2009

Study Quality Assessment in Systematic Reviews of Research on Intervention Effects

Kathleen Wells

Julia H. Littell
Bryn Mawr College, jlittell@brynmawr.edu

Custom Citation

Wells, Kathleen, and Julia H. Littell. "Study Quality Assessment in Systematic Reviews of Research on Intervention Effects." *Research on Social Work Practice* 19, no. 1 (2009): 52-62, doi: 10.1177/1049731508317278.

Running head: STUDY QUALITY

Study Quality Assessment in Systematic Reviews of Research on Intervention Effects

Kathleen Wells

Case Western Reserve University, kathleen.wells@case.edu

Julia H. Littell

Bryn Mawr College

2009

Research on Social Work Practice, 19 (1): 52-62.

Key words: Study quality assessment, systematic reviews, intervention effects

Authors' Note: We thank Jim Baumohl, David Beigel, and Rob Fischer for helpful comments on a prior version of this article.

Abstract

Objective: The goal of this study is to advance an approach to the assessment of the quality of studies considered for inclusion in systematic reviews of the effects of social-care interventions.

Method: To achieve this objective, quality is defined in relation to the widely accepted validity typology; prominent approaches to study quality assessment are evaluated as to their adequacy.

Results: Problems with these approaches are identified.

Conclusion: A formal, yet explicit, multidimensional approach to assessment grounded in substantive issues relevant to the intervention and the broader context in which it is embedded is promoted. Uncritical and exclusive use of indicators of study quality such as publication status, reporting quality, and single summative quality scores are rejected.

"*Scientific evidence is commonly and properly greeted with objections, skepticism, and doubt…[R]esponsible scientists are responsibly skeptical…This skepticism is itself scrutinized. Skepticism must itself be justified, defended. One needs 'grounds for doubt'" (Rosenbaum, 1995, pp. 9-10).*

There is a pressing need for the identification of evidence-based practices and policies to promote pro-social development and well-being and to ameliorate widespread psycho-social and cultural problems such as substance use, crime, and educational failure. It is equally important to identify practices that are ineffective or harmful, programs that work in some situations and not others, and alternative approaches that have similar outcomes. This knowledge can enhance the impact and value of social care and help policymakers, clinicians, and consumers make informed choices.

The identification of an efficacious and effective intervention requires a synthesis of results of empirical investigations of the intervention. Traditional, narrative reviews of research are vulnerable to well-documented biases (Bushman & Wells, 2001; Littell, 2008) that can be countered by *systematic reviews.* These biases are not corrected by meta-analysis alone (Littell, Corcoran, & Pillai, 2008; Petticrew & Roberts, 2006). A systematic review requires investigators to identify studies of sufficient quality to include in the analysis (Juni, Altman, & Egger, 2001); because, "if the 'raw material' is flawed, then the conclusions of systematic reviews cannot be trusted" (Juni, Altman, & Egger, 2001, p. 42). Study quality assessment is also used in systematic reviews to examine variation in the quality of included studies (Aos, Phipps, Barnoski, & Lieb, 2001: Brestan & Eyberg, 1998; Cooper & Hedges, 1994); because variability

in the quality of included studies may account for as much variability in the results of a systematic review as intervention characteristics (Wilson & Lipsey, 2001).

Scholars agree that the quality of intervention studies varies; however, they disagree as to how to conceptualize and measure study quality (Cooper, 1998; Petticrew & Roberts, 2006). Despite the close to 300 measures of study quality that are available, there is no agreement as to which one is best suited for evaluation of intervention studies (Deeks et al., 2003; Moher et al., 1995; West et al., 2002). Moreover, empirical investigations show that when differing assessment tools are used to evaluate the same study, they yield divergent findings (Herbison, Hay-Smith, & Gillespie, 2006; Juni et al., 1999). This situation weakens the confidence the public may place in the quality assessment and in the findings of any one review.

The number (and the prestige) of groups that have considered how to conceptualize and to measure the quality of studies is one index of the significance and complexity of the issue. The groups include international organizations such as the Cochrane Collaboration (Higgins & Green, 2006) and the Campbell Collaboration (Shadish & Myers, 2004); professional societies, such as the Society for Prevention Research (Flay, Biglan, Boruch, Castro, Gottfredson, Kellam, et al., 2005); and governmental agencies, including the U.S. Centers for Disease Control (Zaza, Briss & Harris, 2005), the U.S. Department of Education's What Works Clearinghouse (U. S. Department of Education, 2006), the U.S. Agency for Healthcare Research Quality (AHRQ; West et al., 2002), the U.S. Department of Health and Human Services (DHHS) Substance Abuse and Mental Health Services Administration (SAMHSA) (specifically, the National Registry of Evidence-Based Programs and Practices) (U. S. Department of Health and Human Services, nd); and the United Kingdom National Health Service (NHS) Health Technology Assessment Programme (Deeks et al., 2003; Sutton, Abrams, Jones, Sheldon, & Song, 1998). Yet, none of these groups has resolved definitively how to conceptualize or to measure the

quality of studies to be included in systematic reviews.

Three major reviews of study quality assessment instruments have been completed within the past 12 years. The earliest review was completed by David Moher and his colleagues (Moher et al., 1995). The review's purpose was to evaluate the instruments used to assess the quality of randomized controlled trials of health-care interventions. The second review was completed by Suzanne West and her colleagues (West et al., 2002). With support from the U. S. Agency for Healthcare Research Quality, the review's purpose was to identify measures that were "high performing". The third review was completed by Jonathan Deeks and his colleagues (Deeks et al., 2003). One of the review's purposes was to evaluate tools that could be used to rate the quality of non-randomized studies of treatment effects (i.e., non-experimental and quasi-experimental designs). Taken together, the reviews show that most study quality assessment tools have not been developed or tested using standard scale development techniques and information is lacking, not surprisingly, on the validity and reliability of these scales.

Purposes and Scope of this Article

Our goal is not to provide a detailed review of measurement tools (see Deeks et al., 2003; Moher et al., 1995; West et al., 2002 for work of this type) but, rather, to review and to evaluate broad approaches to study quality assessment and to advocate one for use in systematic reviews. Our discussion is framed primarily in relation to how reviewers might best decide to include or exclude a given study in a systematic review, but we also comment on how reviewers might reflect variation in the quality of included studies.

We focus on intervention research. Intervention studies seek to identify outcomes that can be attributed to an intervention under ideal (efficacy studies) or "real world" conditions (effectiveness studies). In this work, investigators aim to develop plausible, causal inferences about an intervention's effects.

We refer to the health-care as well as to the social-care literature because much of the discussion of study-quality assessment has occurred within the health-care field, and because the interventions for health care and social care have converged in specific areas of practice (e.g., substance use, mental health, child welfare, and aging). Moreover, the central methodological issues in systematic reviews are identical in both (see, for example, Hawe, Shiell, & Riley, 2004).

To achieve our purpose, we first describe broad approaches to assessing the quality of empirical studies of intervention effects and examine their potential for use in evaluating primary studies for inclusion in systematic reviews. Endorsing recent work in this area (Petticrew & Roberts, 2006; Shadish, Cook, & Campbell, 2002), we then advocate one approach, the "risk of bias" approach, to study quality assessment, for consideration and debate by scholars of systematic review (Higgins & Green, 2006). In this approach, we define study quality primarily in relation to a study's internal validity or the strength of inferences one may draw as to an intervention's effects, although we also note how other types of validity – statistical conclusion validity, construct validity, and external validity are relevant to the study quality assessment in the context of systematic reviews.

Assessment Approaches

Publication Status

Historically, reviewers have used publication as an indicator of study quality. The assumption has been that the peer-review process can be relied upon to guarantee "completeness and accuracy of reporting, analyzing, and interpreting the study design and results" (Brestan & Eyberg, 1998, p. 181).

This is often, but not always, the case. Despite the efforts of journal editors to systematize the review process (for an example, one could refer to the review process for the

journal, *Child Abuse and Neglect*), reviewers' judgments of the same study may differ widely. Thus, some published reports describe investigations of low quality, due to the vagaries of the peer review process (Grayson, 2002). Some are insufficiently detailed so that it is not possible to determine the quality of the investigations (Begg et al., 1996).

Moreover, there is evidence that publication decisions may be affected by factors other than study quality and clarity of reporting. For example, scholars do not always evaluate critically evidence that contradicts their views (Mahoney, 1997; Petticrew & Robers, 2006). As a result, some low-quality studies are published and some high-quality reports do not enter the public domain.

Publication status is also confounded with a bias toward specific types of findings. For example, study reports that contain statistically significant results and those that confirm research hypotheses are significantly more likely to be published than those with null or negative findings, all other relevant factors being equal (Mahoney, 1997; Begg, 1994; Rothstein, Sutton, & Bornstein, 2005). This bias is the result of actions taken by authors, reviewers, and editors of journals (Dickersin, 2005): Authors are more likely to submit significant results for publication; peer reviewers are more likely to recommend and journal editors are more likely to accept for publication those reports that contain statistically significant results and to reject those that do not (for reviews of empirical evidence of publication bias, see Dickersin, 2005; Song et al., 2000; Sutton, 2005; Torgerson, 2006). Similarly, there is evidence of selective reporting of outcomes in publications of randomized controlled trials, such that results that are not statistically significant are reported inadequately or omitted altogether (Chan, Hrobjartsson, Haar, Gotzsche, & Altman, 2004; Scherer, Langenberg, & von Elm, 2007). In short, publication processes may introduce a systematic bias that tends to inflate estimates of intervention effects (Hopewell, McDonald, Clarke, & Egger, 2006; Sutton, 2005).

Following Hannah Rothstein and her colleagues (Rothstein, Sutton, & Bornstein, 2005), the Cochrane Collaboration (Higgins & Green, 2006), and others (Moher et al., 2007; Shea et al., 2007), we endorse the position that publication status should not be used to make the decision as to whether an empirical investigation is included or excluded from a systematic review.

Reporting Quality

Reviewers may be tempted to use the completeness of a study report as an indicator of the quality of the study. Several systems have been developed to assess the quality of reporting on empirical investigations. Reporting guidelines include checklists of topics that must be addressed and sample-flow diagrams that provide readers with the information needed to appraise a study, assess its bias, and evaluate the extent to which its findings could be generalized.

Reporting guidelines for primary research include the Consolidated Standards on Reporting Trials (CONSORT; Begg et al. 1996; Moher, Schultz & Altman, 2001), Transparent Reporting on Evaluations with Nonrandomized Controlled Trials (TREND; Caetano, 2004), and Standards for Reporting of Studies of Diagnostic Accuracy (STARD; Bossuyt, Reitsma, Bruns, Gatsonis, Glasziou, Irwig, et al., 2003). Guidelines for reporting systematic reviews and meta-analysis are also available (see Moher et al., 1999). Items considered essential for appraising the quality of reporting on RCTs (Randomized Controlled Trials) (in the CONSORT statement) differ from those required to appraise reporting on non-RCTs (in the TREND statement) and diagnostic studies (in the STARD statement).

Although completeness and quality are often difficult to distinguish, one cannot stand for the other. A high quality study might also be one that has been reported incompletely. Therefore, reviewers should obtain information that is missing and should clarify information that is confusing by contacting the investigators prior to evaluating the quality of their work.

We conclude that while reporting guidelines might help reviewers to identify the information that they need to obtain prior to conducting an evaluation of the quality of a study, these guidelines should not be used to evaluate such quality.

Design Hierarchies

Several groups have proposed hierarchies of research designs that could be used to rank studies of intervention effects in terms of their quality, typically defined as internal validity, as the example displayed in Table 1 shows. In most design hierarchies, the randomized controlled design is considered

---

Insert Table 1 about here

---

the most trustworthy design for efficacy and effectiveness studies, followed by non-randomized comparison group designs that use parallel cohorts (preferably with some form of matching). Observational studies, such as case-control studies and single-group designs are considered weaker than parallel cohort studies (cf. Shadish, Cook & Campbell, 2002). Non-experimental studies, including case studies, and expert opinion are usually placed at the bottom of evidence hierarchies.

The rationale for these hierarchies appears to be twofold. There is some overlap between overall research design and internal validity; that is, some designs tend to produce more credible causal inferences than others. For example, well-executed RCTs provide better controls for selection bias than other methods (Schultz, Chalmers, Hayes, & Altman, 1995). Second, there is substantial evidence of "method effects;" that is, results of studies of intervention effects can be greatly affected by research methods (Wilson & Lipsey, 2001). Results of RCTs are not consistently approximated with other research designs (Glazerman, Levy, & Myers, 2002; Kunz,

Vist, & Oxman, 2002).

The Maryland Scientific Methods Scale (MSMS), shown in Table 2, is one example of this approach. We chose the MSMS to illustrate this approach because it has been used widely in systematic reviews (c.f., Aos, Phipps, Barnowski, & Leib, 2001). This scale was developed by Lawrence Sherman and his colleagues (Sherman et al., 1998) to assess the level of internal

---

Insert Table 2 about here

---

validity of evaluations of crime prevention programs. It is a single, ordinal scale comprised of five categories. Each category corresponds to a type of research design and is associated with a numerical score. The designs are ranked in terms of their ability to handle threats to internal validity. Level 1 refers to a study that produces a correlation between a prevention program and a measure of crime at one point in time. Such a study is considered to have the lowest internal validity because "the design fails to rule out many threats to internal validity and also fails to establish causal order (Sherman et al., 1998, p. 16)." Level 5 refers to a study that uses a randomized controlled trial. Such a study is considered to have the highest level of internal validity because nearly all threats to the internal validity of the study can be eliminated. The scale can be used to rate study quality in conjunction with four additional questions, three of which focus on statistical conclusion validity and one of which focuses on construct validity (Farrington, Gottfredson, Sherman, & Welsh, 2002).

The design hierarchy approach to study quality assessment has the advantage of simplicity. However, it does not account for sources of bias that may compromise the internal validity of a randomized controlled trial or other research designs. For example, important sources of bias or threats to the internal validity of a study include *selection bias,* or differences

between study groups other than exposure to the treatments under investigation, and *attrition*, or systematic differences between groups in withdrawal from treatment or from outcome assessment. (Whenever drop-outs differ from those who remain in treatment, the latter are no longer representative of the initial treatment group. The comparability of groups is further diminished when participants who withdraw from one group are different from those who withdraw from another (differential attrition). This limits the ability to detect differences that are due to treatment.). Of the two, *selection bias* is probably the most important threat to internal validity in intervention research (Larzelere, Kuhn, & Johnson, 2004; Higgins & Green, 2006; Shadish, Cook, & Campbell, 2002). Schultz and colleagues (Schultz & Grimes, 2002; Shultz, Chalmers, Hayes, & Altman, 1995) have shown that even randomized controlled trials do not always protect against selection bias. Moreover, because validity is a property of the *inferences* that can be drawn from empirical results, neither a property of research designs nor of methods (Shadish et al., 2002), we argue the design hierarchy approach to study quality assessment should not be employed by itself.

Design Features

Scholars have focused on features of study design and implementation that could be linked to study quality. Some emphasize specific features sought in randomized controlled trials (c.f., Chalmers et al., 1981). For example, instead of simply classifying a study as a RCT, an analyst might assess how random allocation was generated (for example, a toss of a coin, use of a random numbers table, or reliance on a computer-generated random assignment procedure), whether and how allocation was concealed until after participants enrolled in the study, how recruitment and enrollment took place, or how much attrition occurred in each group at specific points in time.

The Methodological Quality Rating Scale (MQRS), as shown in Table 3, was developed

by William Miller and his colleagues. It is designed to evaluate "the methodological quality of clinical trials in the alcohol field" (Miller & Wilbourne, 2002, p. 266), defined implicitly in relation to the relative absence of threats to the internal validity of a study

---

Insert Table 3 about here

The MQRS is intended for use with studies in which there is at least one treatment group, a comparison condition, a sampling procedure intended to produce equivalent groups before treatment, and an outcome measure.

The MQRS is comprised of 12 items, with each one reflecting a unique aspect of study quality. An item has two or more response categories. Each response is assigned a number, with lower numbers reflecting lower quality, and these numbers are added to produce a total methodological quality score. The coding manual for the scale is available on the web (www.casaa.unm.edu). The total scale score ranges from 0 to 17. A study that receives a total score of 14 (or higher) is considered a well-designed study (Miller, Andrews, Wilbourne, & Bennet, 1998; Miller & Wilbourne, 2002).

We argue that uncritical use of total scale scores in this and other similar scales is problematic because many of the subscales or components of study quality scales appear to be orthogonal. A single summary score may represent therefore very different qualities and not be particularly meaningful.

Nonetheless, the strategy used by William Miller and his colleagues has several advantages over *a priori* judgments about overall research design---a major one being that it produces more complete (and possibly more objective) information about the conduct of a study (Cooper, 1998). Instruments that capture features of study design may help reviewers to assess

plausible sources of bias in a particular study and to make competent judgements as to whether a study is of sufficient quality to be included in a systematic review.

The Validity Framework

Scholars have also equated study quality with the relative absence of threats to the validity of an intervention study. In this framework, validity may be restricted to internal validity, that is the strength of inferences that may be drawn regarding hypothesized causal relationships (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002), or it may also include the strength of inferences that may be drawn regarding external validity, construct validity, and statistical conclusion validity (c.f., Downs & Black, 1998).

The Quality Index (Downs & Black, 1998), shown in Table 4, is one well-regarded example of this later approach (Deeks et al., 2003; West et al., 2002). The QI is a 27-item checklist designed for use with both randomized controlled trials and observational studies. The index is comprised of the following five subscales (with the number of items comprising each one contained in parentheses): Reporting (10), External (3), Internal Validity – Control of Bias

---

Insert Table 4 about here

---

(7); Validity Internal Validity – Confounding (6); and Power (1). Depending on the subscale, the rater indicates whether the item should be checked as “yes”, “no”, or “unable to determine”. Each item is stated positively; that is, it represents a desired design or reporting feature of randomized controlled trials. Responses can be scored and summed to produce subscales as well as a total score, with higher scores indicating higher quality.

As noted above, the use of a single total score to make the decision as to whether a study should be included in a systematic review, as is possible with the Quality Index, is problematic.

However, the use of the QI or other subscales that assess threats to differing types of validity may well help reviewers to evaluate the quality of studies to include in systematic reviews.

The Risk of Bias Framework

Despite the utility of some of the approaches to study quality assessment we describe above, we concur with scholars of research synthesis (Cooper, 1998; Wortman, 1994) that the ideal approach to study quality assessment is one that encourages reviewers to assess features of a study's design in order to assess the risk of bias or threats to the validity of a study.

While scales might be useful in this approach, the emphasis is on the questions reviewers must pose in order to assess competently the major sources of bias that are of greatest concern in studies of intervention effects (Higgins & Green, 2006). These are selection bias and attrition bias which constitute threats to internal validity and performance and detection bias which constitute threats to construct validity. Selection bias refers to systematic differences between study groups other than exposure to the treatments under investigation; attrition bias refers to systematic differences between groups in withdrawal from treatment or from outcome assessment; performance bias refers to systematic differences in the care provided to groups other than the treatments under investigation; and detection bias refers to systematic differences between groups in the assessment or reporting of outcomes. Based on the reviewers' judgments of these risks, studies are included or excluded from inclusion in a systematic review.

To aid in this process, scholars have sought empirical evidence about sources of bias and about the effectiveness of ways to reduce bias in studies of intervention effects. For example, investigations show that if group allocation can be foreseen or altered by researchers, clinicians, or participants, selection biases are likely to limit the comparability of groups (Schultz & Grimes, 2002). Other design features that have been linked empirically to biased results include procedures for assessing outcome that are not "blind" to group assignment, differential loss of

participants from intervention and control or comparison groups, and the exclusion of data from participants who drop-out of the study prior to its completion from outcome analyses (Schultz et al., 1995).

The approach to study quality taken by the *Cochrane Collaboration*, the international organization that produces and distributes systematic reviews in health care, provides one prominent example of a "risk of bias" approach to study quality assessment. The Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2006) describes an approach to quality assessment that focuses on validity of inferences, explicit descriptions of study design and implementation characteristics, and empirical evidence of bias in studies of interventions effects (Higgins & Green, 2006). Cochrane reviewers are encouraged to extract carefully data on multiple study features, using inter-rater reliability checks to ensure accuracy. They then consider several potential sources of bias.

Of particular concern is the presence of selection bias. Given empirical evidence that allocation methods and allocation concealment relate to selection bias (Schultz et al., 1995; Schultz & Grimes, 2002), relevant questions for assessment of selection bias are as follows: Was the method of allocation random (unpredictable)? Was the method of allocation blind (concealed until after enrollment)? Reviewers rate the quality of allocation concealment as "adequate", "unclear", "inadequate", or "not used".

Cochrane reviewers are also encouraged to rate performance bias, attrition bias, and detection bias. Relevant questions for detection bias, for example, are: Who assessed outcomes? Were assessors blind to the treatment condition? Were results reported selectively or for all outcomes measured?

There are several ways to use these guidelines (Higgins & Green, 2006). Reviewers can identify specific criteria for each type of bias and rate whether all, some, or none of the criteria in

each domain were met. They can (and do) draw on existing study quality assessment devices to aid in this process.

In short, the Cochrane guidelines, building on a risk of bias framework, advocate a multi-dimensional yet flexible approach to study quality assessment. The approach taken represents an advance over an assessment of designs, design features, or the validity framework alone, and it emphasizes the critical role of reviewers' judgments in the study-quality-assessment process. This may be particularly important in some fields of practice where specific features of the problem for which an intervention is designed have implications for how best to investigate the efficacy or effectiveness of the intervention.

The Cochrane Collaboration approach has been adopted by the international Campbell Collaboration (C2) (www.campbellcollaboration.org), and it has been applied to systematic reviews of social, educational, and criminological interventions (Boruch, Petrosino, & Chalmers, 1999). C2, as does the Cochrane Collaboration, encourages reviewers to analyze carefully study design and implementation qualities and to assess risk of bias, paying attention to the particular conceptual, contextual, and methodological issues that arise in the empirical investigations under review. It emphasizes the importance of multi-dimensional approaches to quality assessment of intervention studies, the delineation of specific study features, and the use of empirical methods to explore how these features may influence conclusions of a review (Shadish & Myers, 2004). Not surprisingly, C2 discourages the use of scales and indices that produce a single score to judge study quality (Shadish & Myers, 2004), as does the Cochrane Collaboration (Higgins & Green, 2006). Thus, the approach is flexible yet rigorous, emphasizing a pragmatic rather than a uniform approach to study quality assessment (Morgan, 2007).

Additional Observations: Study Quality Assessment in Systematic Reviews

Tailoring Quality Assessment to Broad Study Purpose

Indeed, it would strengthen the practice of systematic reviews and, by implication the assessment of the quality of studies to be included in such reviews, if it were made clear at the outset whether the review was to examine the efficacy or the effectiveness of an intervention. Although one might argue that the two types are not distinct, but that they exist on a continuum, the questions one would ask to assess the quality of efficacy and the quality of effectiveness studies clearly differ.

The Agency for Healthcare Research and Quality (Gartlehner et al., 2006) recently proposed seven domains that could be examined in order to distinguish the two study types: population under study, eligibility criteria, outcomes of interest, duration of follow-up period, adverse events, sample size, and intention-to-treat analysis. They argue that effectiveness studies, in partial contrast to efficacy studies, should include "a diverse population with the condition of interest" (p. 6); should involve "eligibility criteria .. [to] allow the source population to reflect the heterogeneity of the…population" (p. 6); should include "health outcomes, relevant to the condition of interest" (p. 7); should be of sufficient duration in order to "mimic a minimum length of treatment in a clinical setting" (p. 7); should assess adverse events identified in efficacy trials; should include a sample size of sufficient power to "assess a minimally important difference from a patient perspective" (p. 7); and should include the data for all subjects enrolled initially in the study in the analysis.

Casting Inclusion Criteria

Clarification as to whether a review is to focus on efficacy or effectiveness helps reviewers to identify explicit *a priori* eligibility criteria that studies must meet in order to be included in the review. Using the popular "PICO" framework, reviewers determine which populations, interventions, comparisons, and outcomes will be included and excluded in a review (Higgins & Green, 2006). These issues relate to the *external* and *construct validity* of a review,

and guide the search for and identification of relevant studies. In our view, decisions abut the scope and boundaries of a review should be driven by its central questions and objectives.

With clear objectives in mind, reviews of intervention effects must contend with concerns about *internal validity.* Reviewers must decide which overall study designs and/or methodological features are critical to support credible inferences about intervention effects. Reviewers cast inclusion criteria in differing ways. Some limit systematic reviews of intervention effects to RCTs, whereas others include observational studies such as those that depend on cohort or case-control designs. Some reviewers limit included studies to those with specific design features thought to be essential in a particular context (e.g., blinded assessment, low attrition, and/or intention-to-treat analysis). These choices vary, depending in part on the plausible threats to validity and nature of available research in different fields of practices, cultures, and geopolitical contexts.

Regarding *statistical conclusion validity*, it is important to note that systematic reviews in general, and meta-analysis in particular, can overcome some problems that occur in primary studies that lack sufficient statistical power and/or appropriate statistical analyses. Although these issues (power and analysis) are important for critical appraisal of primary studies of intervention effects, they should not constitute initial barriers to inclusion in systematic reviews.

Assessing Variation in Methodological Quality

Irrespective of how investigators cast the inclusion criteria, inevitably there will be variations in design features among studies included in a systematic review. These variations are often considered possible explanations for heterogeneity of effects among studies (Lipsey & Wilson, 2001). As mentioned above, methodological characteristics can account for substantial portions of the variance in effect sizes (Lipsey & Wilson, 2001).

Study qualities are captured during the data extraction phase of a review, by rating or

ranking studies according to pre-specified methodological characteristics. Several study quality assessment instruments could be used for this purpose. Since there is no agreed-upon standard, reviewers have considerable latitude to select and adapt devices for this purpose. Double-extraction and coding of data from primary studies, assessment of inter-rater agreement, and the development of consensus ratings is considered best practice (Higgins & Green, 2006; Littell, Corcoran, & Pillai, 2008). Reviewers should provide a descriptive analysis of variations in study qualities. Many Cochrane and Campbell reviews include a table of study qualities (e.g., Smedslund et al., 2006).

Despite the field's move toward a multi-dimensional approach to study quality assessment, some reviewers have used overall quality scores to weight results of studies in meta-analysis (e.g., Aos et al., 2001). There is little support, however, for this approach and meta-analysts have argued that the use of overall quality scores in meta-analysis should be abandoned (Herbison et al., 2006). Instead, most meta-analysts use individual items that represent design features or risk of bias in sensitivity and moderator analyses (for examples, see Smedslund et al., 2006; Shadish & Baldwin, 2005; Wilson, MacKenzie & Mitchell, 2005)

Discussion and Applications to Social Work

Use of a multi-dimensional approach to study quality assessment will advance the science of systematic reviews of health-care and social-care interventions including those directly relevant to social work or to social welfare programs.

A systematic review of welfare-to-work programs conducted by a Norweigan team is an exemplar of the multidimensional approach (Smedslund, et al., 2006). In order to assess the quality of a study, two members of the team evaluated independently the study's characteristics, design, participants, and interventions in order to determine whether or not the study contained an adequate approach to randomization sequence, concealment of randomization sequence,

prevention of performance bias, prevention of detection bias, and attrition bias, and whether the study's investigators had conducted an intent-to-treat analysis. Raters resolved disagreements in ratings through discussion and consultation with a third reviewer, if needed.

Based on meta- and moderator- analyses of study data, the team concluded their report with a nuanced discussion of the status of randomized controlled studies of welfare-to-work programs within the United States in contrast to those in other industrialized nations; the reliability of administrative data typically used in the investigations under study; the robustness of the effects found in light of the relative absence of knowledge of the extent to which increases in income from work offset decreases in income from welfare payments; the extent to which findings may be generalized to societies outside the United States; and the difficulty of determining whether welfare-to-work programs are voluntary or mandatory and, hence, understanding whether one approach or the other is more effective. (They might have added further caveats reflecting the variability in welfare-to-work programs within the United States.)

Thus, reviewers thought critically and carefully about the strengths and limitations of the welfare-to-work studies they identified. They concluded with a commitment to reevaluate the review every two years thereby showing one way in which science is an ongoing process and systematic reviews, rather than providing the "last word" on a topic, contribute to a continuing dialogue among the community of scholars concerned with a social welfare problem. Systematic reviews can contribute to debates about the importance of methodological features and intervention characteristics; they can identify anomalies in a body of research that may challenge shared beliefs.

In conclusion, efforts to standardize the definition of study quality, to treat it as a uni-dimensional construct amenable to scale development, or to reduce its assessment to simple procedures are unlikely to be successful. Assessment of study quality must be integrated into the

work of practitioners of science. It requires publicly available data-bases, full and complete scientific reports**,** transparent study quality assessment reports, and dialogue and empirical study as to how to evaluate the quality of evidence. These are essential components of a credible evidence-base for policy and practice.

References

Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). The comparative costs and benefits of programs to reduce crime (Version 4.0). *Document Number 01-05-1201. Washington State Institute for Public Policy. http://www.wa.gov.wsipp*.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 399-409). New York: Russell Sage Foundation.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association, 276*, 637-639.

Boruch, R., Petrosino, A., & Chalmers, I. (1999). The Campbell Collaboration: A proposal for systematic, multinational, and continuous reviews of evidence. In P. Davis, A. Petrosino & I. Chalmers (Eds.), *The effects of social and educational interventions: Developing an infrastructure for international collaboration to prepare, maintain and promote the accessibility of systematic reviews of relevant research* (pp. 1-22). London: London University College London School of Public Policy.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41-44.

Brestan, E. V., & Eyberg, S. M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology, 27*(2), 180-189.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personal and Social Psychology Bulletin, 27*, 1123-1130.

Caetano, P. (2004). Standards for reporting non-randomized evaluations of behavioral and public health interventions: the TREND statement. *Society for the Study of Addiction, 99*, 1075-1080.Carlton, P. L., & Strawderman, W. E. (1996). Evaluating cumulated research I: The inadequacy of traditional methods. *Biological Psychiatry, 39,* 65-72.

Chalmers, T. et al. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2*, 31-49.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chan, A. W., Hrobjartsson, A., Haar, M. T., Cotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291, 2457-2465.

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

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews (3rd ed.).* Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. (Eds.). (1994). *The handbook of research synthesis*. NY: Russell Sage Foundation.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technol Assess, 7*(27). http://www.hta.nhsweb.nhs.uk/fullmono/mon727.pdf

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: John Wiley & Sons.

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

Farrington, D., Gottfredson, D., Sherman, L., & Welsh, B. (2002). The Maryland Scientific Methods Scale. In L. Sherman, D. Farrington, B. Welsh, & D. MacKenzie, *Evidence-based crime prevention*, pp. 13-21. London: Routledge.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*, 151-175.

Gartlehner, G., Hansen, R., Nissman, D., Lohr, K., & Carey, T. (2006). *Criteria for distinguishing effectiveness from efficacy trials in systematic reviews*. Technical Report 12 (Prepared by the RTI-International-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). AHRQ Publication No. 06-0046. Rockville, MD: Agency for Healthcare Research and Quality.

Glazerman, S., Levy, D., Myers, D. (2002). *Nonexperimental replications of social experiments: A systematic review*. Washington, DC: Mathematica Policy Research, Inc.

Grayson, L. (2002). Evidence based policy and the quality of evidence: Rethinking peer review. ESRC UK Centre for Evidence Based Policy and Practice, Department of Politics, Queen Mary, University of London. Retrieved December 3, 2006 from http://www.evidencenetwork.org.

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: How "out of control" can a randomised controlled trial be? *British Medical Journal, 328*, 1561-1563.

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*, 1249-1256.

Higgins, J., & Green, S. (Eds.). (2006). *Cochrane handbook for systematic reviews of interventions 4.2.6*. The Cochrane Library, Issue 4. Chichester, UK: John Wiley & Sons, Ltd. Retrieved December 3, 2006, from http://www.cochrane.org/Resourceshandbook/handbookpdf.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2006). Grey literature in meta-analyses of randomized trials of health care interventions. In *The Cochrane Database of Systematic Reviews, 2006, Issue 2*. Chichester, UK: John Wiley & Sons, Ltd.

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *British Medical Journal, 323*(7303), 42-46.

Jüni, P., Witschi, A., Bloch, R., Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1061.

Kunz, R., Vist, G., & Oxman, A. D. (2002). Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Methodology Reviews* 2002, Issue 4. Art. No.: MR000012. DOI: 10.1002/14651858.MR000012.

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin, 130*(2), 289-303.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications.

Littell, J. H. (2008). How do we know what works? The quality of published reviews of evidence-based practices. In D. Lindsey & A. Shlonsky (Eds.), *Child Welfare Research: Advances for Practice and Policy* (pp. 66-93). New York: Oxford University Press.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis.* New York: Oxford University Press.

Mahoney, M. J. (1997). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Miller, W., Andrews, N., Wilbourne, P., & Bennet, M. (1998). A wealth of alternatives: Effective treatments for alcohol problems. In W. Miller & N. Heather (Eds.), *Treating addictive behaviors, 2nd edition* (pp. 203-216). New York: Plenum.

Miller, W., & Wilbourne, P. (2002). Mesa Grande: A methodological analysis of clinical trials of treatments for alcohol use disorders. *Addiction, 97*, 265-277.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F., et al. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *The Lancet, 354*, 1896-1900.

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

Moher, D., Schulz, K., Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of parallel-group randomized trials. *Annals of Internal Medicine, 134*(8), 657-662.

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Med, 4*(3), e78.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48-76.

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences. A practical guide*. Malden, MA: Blackwell.

Rosenbaum, P. (1995). *Observational studies*. New York: Springer.

Rothstein, H., Sutton, A. J., & Bornstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley.

Scherer, R. W., Langenberg, P., & von Elm, E. (2007). *Full publication of results initially presented in abstracts*. Cochrane Database of Systematic Reviews, Issue 2.

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

Schulz, K. F., & Grimes, D. A. (2002). Allocation concealment in randomised trials: Defending against deciphering. *The Lancet, 359*, 614-618.

Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 73*, 6-14.

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

Shadish, W., & Myers, D. (2004). Campbell Collaboration research design policy brief. http://www.campbellcollaboration.org/MG/ResDesPolicyBrief.pdf.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., et al. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*. Available at: http://www.biomedcentral.com/1471-2288/7/10.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1998, July). Preventing Crime: *What works, what doesn't, what's promising*. National Institute of Justice Research in Brief. U. S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Smedslund, G., Hagen, K. B., Steiro, A., Johme, T., Dalsbø, T. K., & Rud, M. G. (2006). Work

programmes for welfare recipients. *The Campbell Collaboration Library*. http://www.campbellcollaboration.org/doc-pdf/Smedslund_Workprog_Review.pdf.

Song, F., Eastwood, A. J., Gilbody, S., Duley, L., & Sutton, A. J. (2000). Publication and related biases. *Health Technology Assessment, 4*(10).

Sutton, A. J. (2005). Evidence concerning the consequences of publication and related biases. In H. Rothstein, A. J. Sutton, & M. Bornstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (1998). Systematic reviews of trials and other studies. *Health Technolology Assessment, 2*(19). http://www.hta.nhsweb.nhs.uk/fullmono/mon219.pdf

Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*, 89-102.

U. S. Department of Education (2006). What works clearinghouse evidence standards for reviewing studies. Retrieved June 1, 2007 from http://www.samhsa.gov/review-criteria.htm.

U. S. Department of Health and Human Services. (nd). National registry of evidence-based programs and practices (NREPP) review criteria. Retrieved June 1, 2006 from http://www.whatworksclearinghouse.org/reviewprocess/standards.html

West, S. et al. (2002). *Systems to rate the strength of scientific evidence*. Evidence Report/Technical Assistance No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Agency for Healthcare Research and Quality Publication No. 02-E016. Rockville, MD. Agency for Healthcare Research and Quality. Retrieved August 25, 2006 from http://www.ahrq.gov/clinic/epcsums/strengthsum.htm#contents

Wilson, D., & Lipsey, M. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413-429.

Wilson, D. B., MacKenzie, D. L., & Mitchell, F. N. (2005). Effects of correctional boot camps on offending. *Campbell Collaboration Library*. http://www.campbellcollaboration.org/doc-pdf/Wilson_bootcamps_rev.pdf

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

Zaza, S., Briss, P. A., & Harris, K. W. (2005). *The guide to community preventive services: What works to promote health?* New York: Oxford University Press. Available at: http://www.thecommunityguide.org/library/book/default.htm

Table 1: A design hierarchy

Systematic review and meta-analysis of randomized controlled trials
Randomized controlled trials
Non-randomized parallel cohort studies
Observational studies
Non-experimental studies
Expert opinion

Adapted from Harbour & Miller (2001).

Table 2: Maryland Scientific Methods Scale (Sherman et al., 1998)

Level	Requirement
5	Random assignment of program and control conditions to units
4	Dependent measures obtained before and after the intervention in multiple experimental and control units, controlling for other variables that influence outcomes
3	Dependent measures obtained before and after the intervention in experimental and comparable control conditions
2	Dependent measures obtained before and after the intervention, with no comparable control conditions
1	Correlation between an intervention and a dependent measure at one point in time

Table 3: Methodological Quality Rating Scale (MQRS; Miller & Wilbourne, 2002)

Group allocation	4 = Randomization 3 = Within S counterbalanced 2 = Case control / matching 1 = Quasi-experimental design, arbitrary assignment, sequential cohorts 0 = Violated randomization or nonequivalent groups
Quality control	1 = Treatment standardized by manual, specific training, content coding, etc. 0 = No standardization of treatment specified
Follow-up rate	2 = 85-100% follow-ups complete 1 = 70-84.9% follow-ups complete 0 = <70% follow-ups complete or longest follow-up < 3 months
Follow-up length	2 = 12 months or longer 1 = 6-11 months 0 = Less than 6 months or unspecified
Contact	1 = Personal or telephone contact for at least 70% of completed follow-ups 0 = Questionnaire, unspecified, or completed < 70% of cases
Collaterals	1 = Collaterals interviewed in > 50% of cases 0 = No collateral verification in most cases or unspecified
Objective	1 = Objective verification (records, serum, breath) in > 50% cases 0 = No objective verification in most cases or unspecified
Dropout	Applies to cases that dropped out of treatment after randomization or treatment assignment 1 = Treatment drop-outs are clearly enumerated and/or characteristics of drop-outs are compared with those for completed cases on baseline characteristics 0 = Treatment drop-outs are not reported, or all non-completers were excluded from outcome analyses
Attrition	Applies to cases lost to follow-up after completion of treatment 1 = Cases lost to follow-up are enumerated and (a) considered in outcome in analyses at some follow-up points, (b) outcomes are imputed for lost cases and included in analyses, or (c) characteristics of lost cases are found to be comparable with those for retained cases at baseline or a prior follow-up point 0 = Cases lost to follow-up are not considered in outcome analyses
Independent	1 = Follow-up conducted by independent interviewers blind to group 0 = Follow-up by nonblind, unspecified, or questionnaire data only

Analyses	1 = Acceptable statistical analyses of group differences 0 = No statistical analysis, inappropriate, or unspecified
Multisite	1 = Parallel replication at 2 or more sites with separate research teams 0 = Single site study or comparison of sites offering different treatments

Table 4: Quality Index (Downs & Black, 1998, pp. 382-383)

Reporting: Were the following clearly described? (Y/N)
1. Study hypothesis/aim/objective
2. Main outcomes
3. Characteristics of the participants
4. Interventions of interest
5. Distributions of principal confounders in each group
6. Main findings
7. Estimates of random variability for main outcomes
8. All the important adverse events that may be a consequence of intervention
9. Characteristics of patients lost to follow-up
10. Actual probability values for main outcomes

External validity (Y/N/unable to determine)
11. Were subjects who were asked to participate representative of the entire population from which they were recruited?
12. Were subjects who were prepared to participate representative of the entire population from which they were recruited?
13. Were the staff, places, and facilities representative of the treatment the majority of subjects received?

Internal validity – bias (Y/N/unable to determine)
14. Was an attempt made to blind subjects to the intervention they received?
15. Was an attempt made to blind those measuring main outcomes of the intervention?
16. If any of the results of the study were based on "data dredging" was this made clear?
17. In trials and cohort studies, do analyses adjust for different lengths of follow-up? Or, in case-control studies, is the period between intervention and outcome the same for cases and controls?
18. Were appropriate statistical tests used to assess the main outcomes?
19. Was compliance with the intervention reliable?
20. Were main outcome measures reliable and valid?

Internal validity – confounding (selection bias) (Y/N/unable to determine)
21. For trials and cohort studies, were patients in different intervention groups? For case-control studies, were cases and controls recruited from the same population?
22. For trials and cohort studies, were subjects in different intervention groups? For case-control studies, were cases and controls recruited over the same period of time?
23. Were subjects randomized to intervention groups?
24. Was the randomized intervention assignment concealed from both patients and staff until recruitment was complete and irrevocable?
25. Was there adequate adjustment for confounding in the analyses from which main findings were drawn?
26. Were losses of subjects to follow-up taken into account?

Power
27. Did the study have sufficient power to detect a clinically important effect where the probability for a difference due to chance was less than 5%?