Methodological Quality Standards for Evaluation Research

The ANNALS of the American Academy of Political and Social Science 2003 587: 49
DOI: 10.1177/0002716202250789

Methodological Quality Standards for Evaluation Research

By
DAVID P. FARRINGTON

Evaluation studies vary in methodological quality. It is essential to develop methodological quality standards for evaluation research that can be understood and easily used by scholars, practitioners, policy makers, the mass media, and systematic reviewers. This article proposes that such standards should be based on statistical conclusion validity, internal validity, construct validity, external validity, and descriptive validity. Methodological quality scales are reviewed, and it is argued that efforts should be made to improve them. Pawson and Tilley's challenge to the Campbell evaluation tradition is also assessed. It is concluded that this challenge does not have any implications for methodological quality standards, because the Campbell tradition already emphasizes the need to study moderators and mediators in evaluation research.

Keywords: methodological quality; evaluation; validity; crime reduction; systematic reviews

The Campbell Collaboration Crime and Justice Group aims to prepare and maintain systematic reviews of impact evaluation studies on the effectiveness of criminological interventions and to make them accessible electronically to scholars, practitioners, policy makers, the mass media, and the general public (Farrington and Petrosino 2000, 2001). It is clear that evalua-

David P. Farrington is professor of psychological criminology at Cambridge University. He is the chair of the Campbell Collaboration Crime and Justice Group and president of the Academy of Experimental Criminology. He is a past president of the American Society of Criminology, the British Society of Criminology, and the European Association of Psychology and Law. He has received the Sellin-Glueck and Sutherland awards from the American Society of Criminology for outstanding contributions to criminology. His major research interest is in the development of offending from childhood to adulthood, and he is the director of the Cambridge Study in Delinquent Development, which is a prospective longitudinal survey of 411 London males from age eight to age forty-eight.

NOTE: I am grateful to Bob Boruch, Tom Cook, Cynthia Lum, Anthony Petrosino, David Weisburd, and Brandon Welsh for helpful comments on an earlier draft of this article.

DOI: 10.1177/0002716202250789

tion studies vary in methodological quality. The preferred approach of the Campbell Collaboration Crime and Justice Group is not for a reviewer to attempt to review all evaluation studies on a particular topic, however poor their methodology, but rather to include only the best studies in systematic reviews. However, this policy requires the specification of generally accepted, explicit, and transparent criteria for determining what are the best studies on a particular topic, which in turn requires the development of methodological quality standards for evaluation research.

Methodological quality depends on four criteria: statistical conclusion validity, internal validity, construct validity, and external validity.

In due course, it is possible that methodological quality standards will be specified by the Campbell Collaboration for all its constituent groups. It is also possible that different standards may be needed for different topics. This article is an attempt to make progress in developing methodological quality standards. Unfortunately, discussions about methodological quality standards, and about inclusion and exclusion criteria in systematic reviews, are inevitably contentious because they are seen as potentially threatening by some evaluation researchers. People whose projects are excluded from systematic reviews correctly interpret this as a criticism of the methodological quality of their work. In our systematic reviews of the effectiveness of improved street lighting and closed-circuit television (CCTV) (Farrington and Welsh 2002; Welsh and Farrington 2003 [this issue]), referees considered that the excluded studies were being "cast into outer darkness" (although we did make a list of them).

What are the features of an evaluation study with high methodological quality? In trying to specify these for criminology and the social and behavioral sciences, the most relevant work—appropriately enough—is by Donald Campbell and his colleagues (Campbell and Stanley 1966; Cook and Campbell 1979; Shadish, Cook, and Campbell 2002). Campbell was clearly one of the leaders of the tradition of field experiments and quasi experimentation (Shadish, Cook, and Campbell 2002, p. xx). However, not everyone agrees with the Campbell approach. The main challenge to it in the United Kingdom has come from Pawson and Tilley (1997), who have developed "realistic evaluation" as a competitor. Briefly, Pawson and Tilley argued that the Campbell tradition of experimental and quasi-experimental evalu-

ation research has "failed" because of its emphasis on "what works." Instead, they argue, evaluation research should primarily be concerned with testing theories, especially about linkages between contexts, mechanisms, and outcomes (see below).

Methodological quality standards are likely to vary according to the topic being reviewed. For example, because there have been many randomized experiments on family-based crime prevention (Farrington and Welsh 1999), it would not be unreasonable to restrict a systematic review of this topic to the gold standard of randomized experiments. However, there have been no randomized experiments designed to evaluate the effect of either improved street lighting or CCTV on crime. Therefore, in our systematic reviews of these topics (Farrington and Welsh 2002; Welsh and Farrington 2003), we set a minimum methodological standard for inclusion in our reviews of projects with before-and-after measures of crime in experimental and comparable control areas. This was considered to be the minimum interpretable design by Cook and Campbell (1979).

This was also set as the minimum design that was adequate for drawing valid conclusions about what works in the book *Evidence-Based Crime Prevention* (Sherman et al. 2002), based on the Maryland Scientific Methods Scale (SMS) (see below). An important issue is how far it is desirable and feasible to use a methodological quality scale to assess the quality of evaluation research and as the basis for making decisions about including or excluding studies in systematic reviews. And if a methodological quality scale should be used, which one should be chosen?

This article, then, has three main aims:

1. to review criteria of methodological quality in evaluation research,
2. to review methodological quality scales and to decide what type of scale might be useful in assisting reviewers in making inclusion and exclusion decisions for systematic reviews, and
3. to consider the validity of Pawson and Tilley's (1997) challenge to the Campbell approach.

Methodological Quality Criteria

According to Cook and Campbell (1979) and Shadish, Cook, and Campbell (2002), methodological quality depends on four criteria: statistical conclusion validity, internal validity, construct validity, and external validity. This validity typology "has always been the central hallmark of Campbell's work over the years" (Shadish, Cook, and Campbell 2002, xviii). "Validity" refers to the correctness of inferences about cause and effect (Shadish, Cook, and Campbell 2002, 34).

From the time of John Stuart Mill, the main criteria for establishing a causal relationship have been that (1) the cause precedes the effect, (2) the cause is related to the effect, and (3) other plausible alternative explanations of the effect can be excluded. The main aim of the Campbell validity typology is to identify plausible alternative explanations (threats to valid causal inference) so that researchers can anticipate likely criticisms and design evaluation studies to eliminate them. If

threats to valid causal inference cannot be ruled out in the design, they should at least be measured and their importance estimated.

Following Lösel and Koferl (1989), I have added descriptive validity, or the adequacy of reporting, as a fifth criterion of the methodological quality of evaluation research. This is because, to complete a systematic review, it is important that information about key features of the evaluation is provided in each research report.

Statistical Conclusion Validity

Statistical conclusion validity is concerned with whether the presumed cause (the intervention) and the presumed effect (the outcome) are related. Measures of effect size and their associated confidence intervals should be calculated. Statistical significance (the probability of obtaining the observed effect size if the null hypothesis of no relationship were true) should also be calculated, but in many ways, it is less important than the effect size. This is because a statistically significant result could indicate a large effect in a small sample or a small effect in a large sample.

The main threats to statistical conclusion validity are insufficient statistical power to detect the effect (e.g., because of small sample size) and the use of inappropriate statistical techniques (e.g., where the data violate the underlying assumptions of a statistical test). Statistical power refers to the probability of correctly rejecting the null hypothesis when it is false. Other threats to statistical conclusion validity include the use of many statistical tests (in a so-called fishing expedition for significant results) and the heterogeneity of the experimental units (e.g., the people or areas in experimental and control conditions). The more variability there is in the units, the harder it will be to detect any effect of the intervention.

Shadish, Cook, and Campbell (2002, 45) included the unreliability of measures as a threat to statistical conclusion validity, but this seems more appropriately classified as a threat to construct validity (see below). While the allocation of threats to validity categories is sometimes problematic, I have placed each threat in only one validity category.

Internal Validity

Internal validity refers to the correctness of the key question about whether the intervention really did cause a change in the outcome, and it has generally been regarded as the most important type of validity (Shadish, Cook, and Campbell 2002, 97). In investigating this question, some kind of control condition is essential to estimate what would have happened to the experimental units (e.g., people or areas) if the intervention had not been applied to them—termed the "counterfactual inference." Experimental control is usually better than statistical control.

One problem is that the control units rarely receive no treatment; instead, they typically receive the more usual treatment or some kind of treatment that is different from the experimental intervention. Therefore, it is important to specify the effect size—compared to what?

It does seem useful...to communicate to scholars, policy makers, and practitioners that not all research is of the same quality.

The main threats to internal validity have been identified often but do not seem to be uniformly well known (Shadish, Cook, and Campbell 2002, 55):

1. Selection: the effect reflects preexisting differences between experimental and control conditions.
2. History: the effect is caused by some event occurring at the same time as the intervention.
3. Maturation: the effect reflects a continuation of preexisting trends, for example, in normal human development.
4. Instrumentation: the effect is caused by a change in the method of measuring the outcome.
5. Testing: the pretest measurement causes a change in the posttest measure.
6. Regression to the mean: where an intervention is implemented on units with unusually high scores (e.g., areas with high crime rates), natural fluctuation will cause a decrease in these scores on the posttest, which may be mistakenly interpreted as an effect of the intervention. The opposite (an increase) happens when interventions are applied to low-crime areas or low-scoring people.
7. Differential attrition: the effect is caused by differential loss of units (e.g., people) from experimental compared to control conditions.
8. Causal order: it is unclear whether the intervention preceded the outcome.

In addition, there may be interactive effects of threats. For example, a selection-maturation effect may occur if the experimental and control conditions have different preexisting trends, or a selection-history effect may occur if the experimental and control conditions experience different historical events (e.g., where they are located in different settings).

In principle, a randomized experiment has the highest possible internal validity because it can rule out all these threats, although in practice, differential attrition may still be problematic. Randomization is the only method of assignment that controls for unknown and unmeasured confounders as well as those that are known and measured. The conclusion that the intervention really did cause a change in the outcome is not necessarily the final conclusion. It is desirable to go beyond this

and investigate links in the causal chain between the intervention and the outcome ("mediators," according to Baron and Kenny 1986), the dose-response relationship between the intervention and the outcome, and the validity of any theory linking the intervention and the outcome.

Construct Validity

Construct validity refers to the adequacy of the operational definition and measurement of the theoretical constructs that underlie the intervention and the outcome. For example, if a project aims to investigate the effect of interpersonal skills training on offending, did the training program really target and change interpersonal skills, and were arrests a valid measure of offending? Whereas the operational definition and measurement of physical constructs such as height and weight are not contentious, this is not true of most criminological constructs.

The main threats to construct validity center on the extent to which the intervention succeeded in changing what it was intended to change (e.g., how far there was treatment fidelity or implementation failure) and on the validity and reliability of outcome measures (e.g. how adequately police-recorded crime rates reflect true crime rates). Displacement of offending and "diffusion of benefits" of the intervention (Clarke and Weisburd 1994) should also be investigated. Other threats to construct validity include those arising from a participant's knowledge of the intervention and problems of contamination of treatment (e.g., where the control group receives elements of the intervention). To counter the Hawthorne effect, it is acknowledged in medicine that double-blind trials are needed, wherein neither doctors nor patients know about the experiment. It is also desirable to investigate interaction effects between different interventions or different ingredients of an intervention.

External Validity

External validity refers to the generalizability of causal relationships across different persons, places, times, and operational definitions of interventions and outcomes (e.g., from a demonstration project to the routine large-scale application of an intervention). It is difficult to investigate this within one evaluation study, unless it is a large-scale, multisite trial. External validity can be established more convincingly in systematic reviews and meta-analyses of numerous evaluation studies. Shadish, Cook, and Campbell (2002, 83) distinguished generalizability to similar versus different populations, for example, contrasting how far the effects of an intervention with men might be replicated with other men as opposed to how far these effects might be replicated with women. The first type of generalizability would be increased by carefully choosing random samples from some population as potential (experimental or control) participants in an evaluation study.

The main threats to external validity listed by Shadish, Cook, and Campbell (2002, 87) consist of interactions of causal relationships (effect sizes) with types of persons, settings, interventions, and outcomes. For example, an intervention designed to reduce offending may be effective with some types of people and in some types of places but not in others. A key issue is whether the effect size varies according to whether those who carried out the research had some kind of stake in the results (e.g., if a project is funded by a government agency, the agency may be embarrassed if the evaluation shows no effect of its highly trumpeted intervention). There may be boundary conditions within which interventions do or do not work, or "moderators" of a causal relationship in the terminology of Baron and Kenny (1986). Also, mediators of causal relationships (links in the causal chain) may be effective in some settings but not in others. Ideally, theories should be proposed to explain these kinds of interactions.

Descriptive Validity

Descriptive validity refers to the adequacy of the presentation of key features of an evaluation in a research report. As mentioned, systematic reviews can be carried out satisfactorily only if the original evaluation reports document key data on issues such as the number of participants and the effect size. A list of minimum elements to be included in an evaluation report would include at least the following (see also Boruch 1997, chapter 10):

1. Design of the study: how were experimental units allocated to experimental or control conditions?
2. Characteristics of experimental units and settings (e.g., age and gender of individuals, sociodemographic features of areas).
3. Sample sizes and attrition rates.
4. Causal hypotheses to be tested and theories from which they are derived.
5. The operational definition and detailed description of the intervention (including its intensity and duration).
6. Implementation details and program delivery personnel.
7. Description of what treatment the control condition received.
8. The operational definition and measurement of the outcome before and after the intervention.
9. The reliability and validity of outcome measures.
10. The follow-up period after the intervention.
11. Effect size, confidence intervals, statistical significance, and statistical methods used.
12. How independent and extraneous variables were controlled so that it was possible to disentangle the impact of the intervention or how threats to internal validity were ruled out.
13. Who knows what about the intervention.
14. Conflict of interest issues: who funded the intervention, and how independent were the researchers?

It would be desirable for professional associations, funding agencies, journal editors, and/or the Campbell Collaboration to get together to develop a checklist of items that must be included in all research reports on impact evaluations.

Methodological Quality Scales

Methodological quality scales can be used in systematic reviews to determine criteria for inclusion or exclusion of studies in the review. Alternatively, they can be used (e.g., in a meta-analysis) in trying to explain differences in results between different evaluation studies. For example, Weisburd, Lum, and Petrosino (2001) found disparities between estimates of the effects of interventions from randomized experiments compared with quasi experiments. Weaker designs were more likely to find that an intervention was effective because in these designs, the intervention is confounded with other extraneous influences on offending.

Descriptive validity refers to the adequacy of the presentation of key features of an evaluation in a research report.

There have been many prior attempts to devise scales of methodological quality for impact evaluations, especially in the medical sciences. Moher et al. (1995) identified twenty-five scales devised up to 1993 for assessing the quality of clinical trials. The first of these was constructed by Chalmers et al. (1981), and it included thirty items each scored from 0 to 10, designed to produce a total methodological quality score out of 100. The items with the highest weightings focused on how far the study was a double-blind trial (i.e., how far the participants and treatment professionals knew or did not know about the aims of the study). Unfortunately, with this kind of a scale, it is hard to know what meaning to attach to any score, and the same score can be achieved in many different ways.

Juni et al. (1999) compared these twenty-five scales to one another. Interestingly, interrater reliability was excellent for most scales, and agreement among the twenty-five scales was considerable ($r = .72$). The authors of sixteen scales defined a threshold for high quality, with the median threshold corresponding to 60 percent of the maximum score. The relationship between methodological quality and effect size varied considerably over the twenty-five scales. Juni et al. concluded that this was because some of these scales gave more weight to the quality of reporting, ethical issues, or the interpretation of results rather than to internal validity.

As an example of a methodological quality scale developed in the social sciences, Gibbs (1989) constructed a scale for assessing social work evaluation studies. This was based on fourteen items, which, when added up, produced a score from 0 to

100. Some of the items referred to the completeness of reporting of the study, while others (e.g., randomization, a no-treatment control group, sample sizes, construct validity of outcome, reliability of outcome measure, and tests of statistical significance) referred to methodological features.

The guidance offered by the Centre for Reviews and Dissemination (2001) of the U.K. National Health Service is intended to assist reviewers in the health field. A hierarchy of evidence is presented:

1. Randomized, controlled, double-blind trials.
2. Quasi-experimental studies (experiments without randomization).
3. Controlled observational studies (comparison of outcomes between participants who have received an intervention and those who have not).
4. Observational studies without a control group.
5. Expert opinion.

This guidance includes many methodological points and discussions about criteria of methodological quality, including key questions that reviewers should ask. The conclusions suggest that quality assessment primarily involves the appraisal of internal validity, that is, how far the design and analysis minimize bias; that a minimum quality threshold can be used to select studies for review; that quality differences can be used in explaining the heterogeneity of results; and that individual quality components are preferable to composite quality scores.

The SMS

The most influential methodological quality scale in criminology is the SMS, which was developed for large-scale reviews of what works or does not work in preventing crime (Sherman et al. 1998, 2002). The main aim of the SMS is to communicate to scholars, policy makers, and practitioners in the simplest possible way that studies evaluating the effects of criminological interventions differ in methodological quality. The SMS was largely based on the ideas of Cook and Campbell (1979).

In constructing the SMS, the Maryland researchers were particularly influenced by the methodological quality scale developed by Brounstein et al. (1997) in the National Structured Evaluation of Alcohol and Other Drug Abuse Prevention. These researchers rated each prevention program evaluation on ten criteria using a scale from 0 to 5: adequacy of sampling, adequacy of sample size, pretreatment measures of outcomes, adequacy of comparison groups, controls for prior group differences, adequacy of measurement of variables, attrition, postintervention measurement, adequacy of statistical analyses, and testing of alternative explanations. They also gave each program evaluation an overall rating from 0 (*no confidence in results*) to 5 (*high confidence in results*), with 3 indicating the minimum degree of methodological rigor for the reviewers to have confidence that the results were reasonably accurate. Only 30 percent out of 440 evaluations received a score of 3 to 5.

Brounstein et al. (1997) found that the interrater reliability of the overall quality score was high (.85), while the reliabilities for the ten criteria ranged from .56 (testing of alternative explanations) to .89 (adequacy of sample size). A principal component analysis of the ten criteria revealed a single factor reflecting methodological quality. The weightings of the items on this dimension ranged from .44 (adequacy of sample size) to .84 (adequacy of statistical analyses). In attempting to improve future evaluations, they recommended random assignment, appropriate comparison groups, preoutcome and postoutcome measures, the analysis of attrition, and assessment of the levels of dosage of the treatment received by each participant.

In constructing the SMS, the main aim was to devise a simple scale measuring internal validity that could easily be communicated. Thus, a simple 5-point scale was used rather than a summation of scores (e.g., from 0 to 100) on a number of specific criteria. It was intended that each point on the scale should be understandable, and the scale is as follows (see Sherman et al. 1998):

> Level 1: correlation between a prevention program and a measure of crime at one point in time (e.g., areas with CCTV have lower crime rates than areas without CCTV).

This design fails to rule out many threats to internal validity and also fails to establish causal order.

> Level 2: measures of crime before and after the program, with no comparable control condition (e.g., crime decreased after CCTV was installed in an area).

This design establishes causal order but fails to rule out many threats to internal validity. Level 1 and level 2 designs were considered inadequate and uninterpretable by Cook and Campbell (1979).

> Level 3: measures of crime before and after the program in experimental and comparable control conditions (e.g., crime decreased after CCTV was installed in an experimental area, but there was no decrease in crime in a comparable control area).

As mentioned, this was considered to be the minimum interpretable design by Cook and Campbell (1979), and it is also regarded as the minimum design that is adequate for drawing conclusions about what works in the book *Evidence-Based Crime Prevention* (Sherman et al. 2002). It rules out many threats to internal validity, including history, maturation/trends, instrumentation, testing effects, and differential attrition. The main problems with it center on selection effects and regression to the mean (because of the nonequivalence of the experimental and control conditions).

> Level 4: measures of crime before and after the program in multiple experimental and control units, controlling for other variables that influence crime (e.g., vic-

timization of premises under CCTV surveillance decreased compared to victimization of control premises, after controlling for features of premises that influenced their victimization).

This design has better statistical control of extraneous influences on the outcome and hence deals with selection and regression threats more adequately.

Level 5: random assignment of program and control conditions to units (e.g., victimization of premises randomly assigned to have CCTV surveillance decreased compared to victimization of control premises).

Providing that a sufficiently large number of units are randomly assigned, those in the experimental condition will be equivalent (within the limits of statistical fluctuation) to those in the control condition on all possible extraneous variables that influence the outcome. Hence, this design deals with selection and regression problems and has the highest possible internal validity.

While randomized experiments in principle have the highest internal validity, in practice, they are relatively uncommon in criminology and often have implementation problems (Farrington 1983; Weisburd 2000). In light of the fact that the SMS as defined above focuses only on internal validity, all evaluation projects were also rated on statistical conclusion validity and on construct validity. Specifically, the following four aspects of each study were rated:

Statistical conclusion validity

1. Was the statistical analysis appropriate?
2. Did the study have low statistical power to detect effects because of small samples?
3. Was there a low response rate or differential attrition?

Construct validity

4. What was the reliability and validity of measurement of the outcome?

If there was a serious problem in any of these areas, the SMS might be downgraded by one point. For example, a randomized experiment with serious implementation problems (e.g., high attrition) might receive a rating of level 4 rather than level 5. The justification for this was that the implementation problems had reduced the comparability of the experimental and control units and hence had reduced the internal validity.

External validity was addressed to some extent in the rules for accumulating evidence from different evaluation studies. The overriding aim was again simplicity of communication of findings to scholars, policy makers, and practitioners. The aim was to classify all programs into one of four categories: what works, what doesn't work, what's promising, and what's unknown.

What works. These are programs that prevent crime in the kinds of social contexts in which they have been evaluated. Programs coded as working must have at least two level-3 to level-5 evaluations showing statistically significant and desirable results and the preponderance of all available evidence showing effectiveness.

What doesn't work. These are programs that fail to prevent crime. Programs coded as not working must have at least two level-3 to level-5 evaluations with statistical significance tests showing ineffectiveness and the preponderance of all available evidence supporting the same conclusion.

What's promising. These are programs wherein the level of certainty from available evidence is too low to support generalizable conclusions but wherein there is some empirical basis for predicting that further research could support such conclusions. Programs are coded as promising if they were found to be effective in significance tests in one level-3 to level-5 evaluation and in the preponderance of the remaining evidence.

What's unknown. Any program not classified in one of the three above categories is defined as having unknown effects.

The SMS has a number of problems arising from its downgrading system, which was not explained adequately by Sherman et al. (1997, 1998), and its method of drawing conclusions about effectiveness based on statistical significance (Farrington et al. 2002). Another problem is that it does not explicitly encompass all possible designs. In particular, time series designs are not incorporated adequately. Arguably, a single interrupted time series design (with no control series) is superior to the one-group, pretest-posttest design (level 2). Equally, a comparison between an interrupted time series (i.e., a time series containing an intervention at a specific point) and a control time series containing no intervention is superior to the simple pretest-posttest, experimental-control design (level 3) because the former clearly deals with threats to internal validity (e.g., history, maturation/trends, regression to the mean) more adequately (e.g., Ross, Campbell, and Glass 1970). In principle, this time series design can also address the neglected issue of the time lag between cause and effect as well as the persistence or wearing off of the effects of the intervention over time.

The SMS criteria are not too dissimilar from the methodological criteria adopted by the Center for the Study and Prevention of Violence at the University of Colorado in developing "blueprints" for exemplary violence prevention programs (see www.colorado.edu/cspv/blueprints). Ten violence prevention programs were initially identified as the basis for a national violence prevention initiative because they met very high scientific standards of program effectiveness, defined as follows:

1. a strong research design, defined as a randomized experiment with low attrition and reliable and valid outcome measures;

2. significant prevention effects for violence or for arrests, delinquency, crime, or drug use;
3. replication in at least one additional site with experimental design and significant effects; and
4. sustained effects for at least one year after the treatment.

Other programs were identified as promising if they had significant preventive effects on violence, delinquency, crime, drug use, or predelinquent aggression (e.g., conduct disorder) in one site with a good experimental or quasi-experimental (with a control group) design. Promising programs did not necessarily have to demonstrate sustained effects.

New Methodological Quality Scales

While the SMS, like all other methodological quality scales, can be criticized, it has the virtue of simplicity. It can be improved, but at the cost of simplicity. It does seem useful to use some kind of index of methodological quality to communicate to scholars, policy makers, and practitioners that not all research is of the same quality and that more weight should be given to higher-quality evaluation studies. It seems highly desirable for funding agencies, journal editors, scholarly associations, and/or the Campbell Collaboration to get together to agree on a measure of methodological quality that should be used in systematic reviews and meta-analyses in criminology. This measure could also be used in systematic reviews of studies of the causes of offending.

My own suggestion, put forward rather tentatively to stimulate discussions, is that a new methodological quality scale might be developed based on five criteria:

1. internal validity,
2. descriptive validity,
3. statistical conclusion validity,
4. construct validity, and
5. external validity.

I have placed the criteria in order of importance, at least as far as a systematic reviewer of impact evaluations is concerned. Internal validity—demonstrating that the intervention caused an effect on the outcome—is surely the most important feature of any evaluation research report. Descriptive validity is also important; without information about key features of research, it is hard to include the results in a systematic review. In contrast, information about the external validity of any single research project is the least important to a systematic reviewer since the main aims of a systematic review and meta-analysis include establishing the external validity or generalizability of results over different conditions and investigating factors that explain heterogeneity in effect size among different evaluation studies.

I suggest that it is important to develop a simple score that can be easily used by scholars, practitioners, policy makers, and systematic reviewers. Lösel and Koferl (1989) rated each of thirty-nine threats to validity on a four-point scale (*no threat*,

low threat, *medium threat*, and *high threat*), but these ratings seem too complex to be easily understood or used. One possibility would be to score each of the above five types of validity 0 (*very poor*), 1 (*poor*), 2 (*adequate*), 3 (*good*), or 4 (*very good*). Possibly, the SMS could form the basis of the five-point scale for internal validity. The problem is that as Shadish, Cook, and Campbell (2002, 100) pointed out, there are no accepted measures of the amount of each type of validity. Nevertheless, efforts should be made to develop such measures.

It is important to develop methodological quality standards for evaluation research.

There are many ways of producing a summary score (0-100) from the individual (0-4) scale scores. For example, consistent with my ordering of the importance of the five types of validity, internal validity could be multiplied by eight (maximum 32), descriptive validity by six (maximum 24), statistical conclusion validity by four (maximum 16), construct validity by four (maximum 16), and external validity by three (maximum 12).

A simpler approach would be to develop just three five-point scales covering design (i.e., internal validity), execution (including construct validity, statistical conclusion validity, and sampling elements of external validity), and reporting. Each project could be rated on all three scales, and the systematic review and meta-analysis would determine the generalizability or external validity of results. However, my purpose in this section is less to propose new scales of methodological quality than to suggest that efforts should be made to develop such scales so that they can be widely accepted and widely used to upgrade the quality of both evaluation research and systematic reviews.

Pawson and Tilley's Challenge

As mentioned, the greatest challenge to the Campbell tradition of evaluation, at least in the United Kingdom, has come from the "realistic evaluation" approach of Pawson and Tilley (1994, 1997, 1998). My exposition of their ideas is based mainly on their publications but also on my discussions with them. Their four most important arguments are

1. past evaluation research has failed because of its focus on what works;
2. instead, researchers should investigate context-mechanism-outcome configurations;
3. these configurations should be studied using qualitative, narrative, ethnographic research focusing on people's choices; and
4. the purpose of evaluation projects is to test theories.

Pawson and Tilley's first argument is that past evaluation research has failed because it has produced inconsistent results and has not influenced criminal justice policy. In support of these points, they cite selected case studies such as Martinson's (1974) critique of correctional effectiveness and the fact that criminal justice policies involving increasing imprisonment are not based on the results of criminological evaluations. The following quotations give the flavor of their arguments:

> For us, the experimental paradigm constitutes a heroic failure, promising so much and yet ending up in ironic anticlimax. The underlying logic . . . seems meticulous, clear-headed and militarily precise, and yet findings seem to emerge in a typically non-cumulative, low-impact, prone-to-equivocation sort of way. (Pawson and Tilley 1997, 8)

> Whilst we are at last cleansed from the absurd notion that there can be no positive social influence on the institutions of criminal justice, we do seem to have reached a different sort of lacuna in which *inconsistent results, non-replicability, partisan disagreement* and above all, *lack of cumulation* remain to dash the hopes of evaluators seeking to establish clear, unequivocal guidelines to policy making. Nowhere is this picture revealed more clearly than in so-called meta-analysis. . . . We submit . . . that *methodological failure* is at the root of the capriciousness of evaluation research. (Pawson and Tilley 1994, 291-92)

Much of this argument might be described as shoot the messenger. Even if we accepted Martinson's (1974) claim that correctional interventions were ineffective, this would not necessarily indicate that evaluation methods were faulty. An evaluation project showing no significant difference between experimental and control groups could nevertheless be described as a successful project (assuming that its statistical conclusion validity was adequate). Personally, I do not find their argument at all credible and believe that systematic reviews and meta-analyses in many cases show that some interventions have consistently desirable effects (e.g., Lipsey and Wilson 1998). It seems to me that, Does it work? is the first and most basic question to address in evaluation research and that not addressing this question is like throwing out the baby with the bath water.

Pawson and Tilley's second argument is that evaluation researchers should not study the effectiveness of interventions but should instead investigate relationships among contexts, mechanisms, and outcomes:

> Programs work (have successful outcomes) only in so far as they introduce the appropriate ideas and opportunities (mechanisms) to groups in the appropriate social and cultural conditions (contexts). (Pawson and Tilley 1997, 57)

> The essential idea is that the successful firing of a program mechanism is always contingent on context. (Pawson and Tilley 1998, 80)

A consequence of these arguments is that since the aim is not to determine effect size but to establish context-mechanism-outcome relationships (what works for whom in what circumstances), control conditions are not needed:

> Instead of comparison with some illusory control group, measurement is directed at expected impacts which would follow if the working theories are correct. Measurement will, thus, invariably focus on changes in behavior *within the program group*. (Pawson and Tilley 1998, 89)

They argue that the Campbell tradition places too much emphasis on internal validity (Pawson and Tilley 1997, 27). My own view is as follows:

> Pawson and Tilley argue that measurement is needed only within the program community and that control groups are not needed, but to my mind the one group pretest-posttest design has low internal validity and fails to control for extraneous variables or exclude plausible alternative explanations. (Farrington 1998, 208)

Another argument against the need for control conditions is that they are unnecessary if large decreases in crime are observed in a one-group, pretest-posttest design (although this argument seems inconsistent with the statement that the effect size is unimportant). For example, crime decreased by 72 percent in the Kirkholt project (Forrester, Chatterton, and Pease 1988; Forrester et al. 1990), and it is true that this large decrease seems convincing. However, if the effect size was used as a criterion for including studies in systematic reviews, the resulting estimates of effect size (e.g., in a meta-analysis) would be biased and misleading. Also, how could we decide what size of percentage decrease was so convincing that a control condition was not needed? And how could we know in advance of designing an evaluation that the effect size would be so large that a control condition would be unnecessary?

Few evaluation researchers would disagree with Pawson and Tilley's argument that contexts and mechanisms (or, more generally, moderators and mediators) should be investigated. Bennett (1996, 568) pointed out that Cook and Campbell (1979) recognized the need to study both. After discussing threats to internal validity in quasi-experimental analysis, I concluded that

> it is important to elucidate the causal chain linking the independent and dependent variables. . . . It is desirable to think about possible links in the causal chain in advance of the research and to make plans to test hypotheses where possible. . . . Attempts to replicate key findings are vital, and it may be possible to identify important boundary conditions within which an independent variable has an effect but outside which it does not. (Farrington 1987, 70)

As stated previously, "I agree that it is desirable to establish what works, for whom, in what circumstances and, hence, that it is desirable to study mechanisms and contexts" (Farrington 1998, 206). However, I think that first, the overall effect size should be estimated (e.g., in a meta-analysis), and second, the influence of moderators on that effect size should be studied (including, but not restricted to,

contexts). For example, a blocking design could be used to investigate how far the effects of treatment differ in different subgroups. It could be that an intervention has similar effects with many different types of people in many different types of contexts; Pawson and Tilley's argument that effects always vary with context seems overstated to me, but it should be empirically tested. There are many examples in the literature of multisite programs where the key results were essentially replicated in different sites (e.g., Consortium for Longitudinal Studies 1983). My extensive efforts to investigate interaction effects of risk factors in predicting delinquency (e.g., Farrington 1994; Loeber et al. 1998) produced rather few moderator effects, and the main effects of risk factors on delinquency are highly replicable in different contexts (Farrington and Loeber 1999).

Pawson and Tilley's insistence that effect size is unimportant seems remarkable to me. For example, I quoted to them the results of Petrosino, Turpin-Petrosino, and Finckenauer (2000): seven randomized experiments providing recidivism data on Scared Straight all showed that this program was harmful in the sense that the experimental group had higher recidivism rates. I therefore suggested that we might recommend to governmental policy makers that this intervention program should be abandoned. However, they disagreed, denying that the overall harmful effect was important; they argued that further research should be carried out on the context-mechanism-outcome configurations involved in Scared Straight.

Pawson and Tilley's third argument is that context-mechanism-outcome configurations should be studied in qualitative, narrative, ethnographic research focusing on people's choices:

> Programs work if subjects choose to make them work and are placed in the right conditions to enable them to do so. (Pawson and Tilley 1994, 294)
>
> Social programs are the product of volition, skilled action and negotiation by human agents. (Pawson and Tilley 1997, 50)
>
> Research would be primarily ethnographic with the researcher observing task-forces and working-groups in order to follow through the decision-making process. . . . Qualitative analysis would thus trace foreseen differences in how such collaboration would alter if conducted in the local area office rather than through the distant town hall. . . . What would be sought in this relatively novel (and under-theorized) research territory would be some preliminary narrative accounts of how certain combinations of contextual conditions lead to success (or otherwise). (Pawson and Tilley 1998, 87)

My own comments are as follows:

> Pawson and Tilley's approach seems to involve the formulation and testing of a large number of idiosyncratic hunches about minute context-mechanism-outcome relationships. . . . Their proposal to study a large number of context-mechanism-outcome configurations seems essentially a correlational design, with all the attendant problems in such designs of inferring causality and excluding plausible alternative explanations. (Farrington 1998, 208-209)
>
> Pawson and Tilley suggest that mechanisms essentially provide reasons and resources (the will?) to change behavior. This seems an idiosyncratic view of causal mechanisms. In particular, it is not clear how reasons could be investigated. Many psychologists are reluctant

> to ask people to give reasons for their behavior, because of the widespread belief that people have little or no introspective access to their complex mental processes. . . . Hence, it is not clear that reasons in particular and verbal reports in general have any validity, which is why psychologists emphasize observation, experiments, validity checks, causes and the scientific study of behavior. (Farrington 1998, 207)

Fourth, Pawson and Tilley argued that the main purpose of evaluation research should be to test theories:

> Realist evaluation begins with theory and ends with further theory. Thus we begin with a program theory, framed in terms of mechanisms, contexts and outcome patterns. Specific hypotheses are derived from the theory and these dictate the appropriate research strategy and tactics such as the choice of where detailed measurements of expected impact need to be undertaken. In the light of this empirical test of the theory, it may be confirmed entirely (a rare eventuality), refuted (seldom at a stroke) or refined (the commonest result). . . . The grand evaluation payoff is thus nothing other than improved theory, which can then be subjected to further testing and refinement, through implementation in the next program. And so the cycle continues. (Pawson and Tilley 1998, 89-90)

It is undoubtedly desirable to test theories about causal mechanisms underlying the effect of an intervention on an outcome. The main problem is that numerous hypotheses can be formulated, and it is difficult to collect adequate data to test many of them. However, it seems to me that Pawson and Tilley have lost sight of the main aim of program evaluation—to assess the effect of an intervention on an outcome—and have converted it into the aim of testing context-mechanism-outcome configurations. I am not at all sure that this should be described as evaluation (still less as "realistic" evaluation), at least as these words are normally defined in the English language. According to the *Shorter Oxford English Dictionary*, "evaluation" means working out the value of something, and "realistic" means representing things as they really are.

My conclusion about Pawson and Tilley's challenge is that it does not require any changes in the Campbell tradition, which already emphasizes the need to study moderators and mediators and to test theories in evaluation research. I would not agree with them that the best method of investigating relationships between contexts, mechanisms, and outcomes is in qualitative, narrative, or ethnographic research. These methods are useful in generating hypotheses, but experimental or quasi-experimental research in the Campbell tradition is needed to test causal hypotheses. Hence, Pawson and Tilley's work does not have any implications for my discussion of methodological quality standards.

Conclusions

It is important to develop methodological quality standards for evaluation research that can be used by systematic reviewers, scholars, policy makers, the mass media, and the general public in assessing the validity of conclusions about the effectiveness of interventions in reducing crime. It is hoped that the develop-

ment of these standards would help to upgrade the quality of evaluation research. All research should not be given equal weight, and criminal justice policy should be based on the best possible evidence. This article has attempted to make progress toward the development of such standards by reviewing types of validity, methodological quality scales, and the challenge of realistic evaluation. The main conclusions are that new methodological quality scales should be developed, based on statistical conclusion validity, internal validity, construct validity, external validity, and descriptive validity, and that Pawson and Tilley's challenge to the Campbell evaluation tradition does not have any implications for methodological quality standards.

References

Baron, Reuben M., and David A. Kenny. 1986. The moderator-mediator variable distinction in social psychology research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology* 51:1173-82.

Bennett, Trevor H. 1996. What's new in evaluation research: A note on the Pawson and Tilley article. *British Journal of Criminology* 36:567-78.

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

Brounstein, Paul J., James G. Emshoff, Gary A. Hill, and Michael J. Stoil. 1997. Assessment of methodological practices in the evaluation of alcohol and other drug (AOD) abuse prevention. *Journal of Health and Social Policy* 9:1-19.

Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Centre for Reviews and Dissemination, U.K. National Health Service. 2001. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews*. 2d ed. York, UK: York Publishing Services.

Chalmers, Thomas C., Harry Smith, Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman, and Alexander Ambroz. 1981. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* 2:31-49.

Clarke, Ronald V., and David Weisburd. 1994. Diffusion of crime control benefits: Observations on the reverse of displacement. In *Crime prevention studies*, Vol. 2, edited by Ronald V. Clarke. Monsey, NY: Criminal Justice Press.

Consortium for Longitudinal Studies. 1983. *As the twig is bent . . . Lasting effects of preschool programs*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Farrington, David P. 1983. Randomized experiments on crime and justice. In *Crime and justice*, Vol. 4, edited by Michael Tonry and Norval Morris. Chicago: University of Chicago Press.

———. 1987. Evaluating area-based changes in policing strategies and laws. *Police Studies* 10:67-71.

———. 1994. Interactions between individual and contextual factors in the development of offending. In *Adolescence in context: The interplay of family, school, peers and work in adjustment*, edited by Rainer K. Silbereisen and Eberhard Todt. New York: Springer-Verlag.

———. 1998. Evaluating "communities that care": Realistic scientific considerations. *Evaluation* 4:204-10.

Farrington, David P., Denise C. Gottfredson, Lawrence W. Sherman, and Brandon C. Welsh. 2002. The Maryland Scientific Methods Scale. In *Evidence-based crime prevention*, edited by Lawrence W. Sherman, David P. Farrington, Brandon C. Welsh, and Doris L. MacKenzie. London: Routledge.

Farrington, David P., and Rolf Loeber. 1999. Transatlantic replicability of risk factors in the development of delinquency. In *Historical and geographical influences on psychopathology*, edited by Patricia Cohen, Cheryl Slomkowski, and Lee N. Robins. Mahwah, NJ: Lawrence Erlbaum.

Farrington, David P., and Anthony Petrosino. 2000. Systematic reviews of criminological interventions: The Campbell Collaboration Crime and Justice Group. *International Annals of Criminology* 38:49-66.

———. 2001. The Campbell Collaboration Crime and Justice Group. *Annals of the American Academy of Political and Social Science* 578:35-49.

Farrington, David P., and Brandon C. Welsh. 1999. Delinquency prevention using family-based interventions. *Children and Society* 13:287-303.

———. 2002. Improved street lighting and crime prevention. *Justice Quarterly* 19:313-42.

Forrester, David H., Michael R. Chatterton, and Ken Pease. 1988. *The Kirkholt Burglary Prevention Project, Rochdale*. Crime Prevention Unit paper 13. London: Home Office.

Forrester, David H., Samantha Frenz, Martin O'Connell, and Ken Pease. 1990. *The Kirkholt Burglary Prevention Project: Phase II*. Crime Prevention Unit paper 23. London: Home Office.

Gibbs, Leonard E. 1989. Quality of study rating form: An instrument for synthesizing evaluation studies. *Journal of Social Work Education* 25:55-66.

Juni, Peter, Anne Witschi, Ralph Bloch, and Matthias Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* 282:1054-60.

Lipsey, Mark W., and David B. Wilson. 1998. Effective intervention for serious juvenile offenders: A synthesis of research. In *Serious and violent juvenile offenders: Risk factors and successful interventions*, edited by Rolf Loeber and David P. Farrington. Thousand Oaks, CA: Sage.

Loeber, Rolf, David P. Farrington, Magda Stouthamer-Loeber, and Welmoet van Kammen. 1998. *Antisocial behavior and mental health problems: Explanatory factors in childhood and adolescence*. Mahwah, NJ: Lawrence Erlbaum.

Lösel, Friedrich, and Peter Koferl. 1989. Evaluation research on correctional treatment in West Germany: A meta-analysis. In *Criminal behavior and the justice system: Psychological perspectives*, edited by Hermann Wegener, Friedrich Lösel, and Jochen Haisch. New York: Springer-Verlag.

Martinson, Robert M. 1974. What works? Questions and answers about prison reform. *Public Interest* 35:22-54.

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

Pawson, Ray, and Nick Tilley. 1994. What works in evaluation research? *British Journal of Criminology* 34:291-306.

———. 1997. *Realistic evaluation*. London: Sage.

———. 1998. Caring communities, paradigm polemics, design debates. *Evaluation* 4:73-90.

Petrosino, Anthony, Carolyn Turpin-Petrosino, and James O. Finckenauer. 2000. Well-meaning programs can have harmful effects! Lessons from experiments of programs such as Scared Straight. *Crime and Delinquency* 46:354-79.

Ross, H. Laurence, Donald T. Campbell, and Gene V. Glass. 1970. Determining the social effects of a legal reform: The British "breathalyzer" crackdown of 1967. *American Behavioral Scientist* 13:493-509.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Sherman, Lawrence W., David P. Farrington, Brandon C. Welsh, and Doris L. MacKenzie, eds. 2002. *Evidence-based crime prevention*. London: Routledge.

Sherman, Lawrence W., Denise C. Gottfredson, Doris L. MacKenzie, John Eck, Peter Reuter, and Shawn Bushway. 1997. *Preventing crime: What works, what doesn't, what's promising*. Washington, DC: U.S. Office of Justice Programs.

———. 1998. *Preventing crime: What works, what doesn't, what's promising*. Research in brief. Washington, DC: U.S. National Institute of Justice.

Weisburd, David. 2000. Randomized experiments in criminal justice policy: Prospects and problems. *Crime and Delinquency* 46:181-93.

Weisburd, David, Cynthia M. Lum, and Anthony Petrosino. 2001. Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Political and Social Science* 578:50-70.

Welsh, Brandon C., and David P. Farrington. 2003. Effects of Closed-Circuit Television on crime. *Annals of the American Academy of Political and Social Science* 587:110-35.