can be used to refer back to a list of the individual study scores and references, such as the one in Table 2. Thus, the user of the Matrix might decide only to consider studies that rated high for descriptions of both internal and external validity, which are labeled with stars on the Matrix. He or she could then refer just to the starred section of the table for further information.

[10] David Weisburd; Jonathan Shepherd and Peter van der Laan were co-authors of two additional studies. A list of Campbell Collaboration Crime and Justice Group steering committee members is published at http://www.campbellcollaboration.org/crime_and_justice_our_group/Who_s_involved_CJ.php.

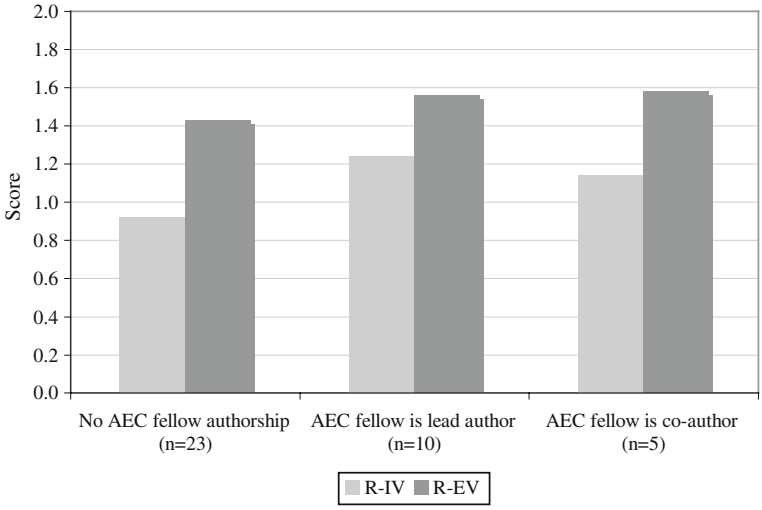


Fig. 7 Reporting indicators by AEC Fellow authorship

Figure 8 shows the application of the Matrix to the present sample. While the results above have indicated that overall, reporting quality has been medium for items relevant to internal validity and high for items relevant to external validity, the Matrix clearly demonstrates that the majority of studies in the sample had reasonably high descriptive validity. In all, eight of 38 studies (21.1%) scored 'high' on both indicators, and 26 of 38 (68.4%) scored highly on at least one. Only nine studies (23.7%) scored high on R-IV, compared to 25 (65.8%) scoring high on R-EV. However, the overall distribution of scores is in the direction of the top-right corner, which is a promising assessment for reporting quality in the field.

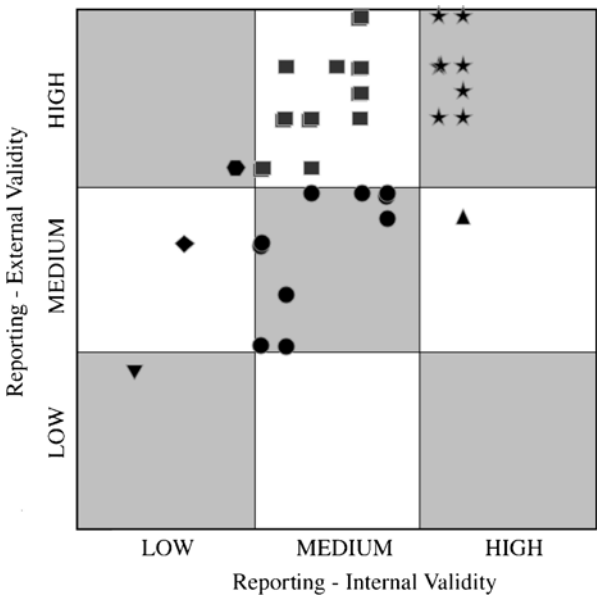


Fig. 8 Descriptive validity matrix

Discussion

Despite concerns about descriptive validity in the criminological literature, and findings from empirical research (Perry and Johnson 2008; Perry et al. 2010) indicating that criminal justice research has some way to go to catch up to standards in the healthcare field, the results of this study are fairly promising. This sample of RCTs published between 2002 and 2008 in leading criminology journals provides at least partial details of information crucial to assessments of internal and external validity, and thus to the policy relevance of crime and justice evidence. The studies were particularly strong in reporting on items relevant to external validity or generalizability, which is of paramount importance in translating evidence into practice across different populations and settings. Furthermore, despite the need to borrow a reporting validity checklist from the medical field, studies conducted by criminologists and sociologists or focused on more traditional criminal justice strategies and settings performed as well, and sometimes better, than crime-related studies conducted within health science disciplines. The recent focus on experimental methods and evidence-based practice in criminology, and the founding of organizations like the Campbell Collaboration Crime and Justice Group and the Academy of Experimental Criminology, may have led to an improvement in reporting methods despite the lack of an agreed-upon standard. Studies in the present sample that were authored or co-authored by AEC Fellows had better reporting quality than those that were not, particularly on items related to internal validity.

However, the results of this study also indicate that much more needs to be done to improve reporting quality even further. A consistent finding across all the results presented above is that the R-IV score is always distinctly lower than the R-EV score. Only one study scored higher on R-IV than R-EV (Gottfredson and Exum 2002). There is substantial variability in the extent to which the individual elements of the R-IV indicator were reported. Arguably, many of the details that comprise the R-EV indicator are more obvious or easier to capture than those comprising R-IV. In a complex field RCT, researchers will certainly know the details of the intervention and the eligibility criteria for participants, but it may be much more difficult to track information about the flow of participants through the trial, especially if they are reliant on staff who work in the field to provide information over and above their normal duties. However, participant flow is vital in showing how representative the final sample of participants was of the full population, and how many were lost at each stage of the experiment. Differential attrition of participants and treatment crossover are major threats to internal validity, especially when those who do not drop out or who end up receiving the intervention are those most likely to respond positively. Fewer than half of the studies in the sample reported the numbers of participants actually receiving the intended treatments and completing the study protocol separately from the numbers randomly assigned or analyzed. Only half of the studies indicated whether the analysis was based on intention-to-treat or per protocol. The studies also provided very little information about the random assignment sequence; for example, only one study fully reported the methods of implementation and concealment of the sequence (Zhang and Zhang 2005), and just three more provided partial information (Haapanen and Britton 2002; Labriola et al. 2008; Watt et al. 2008). Allocation concealment is crucial to internal validity

because it prevents selection bias by ensuring that the random assignment sequence is not known in advance. Prior knowledge of the sequence could, for example, result in participants thought to be 'deserving' of treatment being deliberately selected for the treatment group, which biases effect estimates (see Altman et al. 2001: 673).

Overall, the issues with reporting information about details relevant to internal validity in this sample are a cause for concern. Internal validity is considered to be the most important dimension of scientific validity (Farrington 2003a; Shadish et al. 2002), so it follows that without a good R-IV score, a high R-EV rating would not be as meaningful. Arguably, a study's applicability to other populations and settings is irrelevant if the causal relationships it demonstrates are unreliable.

The key contribution of this study is the development of the Descriptive Validity Matrix, which visually organizes studies according to their R-IV and R-EV scores. The Matrix is a simple, intuitive way to convey information to decision-makers about whether a set of evaluations provide sufficient information to judge their internal and external validity. The most obvious application of the Matrix would be as an organizing scheme for a set of studies examining the same intervention or treatment: for example, a matrix could be produced that classifies all the rigorous evaluations conducted on drug courts according to R-IV and R-EV. A decision-maker who is considering implementing drug courts in his or her jurisdiction could use the Matrix to identify a subset of evaluations meeting a minimum standard of reporting quality, which would save the time of reading through reports that do not contain sufficient information. Alternatively, the Matrix could be taken in its entirety as an indicator of reporting quality across the evidence-base, providing the user with a basis for assessing and articulating confidence in his or her decisions based on the available research. As well as being a decision-making tool, the Matrix could be used by scholars of scientific validity to identify areas for improvement and develop checklists and standards in those areas.

Of course, producing a Matrix for each type of intervention would be quite a time-consuming task. However, it could in theory be successfully combined with the systematic reviews produced on behalf of the Campbell Collaboration. One of the essential steps of systematic review is the development of a coding protocol to extract information from each study about the intervention, population, and outcomes. The R-IV and R-EV indicators used in this study consist of 26 items that could be easily judged while reading the study, and recorded on the protocol. The indicators themselves are based on a simple mean and can be calculated in seconds with any statistical software or spreadsheet. Systematic review authors could generate the Matrix and include it in their reports along with the list of references. It is even concise enough to be included in shorter 'user abstracts.' In this way, the discipline of systematic review contributes to the development of evidence-based policy by providing summaries of both the overall effects of an intervention, and the confidence that can be placed in those effects based on the extent to which the review authors could glean information from the primary research.

This study has several limitations. It was not always possible to distinguish CONSORT items that were not reported from those that did not apply. Although all the items were relevant to criminological trials in general, it is not necessarily the case that all the issues would apply to all trials. For example, a report might fail to discuss the results in the context of current evidence, but the study may represent the first attempt to assess a particular strategy. In addition, the coding of CONSORT

items was conducted by one person, and as such is based on personal judgment. Other readers of the same study reports may disagree with my assessments. However, I have been careful to apply an objective understanding of the concepts of internal and external validity based on prior literature.

The sample is a small subset of criminal justice experiments, so the studies reviewed here may not be representative of the overall quality of reporting in the field as a whole. The limited timeframe does not encompass some of the more productive eras in experimental research in criminology. Although evidence-based crime policy gained prominence relatively recently, Farrington and Welsh (2005) found 83 criminological RCTs published between 1982 and 2004, and a further 35 conducted between 1957 and 1981. More importantly, for reasons described above, the sampling criteria excluded place-based experiments, which eliminates much of the recent research on policing, a key domain of criminology that has provided a fruitful output of experimental research. Several high-quality studies (e.g., Weisburd et al. 2006; Braga and Bond 2008; Weisburd, Morris, and Ready 2008) representative of the field were excluded as a result.

Furthermore, only RCTs published in academic journals are included. Journal articles may be constrained by space and themes, and focus more on results and contributions to scholarship and criminological theory than the finer details of the project. This may explain why fewer authors specifically reported their own assessments of external validity in this study. Policymakers may be more likely to read research from their own governmental organizations, private research organizations, and technical reports submitted by academics (for example, the grant report on which a journal article may be based), which may contain more information about the full details of the study. Thus, this study may actually underestimate the quality of information available to policymakers. It is conceivable that when good research comes across their desks, it is in the form of more detailed technical reports.

Future research in this area should focus on refining the indicator system developed in this study to better capture information vital to the assessment of internal and external validity and increase its relevance to criminological trials. More work is required to unpack the definitions of internal and external validity themselves before they can be fully incorporated into reporting standards. The present indicator system also assumes that all the elements of internal and external validity are of equal importance, which may be unjustified. A refinement to this system, with the guidance of further research on the nature of scientific validity, might incorporate a weighted average to rank certain elements of validity as more or less important. In addition, this study does not examine the other important types of validity—statistical conclusion and construct validity—both of which are also important to policy relevance. For example, low statistical power is a major threat to statistical conclusion validity (Farrington 2003a: 52) and a chronic problem in criminological research (Brown 1989; Weisburd et al. 1993), yet fewer than 25% of the studies reviewed here offered information on how the sample size was determined. As discussed above, there is also an urgent need for a modified CONSORT-type reporting checklist designed specifically for the field of criminology, which takes into account the different research designs and units of analysis that are not found in the health sciences, the most obvious of which are the place-based experiments.

It would also be instructive to conduct a similar study of internal and external validity reporting in healthcare trials and compare it to these findings (Perry et al.

2010). This would help us to learn whether criminology does need to catch up with the medical field, especially since this study suggests that criminological trial reports authored by scholars trained or working in medicine were not always better reported than trials written by those from social science backgrounds. Furthermore, given the extent to which research and practice are connected in the health sciences (Shepherd 2003), it would be interesting to contrast health and criminology trials on the Matrix to compare the amount of policy-relevant information they provide.

Conclusions

This paper makes the case for the importance of descriptive validity as a foundation for drawing conclusions about scientific validity in criminological research. I constructed indicators designed to help research consumers assess whether a study provided sufficient information to assess the trial's internal and external validity. I applied the indicators to 38 randomized controlled trials of criminological interventions. Reporting quality results were mixed, with factors relevant to external validity well reported, but important information about technical aspects of study design that greatly impact conclusions about internal validity routinely missed. Although the reporting standard applied was borrowed from the healthcare field, the elements that formed the reporting indicators were equally applicable to the effective reporting of criminal justice trials, and those items that were missed were not omitted because they were irrelevant. The indicators developed were used to map studies onto a Descriptive Validity Matrix, which could be provided to policymakers to help them assess the quality of information available in the evidence-base for a particular intervention or strategy.

Although this study has some limitations, it represents an important first step in assessing how descriptive validity relates to internal and external validity, and the value of criminological research to policy and practice. The indicators developed are based on a respected, well-documented framework and have been applied to a group of studies that is representative of much of the experimental research in criminology. As such, this is a useful starting point and framework for continued assessment of descriptive validity. The General Accounting Office report has indicated that the field of criminology still has some distance to go in improving the quality of the research it offers to policy decision-makers. While descriptive validity indicators do not address the fundamental difficulties of conducting field experiments in criminal justice settings, attention to good reporting of the problems that inevitably arise could go a long way toward helping decision-makers to make sense of research quality. As criminologists continue to hold up the randomized controlled trial as the authoritative evaluation design, and expand efforts to disseminate the results of experiments and systematic reviews to policymakers, we must recognize the "moral imperative" (Weisburd 2003) not only to produce the best research, but to clearly report it to enhance the objectives of evidence-based crime policy.

Acknowledgments I would like to thank Richard Berk, Bob Boruch, John MacDonald, Larry Sherman, David Weisburd, and the anonymous reviewers for their helpful comments and support in developing and improving this paper.

Appendix

Coding Protocol for Descriptive Validity in Criminological Trials

Study ID:

Article title:
Article year:
Journal:
Journal country:
General topic of study:

Lead author (LA):
LA department:
LA institution:
LA institution country:
Level and substantive area of
LA highest degree:
CONSORT mentioned: 0 (No) 1 (Yes)

For the following CONSORT -derived items, code '0' if article authors do not report on the item, '1' if it is partially reported (include a brief explanation of this assessment on the line), and '2' if it is fully reported.

Individual Element #	CONSORT Item #				
		Abstract/Title/Keywords			
1	1	"Random assignment" etc. in abstract/title/keywords	0	1	2
		Methods			
2	2	Background/rationale	0	1	2
3	3	Eligibility criteria for participants	0	1	2
4	3	Setting/location where data collected	0	1	2
5	4	Details of intervention for each group	0	1	2
6	4	How intervention was administered for each group	0	1	2
7	5	Objectives of the study	0	1	2
8	5	Explicit hypotheses	0	1	2
9	6	Primary outcome measures	0	1	2
10	6	Secondary outcome measures	0	1	2
11	6	Any methods used to enhance the quality of measurement	0	1	2
12	13	Number of participants in the trial	0	1	2
13	7	How the sample size was determined	0	1	2
14	7	Explanation for any interim analyses	0	1	2
15	7	Explanation for any stopping rules	0	1	2
16	8	Method for generating the random assignment sequence	0	1	2
17	8	Details of random assignment restriction (e.g., blocking)	0	1	2
18	9	Method of implementation and concealment of RA sequence	0	1	2
19	10	Who generated the random assignment sequence	0	1	2
20	10	Who enrolled participants	0	1	2
21	10	Who assigned participants to their groups	0	1	2
22	11	Description of any blinding	0	1	2
23	11	Evaluation of the success of blinding	0	1	2
24	12	Statistical methods used to compare groups for primary outcomes	0	1	2
25	12	Separate description of methods for additional analyses (e.g. subgroups)	0	1	2
		Results			
26	13	Flow of participants through each stage of the trial	0	1	2

27	13	Details of number of participants randomly assigned	0	1	2
28	13	Details of number of participants receiving intended treatment	0	1	2
29	13	Details of number of participants completing study protoc ol	0	1	2
30	13	Details of number of participants analyzed for primary outcome	0	1	2
31	13	Description of any deviations from planned study protocol	0	1	2
32	14	Dates of recruitment period	0	1	2
33	14	Dates or timing of follo w-up period	0	1	2
34	15	Baseline demographic characteristics of participants in both groups	0	1	2
35	15	Relevant 'clinical' information about participants at baseline	0	1	2
36	16	Was analysis based on intention to treat or per proto col	0	1	2
37	16	Number of participants (denominator) in each group used for analysis	0	1	2
38	17	Summary of outcomes for each group	0	1	2
39	17	Estimated effect size and precision (e.g., 95% CI) for outcomes (or enough information to calculate it)	0	1	2
40	18	Report on other analyses, e.g., subgroup or adjusted analyses	0	1	2
41	19	Report on any adverse effects	0	1	2
		Discussion			
42	20	Interpretation of results in light of study hypotheses	0	1	2
43	20	Description of potential bias or confounding in the results	0	1	2
44	21	Discussion of generalizability (external validity) of findings	0	1	2
45	22	Interpretation of results in context of current evidence	0	1	2

References

* denotes a report of an experiment included in the study

Altman, D. G., Schultz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134*(8), 663–694.

*Armstrong, T. A. (2002). The effect of environment on the behavior of youthful offenders: A randomized experiment. *Journal of Criminal Justice, 30*, 19–28.

*Armstrong, T. A. (2003). The effect of moral reconation therapy on the recidivism of youthful offenders: A randomized experiment.

*Banks, D., & Gottfredson, D. C. (2004). Participation in drug treatment court and time to rearrest. *Justice Quarterly, 21*(3), 637–658.

Barnes, G. C., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E., & Malvestuto, R. (2010). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology, 6*, 159–189.

Berk, R. A., & Rossi, P. H. (1999). *Thinking about program evaluation* (2nd ed.). Thousand Oaks, CA: Sage.

*Biggam, F. H., & Power, K. G. (2002). A controlled, problem-solving, group-based intervention with vulnerable incarcerated young offenders. *International Journal of Offender Therapy and Comparative Criminology, 46*(6), 678–698.

*Bijleveld, C. (2007). Fare dodging and the strong arm of the law: An experimental evaluation of two different penalty schemes for fare evasion. *Journal of Experimental Criminology, 3*, 183–199.

*Blais, E., & Bacher, J. L. (2007). Situational deterrence and claim padding: Results from a randomized field experiment. *Journal of Experimental Criminology, 3*, 337–352.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Boruch, R., Snyder, B., & DeMoya, D. (2000). The importance of randomized field trials. *Crime and Delinquency, 46*(2), 156–180.

*Bottcher, J., & Ezell, M. E. (2005). Examining the effectiveness of boot camps: A randomized experiment with a long-term follow up. *Journal of Research in Crime and Delinquency, 42*(3), 309–332.

Braga, A. A., & Bond, B. J. (2008). Policing crime and disorder hot spots: A randomized controlled trial. *Criminology, 46*(3), 577–607.

Brown, S. E. (1989). Statistical power and criminal justice research. *Journal of Criminal Justice, 17*, 115–122.

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

Farrington, D. P. (2003a). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science, 587* (May), 49-58.
Farrington, D. P. (2003b). A short history of randomized experiments in criminology: A meager feast. *Evaluation Review, 27*(3), 218–227.
Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2006). The Maryland Scientific Methods Scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (2nd ed., pp. 13–21). New York, NY: Routledge.
Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.
*Feder, L., & Dugan, L. (2002). A test of the efficacy of court-mandated counseling for domestic violence offenders: The Broward experiment. *Justice Quarterly, 19*(2), 343–375.
*Feske, U. (2008). Treating low-income and minority women with posttraumatic stress disorder: A pilot study comparing prolonged exposure and treatment as usual conducted by community therapists. *Journal of Interpersonal Violence, 23*(8), 1027–1040.
*Giblin, M. J. (2002). Using police officers to enhance the supervision of juvenile probationers: An evaluation of the Anchorage CAN program. *Crime and Delinquency, 48*(1), 116–137.
Goldkamp, J. S. (2008). Missing the target and missing the point: 'Successful' random assignment but misleading results. *Journal of Experimental Criminology, 4*, 83–115.
*Goldkamp, J. S., & White, M. D. (2006). Restoring accountability in pretrial release: the Philadelphia pretrial release supervision experiments. *Journal of Experimental Criminology, 2*, 143–181.
*Gondolf, E. W. (2007). Culturally-focused batterer counseling for African-American men. *Criminology and Public Policy, 6*(2), 341–366.
*Gottfredson, D. C., & Exum, M. L. (2002). The Baltimore City drug treatment court: One-year results from a randomized study. *Journal of Research in Crime and Delinquency, 39*(3), 337–356.
*Gottfredson, D. C., Najaka, S. S., & Kearley, B. (2003). Effectiveness of drug treatment courts: Evidence from a randomized trial. *Criminology and Public Policy, 2*(2), 171–196.
*Gottfredson, D. C., Najaka, S. S., Kearley, B. W., & Rocha, C. M. (2006). Long-term effects of participation in the Baltimore City drug treatment court: Results from an experimental study. *Journal of Experimental Criminology, 2*, 67–98.
*Haapanen, R., & Britton, L. (2002). Drug testing for youthful offenders on parole: An experimental evaluation. *Criminology and Public Policy, 1*(2), 217–244.
*Kilmer, B. (2008). Does parolee drug testing influence employment and education outcomes? Evidence from a randomized experiment with noncompliance. *Journal of Quantitative Criminology, 24*, 93–123.
*Kinlock, T. W., Gordon, M. S., Schwartz, R. P., & O'Grady, K. E. (2008). A study of methadone maintenance for male prisoners: 3-month postrelease outcomes. *Criminal Justice and Behavior, 35*(1), 34–47.
*Labriola, M., Rempel, M., & Davis, R. C. (2008). Do batterer programs reduce recidivism? Results from a randomized trial in the Bronx. *Justice Quarterly, 25*(2), 252–282.
*Lane, J., Turner, S., Fain, T., & Seghal, A. (2005). Evaluating an experimental intensive juvenile probation program: Supervision and official outcomes. *Crime and Delinquency, 51*(1), 26–52.
Lauritsen, J. L. (2006). Assessing problematic research: How can academic researchers help improve the quality of anticrime program evaluations? *Journal of Experimental Criminology, 2*, 363–373.
*LeSure-Lester, G. E. (2002). An application of cognitive-behavior principles in the reduction of aggression among abused African American adolescents. *Journal of Interpersonal Violence, 17*(4), 394–402.
Lipsey, M., Petrie, C., Weisburd, D., & Gottfredson, D. (2006). Improving evaluation of anti-crime programs: Summary of a National Research Council report. *Journal of Experimental Criminology, 2*, 271–307.
*Little, M., Kogan, J., Bullock, R., & van der Laan, P. (2004). ISSP: An experiment in multi-systemic responses to persistent young offenders known to children's services. *British Journal of Criminology, 44*, 225–240.
Lösel, F., & Köferl, P. (1989). Evaluation research on correctional treatment in West Germany: A meta-analysis. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system: Psychological perspectives* (pp. 334–355). New York, NY: Springer.
Lum, C., Koper, C. S., and Telep, C. W. (in press). The Evidence-Based Policing Matrix. *Journal of Experimental Criminology.*
Lum, C., & Yang, S.-M. (2005). Why do evaluation researchers in crime and justice choose non-experimental methods? *Journal of Experimental Criminology, 1*, 191–213.
MacKenzie, D. L. (2000). Evidence-based corrections: Identifying what works. *Crime and Delinquency, 46*(4), 457–471.

*MacKenzie, D. L., Bierie, D. L., & Mitchell, O. (2007). An experimental study of a therapeutic boot camp: Impact on impulses, attitudes and recidivism. *Journal of Experimental Criminology, 3*, 221–246.

*Marlowe, D. B., Festinger, D. S., Lee, P. A., Schepise, M. M., Hazzard, J. E. R., Merrill, J. C., et al. (2003). Are judicial status hearings a key component of drug court? During-treatment data from a randomized trial. *Criminal Justice and Behavior, 30*(2), 141–162.

*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* (1), 79–107.

*McGarrell, E. F., & Hipple, N. K. (2007). Family group conferencing and re-offending among first-time juvenile offenders: The Indianapolis experiment. *Justice Quarterly, 24*(2), 221–246.

Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *Journal of the American Medical Association, 285*(15), 1992–1995.

*Moynahan, L., & Strømgren, B. (2005). Preliminary results of Aggression Replacement Training for Norwegian youth with aggressive behavior and with a different diagnosis. *Psychology, Crime & Law, 11*(4), 411–419.

Perry, A. E. (2010). Descriptive validity and transparent reporting in randomised controlled trials. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 333–352). New York, NY: Springer.

Perry, A. E., & Johnson, M. (2008). Applying the Consolidated Standards of Reporting Trials (CONSORT) to studies of mental health provision for juvenile offenders: A research note. *Journal of Experimental Criminology, 4*, 165–185.

Perry, A. E., Weisburd, D., & Hewitt, C. (2010). Are criminologists describing randomized controlled trials in ways that allow us to assess them? Findings from a sample of crime and justice trials. *Journal of Experimental Criminology, 6*, 245–262.

Petrosino, A., Kiff, P., & Lavenberg, J. (2006). Randomized field experiments published in the *British Journal of Criminology*, 1960-1994. *Journal of Experimental Criminology, 2*, 99–111.

*Phillips, L. A. (2004). Evaluating the effects of a brief program for moral education in a county jail. *Journal of Offender Rehabilitation, 39*(2), 59–72.

Plint, A. C., Moher, D., Morrison, A., Schultz, K., Altman, D. G., Hill, C., et al. (2006). Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *The Medical Journal of Australia, 185*(5), 263–267.

*Sacks, J. Y., Sacks, S., McKendrick, K., Banks, S., Schoeneberger, M., Hamilton, Z., et al. (2008). Prison therapeutic community treatment for female offenders: Profiles and preliminary findings for mental health and other variables (crime, substance use and HIV risk). *Journal of Offender Rehabilitation, 46* (3), 233–261.

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

Shepherd, J. P. (2003). Explaining feast or famine in randomized field trials: Medical science and criminology compared. *Evaluation Review, 27*(3), 290–315.

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (Eds.). (2006). *Evidence-based crime prevention* (2nd ed.). New York, NY: Routledge.

Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising.* Washington, D.C.: United States Department of Justice, National Institute of Justice. Retrieved from http://www.ncjrs.gov/works.

Sherman, L. W., Strang, H., Angel, C., Woods, D., Barnes, G. C., Bennett, S., et al. (2005). Effects of face-to-face restorative justice on victims of crime in four randomized, controlled trials. *Journal of Experimental Criminology, 1*, 367–395.

*Stickle, W. P., Connell, N. M., Wilson, D. M., & Gottfredson, D. (2008). An experimental evaluation of teen courts. *Journal of Experimental Criminology, 4*, 137–163.

*Sullivan, C. M., Bybee, D. I., & Allen, N. E. (2002). Findings from a community-based program for battered women and their children. *Journal of Interpersonal Violence, 17*(9), 915–936.

*Taxman, F. S., & Thanner, M. (2006). Risk, need, and responsivity (RNR): It all depends. *Crime and Delinquency, 52*(1), 28–51.

*van Voorhis, P., Spruance, L. M., Ritchey, P. N., Listwan, S. J., & Seabrook, R. (2004). The Georgia Cognitive Skills experiment: A replication of Reasoning and Rehabilitation. *Criminal Justice and Behavior, 31*(3), 282–305.

*Vannoy, S. D., & Hoyt, W. T. (2004). Evaluation of an anger therapy intervention for incarcerated adult males. *Journal of Offender Rehabilitation, 39*(2), 39–57.

*Watt, K., Shepherd, J., & Newcombe, R. (2008). Drunk and dangerous: A randomised controlled trial of alcohol brief intervention for violent offenders. *Journal of Experimental Criminology, 4*, 1–19.
Weisburd, D. (2003). Ethical practice and evaluation of interventions in crime and justice: The moral imperative for randomized trials. *Evaluation Review, 27*(3), 336–354.
*Weisburd, D., Einat, T., & Kowalski, M. (2008). The miracle of the cells: An experimental study of interventions to increase payment of court-ordered financial obligations. *Criminology and Public Policy, 7*(1), 9–36.
Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Social and Political Science, 578* (November), 50-70.
Weisburd, D., Morris, N. A., & Ready, J. (2008). Risk-focused policing at places: An experimental evaluation. *Justice Quarterly, 25*(1), 163–200.
Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice, 17*, 337–379.
Weisburd, D., Wyckoff, L. A., Ready, J., Eck, J. E., Hinkle, J. C., & Gajewski, F. (2006). Does crime just move around the corner? A controlled study of spatial displacement and diffusion of crime control benefits. *Criminology, 44*(3), 549–592.
*White, M. D., Goldkamp, J. S., & Robinson, J. B. (2006). Acupuncture in drug treatment: Exploring its role and impact on participant behavior in the drug court setting. *Journal of Experimental Criminology, 2*, 45–65.
*Zhang, S. X., & Zhang, L. (2005). An experimental study of the Los Angeles County Repeat Offender Prevention Program: Its implementation and evaluation. *Criminology and Public Policy, 4*(2), 205–236.

Charlotte Gill is a post-doctoral fellow in the Center for Evidence-Based Crime Policy, George Mason University. She completed her Ph.D. at the Jerry Lee Center of Criminology, University of Pennsylvania, in 2010. Dr. Gill is the managing editor of the Campbell Crime and Justice Group. Her research interests include evidence-based crime prevention, quantitative methods, and research synthesis.