Research in Nursing & Health, 2003, 26, 322–333

Focus on Research Methods

Research Methods: Managing Primary Study Quality in Meta-Analyses

Vicki S. Conn,* Marilyn J. Rantz*

School of Nursing, University of Missouri–Columbia, Columbia, MO 65211
Received 3 March 2003; accepted 8 May 2003

Abstract: Meta-analyses synthesize multiple primary studies and identify patterns of relationships. Differences in primary study methodological quality must be addressed for meta-analysis to produce meaningful results. No single standard exists for addressing these quality variations. Quality measurement scales are fraught with development and application problems. Several strategies have been proposed to address quality. Researchers can set minimum levels for inclusion or require that certain quality attributes be present. An inclusive method is to weight effect sizes by quality scores. This allows the inclusion of diverse studies but relies on questionable quality measures. By considering quality an empirical question, meta-analysts can examine associations between quality and effect sizes and thus preserve the purpose of meta-analysis to systematically examine data. Researchers increasingly are combining strategies to overcome the limitations of using a single approach. Future work to develop valid measures of primary study quality dimensions will improve the ability of meta-analysis to inform research and nursing practice. © 2003 Wiley Periodicals, Inc. Res Nurs Health 26:322–333, 2003

Keywords: meta-analysis; research methods

An information explosion has occurred over the last 50 years as the volume of scientific literature has grown exponentially. Researchers strive to design new studies based on existing knowledge but face the daunting task of summarizing what is known from extant research. Problems with narrative reviews have stimulated interest in quantitative integration of existing research, both as a foundation for future research and as a basis for nursing practice. As the need to systematically synthesize research grows more critical, use of the powerful tool known as meta-analysis is becoming increasingly common (Cooper, 1998). Well-conducted meta-analyses can guide future research and inform practice (Conn & Armer, 1996). The results of meta-analyses are determined both by the studies included and their management in the meta-analysis process. The scientific rigor of potential primary studies varies dramatically, and several strategies have been proposed to address quality (Moher et al., 1999; Saunders, Soomro, Buckingham, Jamtvedt, & Raina, 2003). One solution is to exclude all but the most rigorous studies. Another approach is to include studies of varied

Contract grant sponsor: NIH NINR (to Vicki Conn, principal investigator); Contract grant number: RO1NR07870.
Correspondence to Vicki S. Conn, S317 School of Nursing–MU, Columbia, MO 65211.
*Professor.
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/nur.10092

quality and then address quality through weighting or moderator analysis procedures (Cooper, 1998). No single standard exists for managing this complex issue. In this article we examine strategies for managing the varied quality of primary studies in meta-analysis.

BRIEF OVERVIEW OF META-ANALYSIS

In meta-analysis research, the pooled results of several primary studies are analyzed to provide a quantitative review of existing empirical evidence. Meta-analysis follows a systematic process: (a) formulate the research problem, (b) search for eligible studies, (c) evaluate available data, (d) pool results, (e) quantitatively analyze, and (f) interpret findings taking into account the strengths and limitations of the existing studies (Cooper, 1998). Meta-analysts calculate an overall estimate of the magnitude of association between the variables they study.

Although the overall effect estimate is very important, it is sometimes of equal interest to investigate differences in effect size associated with variations between studies by conducting a moderator analysis. Moderator analysis estimates effect sizes separately for different values of the moderator variable under study. Intervention attributes and characteristics of samples are typical variables for moderator analyses. For example, in a recent meta-analysis Conn, Valentine, & Cooper (2002) reported the overall effect size of interventions to increase physical activity among aging adults. However, the researchers calculated significantly larger effect sizes for particular intervention components (e.g., self-monitoring) and for studies with particular subject characteristics (e.g., patients with specific chronic illnesses). This moderator analysis is especially useful for nursing intervention research, in which a basic intervention often varies somewhat between studies.

The discipline of nursing will realize the immense potential benefits of quantitative synthesis only when meta-analytic methods are applied appropriately to primary studies. Meta-analysts must address the critically important issue of primary study quality during study selection or data management or both. Generally, some of the research reports that are retrieved and assessed for inclusion in meta-analyses will be strong, and others will possess weaknesses. The challenge is to generate the most useful information possible from the existing empirical evidence. In this article we discuss (a) ways that researchers conceive of quality and assess it, (b) associations between study quality and outcomes, and finally, (c) strategies to manage study quality in quantitative syntheses. Although meta-analysis is useful in both intervention and descriptive research, we address the intervention category because studies in this category are often used to direct nursing practice or to suggest further intervention research. Readers interested in further information about meta-analysis may refer to Cooper (1998) or to frequently updated Web sites (e.g., that of the University of Maryland, http://ericae.net/meta/).

METHODOLOGICAL QUALITY

Both consumers of research and the researchers themselves consistently express concerns about methodological quality. The emphasis on quality is consistent with the goals of science to produce valid knowledge (Petersen & White, 1989). This discussion focuses on internal validity aspects of quality because external validity cannot be present without internal validity and because external validity is not an inherent attribute of individual studies (Juni, Altman, & Egger, 2001).

Explicit definitions of quality generally focus on the extent to which studies generate reproducible information (Moher et al., 1995). ''Quality gives us an estimate of the likelihood that the results are a valid estimate of the truth,'' according to Moher et al. (1995, p. 62). Quality is determined by the extent to which study design, conduct, and analysis systematically avoid or minimize potential sources of bias (Moher et al., 1995). Systematic bias can contribute to error, which could favor either the experimental or the control/comparison treatment. A loss of precision may contribute to error, in which potentially efficacious treatments are abandoned as ineffective. Studies with methodological problems can contribute added variability that reduces precision and hampers scientific progress (Detsky, Naylor, O'Rourke, McGeer, & L'Abbe, 1992; Lohr & Carey, 1999; West et al., 2002).

Unfortunately, no gold standard exists for determining the ''true'' scientific rigor of primary studies (Detsky et al., 1992). Most quality dimensions have to do with preventing bias in selection, performance, detection, or attrition (Juni et al., 2001). Table 1 summarizes commonly noted components of intervention research quality (Balk et al., 2002; Chalmers, Celano, Sacks, & Smith, 1983; Juni et al., 2001; Kunz & Oxman, 1998; Moher et al., 1998; Schulz, Chalmers, & Altman,

Table 1. Components of Primary Intervention Research Methodological Quality for Meta-Analysis

Concept	Issues
Sample selection	Sample attributes appropriate for study purpose
	Intervention tested with important subgroups
Recruitment	Recruitment strategy prevents bias
	Description of potential subjects who declined participation
Sample size adequacy	Size adequate to provide a sufficiently precise estimate of effect size
Random assignment	Central system generates an unpredictable assignment sequence
	Allocation concealment/randomization blinding
Comparison group	Nature of the comparison group appropriate for the area of science
	Management of preintervention differences between comparison groups
Blinding/masking	Participants
	Care providers
	Assessors measuring outcomes
	Data analysts
Interventions	Intervention reproducible by others
	Intervention consistent with theory
	Treatment integrity
	Prevention of treatment contamination
Attrition management	Attrition prevented and reported
	Intention to treat analysis
Outcome measures	Objective measures when possible
	Construct validity of instruments ascertainable
	Adequate reliability to provide sufficiently precise estimate of effect size
	Appropriate follow-up period to measure outcomes
	Avoid mono-operation bias, if appropriate
Statistical analysis	Assumptions of analysis consistent with data
	Significance level appropriate given number of tests conducted on data
	Potential confounders not controlled in design addressed in analysis
	Exact test statistic values and *p* levels presented

2002; Schulz, Chalmers, Hayes, & Altman, 1995; Sindhu, Carpenter, & Seers, 1997; West et al., 2002).

The notion of research quality is complex. Although research methods experts list many similar components of quality, their lists are rarely identical, and definitions vary substantially. For example, ''single blinding'' may refer either to participants being unaware of their assignment or to the masking of the people conducting the outcome assessments. Studies may report single, double, triple, or quadruple blinding (Schulz et al., 2002). Some experts have suggested assessing all components of quality (Table 1), but others have argued that only selected aspects of quality are critical (Juni, Witschi, Bloch, & Egger, 1999). Most investigators agree that they must assess both the design and execution of the study for quality. Concerns about quality have contributed to interest in strategies to measure methodological rigor.

Instruments to Measure Quality

Because meta-analysts have sometimes been unsure of how or whether to include primary studies of varying quality in their work, they have created a number of scales for assessing study quality (Balk et al., 2002). Meta-analysts began to develop primary study quality scales in the 1980s (Chalmers et al., 1981). More than 100 of these scales exist for measuring the quality of primary studies (West et al., 2002), and they vary dramatically in size, composition, complexity, and extent of development (Moher, Jadad, & Tugwell, 1996). For instance, the relative weight between categories differs greatly among scales. Juni et al. (1999) found that the relative weight assigned to three common measures (blinding, randomization, and management of dropouts) varied from 0% to 100%. Scoring of items also varied, with some scales using gradation of scores within each item and others scoring only presence or absence.

Few quality measures have been developed using established scale development techniques (Jadad et al., 1996; Moher et al., 1996). The West et al. (2002) review of more than 100 scales found only two instruments developed using standard procedures (Downs & Black, 1998; Sindhu et al., 1997). When developing scales, researchers typically include estimations of reliability and assessments of content and criterion validity.

Establishing the validity of these instruments is challenging work. Unfortunately, no gold standard exists, so criterion validity cannot be established. Psychometric properties of instruments, including interrater reliability, rarely have been documented (Moher et al., 1996).

Meta-analysts should base their selection of a quality scale on instrument attributes and complexity, consistency between items, and an understanding of which characteristics of primary studies are key for the current projects in that area of science. For example, breast cancer treatment researchers often encounter poor compliance and high dropout because of both the acute and the chronic effects of treatment (Liberati, Himel, & Chalmers, 1986). In contrast, typical limitations of pain research include statistically underpowered trials and insufficient follow-up given the common use of wait-list control groups (Morley, Eccleston, & Williams, 1999). In other areas of science, the use of outcome measures with questionable psychometric properties is common (Conn et al., 2002). Meta-analysts working in different areas of science may choose different measures of quality based on their particular concerns. Readers interested in more details on quality measures can review published copies of quality scales (de Vet et al., 1997; Downs & Black, 1998). West et al. (2002) provide an excellent review of existing scales. Saunders et al. (2003) provide an overview of instruments applicable to nonrandomized intervention studies. The proliferation of scales has been accompanied by considerable debate over their usefulness.

Problems with Scale Measures of Overall Quality

As scale developers have failed to use appropriate processes, problems have developed. Some variations among instruments reflect differing conceptions of quality that are seldom explicitly defined (Juni et al., 1999). Only moderate agreement exists about which domains of quality should be included in scales. Most scale items are based on expert opinion because there has been only scant empirical work documenting the link between quality dimensions and study outcomes (Lohr & Carey, 1999). Even within fairly narrow areas of research, investigators may agree only partially on which attributes are the most important. For example, Cooper (1986) asked experts in one substantive field to evaluate the importance of design attributes. An overall average correlation of $r = .47$ was found, reflecting the difficulty in reaching a consensus on how to generate valid knowledge, even within a well-defined area of science.

Because varied ideas of quality give rise to varied scales, it is not surprising that these scales generate discrepant results when applied to studies (Juni et al., 1999). Moher et al. (1996) compared six primary-study quality scales when they were applied to 12 individual studies. Scoring was completed by several raters, who then resolved score differences through consensus and arbitration. They found that scores differed dramatically across scales, with scores ranging from 23% to 74% of the maximum possible score for individual studies. It remains unclear how these very different quality indices should be interpreted.

Some of the scales' scoring procedures are problematic. Assessments of report quality have often been confounded with design quality (Detsky et al., 1992; West et al., 2002). For instance, some scales require that a research report explicitly address a particular quality dimension before allocating points, but other scales automatically provide points unless a report explicitly states that the quality feature was absent from the study (Sindhu et al., 1997). Missing data is common, as research reports often include insufficient design details. Problems with missing data may be managed by using mean rather than summed scores. Sometimes, instruments have been constructed such that weaker research designs that are fully revealed and acknowledged as a limitation may score better in that category than a study with an apparently better design but one about which little detail is offered and whose limitations might be unknown (Sindhu et al., 1997).

Expert opinion determines the weighting of individual items on scales because empirical evidence is lacking (de Vet et al., 1997). It remains unclear how to interpret the highest quality-scale score for a given set of studies. If the highest score is 60 points out of a possible 100, should 60 be considered the highest score for this area of science? The scoring systems of some scales adjust for criteria that are not applicable to a given study, resulting in relative rather than absolute scores (Chalmers et al., 1981). It is unclear whether study quality is a continuous characteristic or whether a threshold effect exists for quality (Detsky et al., 1992). The interpretation of scores is ambiguous.

Most scales result in a single, overall score for quality. Only a few scales contain subscales that profile strengths and weaknesses (Downs & Black,

1998). This is a major limitation. For example, a small study with inadequate power might be an important source of information even though findings lacked statistical significance, whereas a large study with selection bias might be a less valid source. In such cases, relying on a single global quality score may obscure the underlying structure of the multidimensional concept of methodological quality (Lohr & Carey, 1999).

Beyond problems with the scales themselves, the instruments have proven difficult to apply consistently, even in randomized controlled trials. For example, Balk et al. (2002) reported low interrater reliability for the presence of intention-to-treat analysis, randomization location, outcome assessor blinding, inclusion of a statistician, and accounting for confounding variables. Also, some scales are reliable and accurate in some areas of science but not in others (Moher & Olkin, 1995). The potential for bias in the rating process is another concern, prompting some to suggest masking research reports (concealing authorship and institutional affiliation and methods or results sequentially). Little empirical evidence has accumulated to support these procedures (Jadad et al., 1996; McNutt, Evans, Fletcher, & Fletcher, 1990; Moher et al., 1998).

RELATIONSHIPS BETWEEN QUALITY AND STUDY OUTCOMES

Associations between quality dimensions and effects are observational findings. Confounding between quality dimensions and other important aspects of the studies (such as treatments tested) could exist (Juni et al., 1999). Given those limitations, several researchers have examined the link between study quality and outcomes. Findings often have been contradictory.

Scale-Measured Quality and Study Outcomes

When comparing low-quality studies with high-quality studies, some researchers found that low-quality studies underestimated effect sizes compared to high-quality studies (Ortiz et al., 1998). In contrast, other researchers have documented effect sizes 30–50% larger among the low-quality studies (Juni et al., 1999; Kjaergard, Villumsen, & Gluud, 2001; Moher et al., 1995, 1998, 1999). For example, Moher et al. (1998) reported an overall treatment benefit of 39% for all trials, 52% for low-quality trials, and 29% for high-quality trials. The overall effects were reduced to 35% when quality scores weighted estimates. Other researchers found no or limited association between overall quality scores and effect sizes (Balk et al., 2002; Emerson, Burdick, Hoaglin, Mosteller, & Chalmers, 1990; Fergusson, Laupacis, Salmi, McAlister, & Huet, 2000; Sharpe, 1997; Sterne et al., 2002).

The quality of primary studies included in meta-analyses can influence results in unpredictable directions, including masking or even reversing the effect direction (Sterne et al., 2002). For example, Khan, Daya, and Jadad (1996) found that a fertility effect disappeared when they excluded studies of low quality. In other cases, treatment effects were only manifested in high-quality studies. For example, Brown (1992) found a positive effect for educating people with diabetes, but only when low-quality studies were excluded.

Primary study quality may be confounded with other attributes of studies. For instance, Sterne et al. (2002) found that differences in effect-size estimates between unpublished and published trials increased after controlling for trial quality. In contrast, they found a decrease in effect-size-estimate differences between trials in English versus other languages after study quality was controlled for (Sterne et al., 2002). Bias related to study quality is a bigger problem when overall effect sizes are small in randomized controlled trials (Kunz, Neumayer, & Khan, 2002).

Different scales generate diverse assessments of study quality, which may cause inconsistency in efforts to relate study quality to outcome. Juni et al. (1999) compared the results from 25 quality scales that were applied to studies comparing low-molecular-weight heparin with standard heparin. For six quality scales the relative risks were nearly identical for both treatments in high-quality trials, whereas better effects for low-molecular-weight heparin were documented in low-quality trials. Seven scales documented an opposite trend: No intervention differences for low-molecular-weight heparin were found in low-quality trials, but high-quality trials showed evidence of improved outcomes. For the remaining 12 studies, no differences by study quality were documented. The authors noted that these discrepant results were not surprising given the heterogeneous nature of the quality scales.

Inconsistencies may also be related to differences between areas of science. Balk et al. (2002) found that overall measures of study quality were not associated with effect-size differences across

four medical areas. When meta-analyses within each area were considered separately, quality components were related to effect estimates, but the direction of the association was not consistent across the four specialties. These findings suggest associations between quality and effect-size estimates may vary by area of science.

These inconsistent findings regarding the importance of study quality may point to the questionable appropriateness of using overall measures or to problems of grouping studies across variables. Overall quality scores may obscure associations between quality and outcomes. Existing associations between selected dimensions of quality and outcomes may not be apparent because the most important items may have little overall effect on the total quality scores (Juni et al., 2001). The lack of consistent association between overall quality scores and effect-size values may reflect the possibility that different aspects of lower quality may operate to either increase or decrease effect-size estimates. It is possible that associations between individual items cancel each other out when overall scores are used. These concerns have contributed to attempts to link individual quality dimensions with outcomes.

Randomization and Effect Sizes

Most researchers have examined differences among randomized controlled trials, with only a few studies examining the impact of randomization itself. In a meta-analysis of single-intervention trials including both randomized and nonrandomized primary studies, Kunz and Oxman (1998) reported both under- and overestimation of effect sizes in nonrandomized trials. The magnitude of the differences was sizable, ranging from 76% underestimation of effects to 160% overestimation.

Blinding/Masking/Concealment and Study Outcomes

Adequate randomization requires both adequate generation of the allocation sequence and allocation concealment. Failure to mask randomization has been associated with notably larger (e.g., 41% greater) effect sizes (Chalmers et al., 1983; Colditz, Miller, & Mosteller, 1989; Kunz & Oxman, 1998; Schulz et al., 1995; West et al., 2002). In contrast, others have found no association between allocation concealment and effect sizes (Balk et al., 2002; Juni et al., 1999; Linde et al., 1999). Allocation concealment may be most important when investigators possess strong beliefs about the superiority of one treatment or when treatments are compared to control conditions instead of alternative treatments. It is also possible that inadequate concealment is a surrogate measure for other quality aspects of the study (Schulz et al., 1995).

Although experts often suggest masking group assignments from health care providers, few researchers have examined this issue. Van der Heijden, van der Windt, Kleijnen, Koes, and Bouter (1996) found that poor blinding of providers was the most prevalent weakness in studies of shoulder steroid injections. This type of masking may be most important in studies where providers offered compensatory treatments or otherwise confounded the assignment.

A common strategy is to mask those conducting outcome assessments. Researchers have documented increased effect sizes (up to 35%) among studies without masked data collectors (Chalmers et al., 1983; Juni et al., 1999; Kunz & Oxman, 1998; Schulz et al., 1995; West et al., 2002). Blinding may be especially important when the outcome assessment requires at least some subjective judgment.

Management of Dropouts/Withdrawals and Effect Sizes

Although intention-to-treat analysis prevents selective attrition from biasing results, it is particularly difficult to assess (Balk et al., 2002). Neither Schulz et al. (1995) nor Kjaergard et al. (2001) detected a consistent pattern in effect sizes related to exclusions after randomization. The authors noted that this issue is especially poorly reported in primary studies, thus rendering their results unclear.

Conclusions About the Relationship Between Quality and Outcomes

Findings linking quality with outcomes are inconclusive. Problems with scales may contribute to the inconsistency. Further, differences by area of science make generalizations risky. These findings do not suggest that quality is unimportant, rather, that different aspects of quality may be important in different areas of science and that further development of valid scales is essential.

STRATEGIES TO MANAGE QUALITY

To ensure the quality of meta-analysis results, it is important that explicit systems are in place to handle the variable quality of primary studies (Assendelft, Koes, Knipschild, & Bouter, 1995; Moher et al., 1999). Three basic strategies address quality in primary studies. Setting quality thresholds for inclusion in meta-analysis is based on the idea that only studies with certain quality features can contribute valid answers to the research question. Weighting estimates by quality scores allows studies with stronger research methods to make larger contributions to effect-size estimates. Considering quality an empirical question allows examination of differences in effect sizes in relationship to either overall quality or specific quality attributes (e.g., intention-to-treat analysis).

Setting a Quality Threshold

A priori decisions to use explicit selection criteria for whether to include or exclude a study based on its methods are common in meta-analyses. Decisions to include or exclude studies are critical determinants of the validity and generalizability of findings. When quality thresholds are not explicitly stated, they are often implied in exclusion criteria. For example, in Sharpe's (1997) review of 32 published clinical psychology meta-analyses quality exclusions most commonly occurred because of the absence of control groups, confounding of treatment conditions, a lack of random assignment, and invalid measures. Determining which studies to exclude is a challenge. Decisions about exclusion on the basis of quality are often at least somewhat arbitrary. This allows for a potential inclusion bias because study procedures are complex, and research reports frequently lack pertinent details (Fergusson et al., 2000; Sharpe, 1997).

The threshold approach can involve inclusion criteria that are particularly important for an area of science. Another approach is to select *a priori* particular quality-scale cutoff scores. These criterion-referenced approaches are common. Alternatively, a norm-referenced strategy could be used, in which the quality threshold score is determined as derived from a particular set of studies being considered, such as the median score. This strategy would yield the highest-quality studies from a set of primary reports.

It is inadequate to include only published primary studies in meta-analyses. That an article is published in a peer-reviewed journal is an unsatisfactory proxy measure of its quality because the most consistent difference between published and unpublished reports is the statistical significance of the findings, not the methodological quality of the study (Conn, Valentine, Cooper, & Rantz, in press). In some areas of science, dissertation research is poor quality (Vickers & Smith, 2000), whereas in other areas dissertation studies are similar in quality to published research (Conn et al., 2002). Setting explicit criteria will ensure that studies are fairly evaluated for inclusion.

The threshold approach sometimes, to use a colloquial term, ''throws the baby out with the bathwater.'' That is, among the poorer-quality studies excluded are likely be some studies that vary in ways beyond quality (Moher et al., 1996). For example, excluding studies with very small samples may omit projects with highly innovative interventions or studies with difficult-to-recruit subjects. This exclusion could limit the usefulness of moderator analysis to determine whether variations in interventions or samples are associated with effect-size differences.

The threshold approach has a major limitation: Excluding research that may be of lower rigor goes against the scientific habit of examining data—letting the data speak. A strength of meta-analysis is its ability to examine the association between design attributes and effect-size estimates as an empirical question. If strong studies produce systematically different results from weak studies, the results of the strong ones can be believed (Cooper, 1998). But if no differences are detected, studies of varied strength can be included in the analysis because they likely provide other variations that may contribute to the value of the findings; these could include different approaches to measuring ambiguous outcomes or varied samples (Cooper, 1998). The practices of including all studies and empirically examining methodology-related differences are consistent with the scientific discovery process (Cooper, 1998).

When arguments are made to exclude studies of inferior quality, it begs the questions of what constitutes research of high quality and how to validly assess quality (Sharpe, 1997). Decisions to exclude studies may be too subjective to be reliable. Studies chosen with the threshold approach often contain quality variations often not addressed in the meta-analysis (Tritchler, 1999). In practice, meta-analysts often combine the threshold approach with other strategies that will be discussed next.

Weighting by Study Quality Scores

The second major strategy for managing the quality of primary studies is the common strategy of weighting effect-size estimates by study attributes. Most syntheses weight effect sizes by some indicator of sample size, such as the inverse of the variance, which gives larger studies more weight in effect-size estimates. Similarly, individual effect estimates can be weighted by quality scores. This would yield a larger impact from higher-quality studies on overall pooled results (Detsky et al., 1992).

Weighting by quality-scale scores is the most common approach in medical meta-analyses (Moher et al., 1999). This practice follows the assumption that studies with deficiencies are less informative and should have less influence on overall outcomes (Tritchler, 1999), and it offers several potential advantages. Weighting by quality scores allows all studies to be included in the synthesis, even very diverse studies. This prevents potential bias in the selection of primary studies. Weighting places more emphasis on studies of greater rigor so that these studies affect the findings more heavily. Use of quality-scale scores as a weight produce less statistical heterogeneity because better-quality trials likely result in a higher signal-to-noise ratio as random variation decreases (Moher et al., 1998).

Decisions about weighting may be difficult. Statistical and empirical justification is lacking when it comes to incorporating quality scores as weights (Detsky et al., 1992; Juni et al., 2001). Using quality scores as weights assumes there is a linear relation between estimates of quality and the weights assigned to response options on the scale (Moher et al., 1998). The scaling relation may not be linear but rather may require more complex scoring and weighting systems (Moher et al., 1998). Further, weighting strategies appropriate in one area of science may not generalize to other fields (Balk et al., 2002).

Weighting by quality scores is impeded by the problems of the scales themselves, as previously described. The lack of interrater reliability among quality scales suggests that considerable subjective assessments are required to complete the scales. Using quality scores that are produced with considerable observer inference muddles objective measures with arbitrary judgments (Greenland, 1994). Another problem with the weighting strategy is that an overall score for quality may mask the effect of potentially important individual components of quality (Fergusson et al., 2000). For example, Fergusson found that overall study quality was not related to effect size, but the exclusion of studies based on control group management was associated with differences in effect-size estimates. Weighting by overall quality may obscure important sources of heterogeneity among study results (Greenland, 1994).

Considering Quality an Empirical Question

The third major approach to addressing the quality of primary studies is to examine the association between quality and effect sizes. In this approach the relationship between outcomes and overall quality scores and/or between effect-size outcomes and specific components of quality can be examined. Considering study quality as a potential moderator of sources of heterogeneity is consistent with meta-analysis as a study of studies, rather than as just a statistical system for combining primary study outcomes into a single effect-size estimate (Greenland, 1994).

Examining the impact of research methods on effect sizes makes quality an empirical question. This is consistent with the rigorous approach to research synthesis that moves beyond narrative reviews to examine effect-size moderators quantitatively (Tritchler, 1999). If the results of methodologically sound studies are different from studies with flaws, the results of the high-quality studies can be believed (Cooper, 1998). If methodological differences are not associated with effect-size variations, then the studies of lower quality may be included because they will likely vary from the higher-quality studies in other ways that may be important for the meta-analysis. For example, studies with small samples could contain pilot tests of novel interventions or might include hard-to-recruit participants. These studies would represent a valuable source of information for the overall synthesis.

Retaining all studies that meet the inclusion criteria allows readers to understand the full range of evidence in the area of science and to decide how much importance to give the evidence. Analyzing the effects of methodological components increases confidence in findings when studies of diverse designs are included. Excluding studies other than randomized controlled trials can be problematic in some areas of science, especially in behavioral research that includes patient populations because withholding treatment may pose ethical problems. Meta-analysts have more confidence in robust findings across diverse designs

that are subject to varied threats to validity. Including all possible studies that address the research question allows maximal use of existing data.

One approach to looking at quality as an empirical question is to focus on the quality scores that scales generate. Researchers can graphically plot the association between scores and effect-size estimates. An alternative is to initiate the meta-analysis with high-quality studies and sequentially add studies of progressively lower quality to graphically portray the relationship between quality and effect-size estimates (Moher et al., 1996). As a form of sensitivity analysis, this strategy tests how robust the results are relative to quality attributes (Oxman, 1994). This analysis is often conducted with overall quality scores but could be completed with quality components.

The component approach identifies specific research methodology dimensions that may be coded reliably from research reports and then subjected to moderator analysis (Juni et al., 2001; Moher et al., 1996). Generally, this approach requires less inferential judgment than formulating overall study quality ratings, and so greater reliability is possible (Cooper, 1998). The component approach allows the synthesist to address topics of relevance to that area of science without having to cope with missing data on scales designed for other areas of science. The component approach may avoid problems associated with overall quality scores related to different weaknesses, which can affect results in diverse directions. It is very important to select components relevant to the area of science for coding and analysis. Wortman (1994) provides an excellent overview of the component moderator analysis threats to validity approach based on the classic Cook and Campbell text (1979).

Generally, the meta-analyst begins moderator analysis by examining associations between quality measures and outcomes because methodological differences may be correlated with substantive differences. Researchers must interpret these findings with care so that they are sure not to interpret a confounded situation as being attributable to the substantive difference. When methodological attributes are found to be important, they can be controlled in the subsequent moderator analysis. For example, Conn et al. (2002) controlled for the interval between intervention completion and outcome measure in a synthesis of exercise behavior change interventions. Subsequent moderator analyses may address substantive differences in interventions or differences related to the samples studied.

Increasingly, meta-analysts are using random-effects models for analysis. Researchers using the random-effects model assume that a study-level variance component is present as an additional source of random influence. The random-effects model allows broader generalization of findings than does the fixed-effects model (Cooper, 1998), which might make the former appropriate when studies of diverse quality are included.

A major strength of this approach is its potential to allow researchers to examine the association between outcomes and specific quality components. As researchers learn more about which components of quality are important, this knowledge will enhance the interpretation of existing research and inform future research methods. The decision to examine the components of quality acknowledges that concerns may vary across areas of science. This strategy hews to the goals of science in that it allows the scientific community to evaluate findings within the context of maximum information about the association between methodological decisions and findings. Cochrane Collaboration reviews predominantly use the component approach to examine quality and effect sizes (Moher et al., 1999).

Limitations exist for this strategy. Problems with scale measures of overall quality (lack of interrater reliability, validity questions) may limit confidence in findings related to overall quality measures. Overall measures may mask interesting effects of individual components of quality on effect-size estimates. The component approach does not suffer from these limitations.

Combination Strategies

Sometimes meta-analysts use a combination of strategies, or mixed approach. For example, researchers may judge some studies as totally inadequate for one area of science and exclude them from the analysis while including studies with varied strengths and weaknesses. This allows investigators to exclude studies so poorly designed or executed that it is difficult to come to any conclusions based on their findings. Studies that are retained may be weighted by quality scores. Alternatively, researchers may analyze quality components for their association with effect size (Petersen & White, 1989). For example, in a recent meta-analysis of interventions to increase physical activity among aging adults, studies in which the measure of the dependent variable was confounded with the independent variable were excluded (Conn et al., 2002). This occurred when

researchers provided frequent center-based supervised exercise sessions as an intervention to increase activity behavior and then measured frequency of participation in the sessions as the indicator of physical activity. Studies like this were excluded. However, the meta-analysts did include studies with single-group pre–post designs because several small pilot studies or demonstration projects added interesting intervention variations to the synthesis. Conn et al. (2002) controlled for this design feature in the analysis. Moher et al. (1998) provided an example of conducting several strategies to deal with study quality. These combination approaches allow meta-analysts to retain studies that have potential to address the research question while examining consistencies and accounting for the variability related to methodological features in a collection of primary studies.

General Considerations in Assessing Methodological Quality

Regardless of which system is used to assess the quality of primary studies, extensive training and pretesting are essential. Reviewers must put their predispositions aside and evaluate primary studies objectively (Cooper, 1998). Multiple raters should assess quality. Researchers should decide *a priori* how they will adjudicate differences in reviewers' opinions. As the complexity of rating systems increases, so do discrepancies (Lohr & Carey, 1999). Researchers should manage rating scales as they do other research instruments: select an instrument with sound psychometric properties and evidence of applicability to the topic, extensively train for and supervise instrument application, and fully disclose instrument-related problems in research reports.

The inadequacy of research reporting is an ongoing challenge (Beck, 1999; Moher, Schulz, & Altman, 2001). The quality of research reports invariably affects assessments of methodological features. Recent efforts to develop standards for research reporting in scientific journals may improve this situation (Beck; Moher et al.). In this article we have discussed strategies that researchers can use to address the quality of primary studies considered for inclusion in meta-analyses. Strategies that assess quality allow conclusions about the cumulative strength of information in a particular area of science. This information can be useful for planning future research, as well as for practice and policy implications.

Future Methodological Research

Findings linking quality attributes to effect-size estimates are contradictory and inconclusive. Future research within areas of science may provide more reliable evidence regarding these associations. The previously described limitations of existing instruments to measure overall quality provide an excellent background for future work to develop better measures of quality. Measures with subscales to address distinct dimensions of study quality are urgently needed. For example, the What Works Clearinghouse funded by the U.S. Department of Education is currently developing a set of standards for evaluating the validity of causal claims. Their instrument to assess quality will possess subscales such as intervention construct validity, comparability of treatment groups, contamination, outcome measure construct validity, and statistical validity (Valentine & Cooper, 2003). Their instrument will yield information consistent with Cook and Campbell's (1979) well-known sets of threats to validity: construct validity, internal validity, external validity, and statistical validity. The presence of explicit definitions of terms is another strength of this approach. This approach will allow meta-analysts to consider design quality components individually as an empirical question while providing tested coding items for assessing quality. Ideal instruments to assess quality will require low inference judgments by focusing on concrete assessments rather than abstract judgments. Psychometric evaluation of such instruments is essential. It is crucial that future work produce valid measures of primary-study quality dimensions if meta-analysis is to better inform research and nursing practice.

CONCLUSIONS

Questions about the quality of studies included in meta-analyses have existed since Glass coined the term *meta-analysis* in 1976. Emphasis on the primary study quality is vital to ensure that future research and nursing practice are based on valid syntheses of existing studies. It is essential that researchers lay out explicit processes for addressing the quality of primary studies so that others may assess how well the process protected against bias or errors (Oxman, 1994). Using the watchwords *caution* and *inclusion* as a rule of thumb, meta-analysts should have a goal of including as much data as possible in their work. At times this may mean including studies that have

methodological weaknesses but that nonetheless may offer valuable information. By including studies of varied quality, meta-analysts can delineate the impact of these variations on outcomes (Detsky et al., 1992). Assessing the quality of primary studies adds a crucial layer of complexity to the process of conducting meta-analysis (Moher et al., 1998). As researchers carefully attend to issues of quality, their work furthers not only substantive science but also our understanding of how methodological decisions influence outcomes. Thus, the science goals of examining all available evidence and building cumulative knowledge may be realized.

REFERENCES

Assendelft, W., Koes, B., Knipschild, P., & Bouter, L. (1995). The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA, 274, 1942–1948.

Balk, E., Bonis, P., Moskowitz, H., Schmid, C., Ioannidis, J., Wang, C., et al. (2002). Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA, 287, 2973–2982.

Beck, C. (1999). Facilitating the work of a meta-analyst. Research in Nursing & Health, 22, 523–530.

Brown, S. (1992). Meta-analysis of diabetes patient education research: Variations in intervention effects across studies. Research in Nursing & Health, 15, 409–419.

Chalmers, T., Celano, P., Sacks, H., & Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. New England Journal of Medicine, 309, 1358–1361.

Chalmers, T., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., et al. (1981). A method for assessing the quality of a randomized control trial. Controlled Clinical Trials, 2, 31–49.

Colditz, G., Miller, J., & Mosteller, F. (1989). How study design affects outcomes in comparison of therapy. Medical Statistics in Medicine, 8, 441–454.

Conn, V., & Armer, J. (1996). Meta-analysis and public policy: Opportunity for nursing impact. Nursing Outlook, 44, 267–271.

Conn, V., Valentine, J., & Cooper, H. (2002). Interventions to increase physical activity among aging adults: A meta-analysis. Annals of Behavioral Medicine, 24, 190–200.

Conn, V., Valentine, J., Cooper, H., & Rantz, M. (In press). Should grey literature be included in meta-analysis? Nursing Research.

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

Cooper, H. (1986). On the social psychology of using research reviews: The case of desegregation and black achievement. In R. Feldman (Ed.), The social psychology of education (pp. 341–363). Cambridge, UK: Cambridge University Press.

Cooper, H. (1998). Synthesizing research (3rd ed.) Thousand Oaks, CA: Sage.

Detsky, A., Naylor, C., O'Rourke, K., McGeer, A., & L'Abbe, K. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. Journal of Clinical Epidemiology, 45, 255–265.

de Vet, H., de Bie, R., van der Heijden, G., Verhagen, A., Sijpkes, P., & Kipschild, P. (1997). Systematic review on the basis of methodological criteria. Physiotherapy, 1997, 284–289.

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

Emerson, J., Burdick, E., Hoaglin, D., Mosteller, F., & Chalmers, T. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clinical Trials, 11, 339–352.

Fergusson, D., Laupacis, A., Salmi, L., McAlister, F., & Huet, C. (2000). What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. International Journal of Technology Assessment in Health Care, 16, 1109–1119.

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

Jadad, A., Moore, R., Carroll, D., Jenkinson, C., Reynolds, D., Gavaghan, D., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials, 17, 1–12.

Juni, P., Altman, D., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. British Medical Journal, 323, 42–46.

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association, 282, 1054–1060.

Khan, K., Daya, S., & Jadad, A. (1996). The importance of quality of primary studies in producing unbiased systematic reviews. Archives of Internal Medicine, 156, 661–666.

Kjaergard, L., Villumsen, J., & Gluud, C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of Internal Medicine, 135, 982–989.

Kunz, R., Neumayer, H., & Khan, K. (2002). When small degrees of bias in randomized trials can mislead clinical decisions: An example of individualizing preventive treatment of upper gastrointestinal bleeding. Critical Care Medicine, 30, 1503–1507.

Kunz, R., & Oxman, A. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. British Medical Journal, 317, 1185–1190.

Liberati, A., Himel, H., & Chalmers, T. (1986). A quality assessment of randomized control trials of primary treatment of breast cancer. Journal of Clinical Oncology, 4, 942–951.

Linde, K., Scholz, M., Ramirez, G., Clausius, N., Melchart, D., & Jonas, W.B. (1999). Impact of study quality on outcome in placebo-controlled trials of homeopathy. Journal of Clinical Epidemiology, 52, 631–636.

Lohr, K., & Carey, T. (1999). Assessing ''best evidence'': Issues in grading the quality of studies for systematic reviews. Joint Commission Journal on Quality Improvement, 25, 470–479.

McNutt, R., Evans, A., Fletcher, R., & Fletcher, S. (1990). The effects of blinding on the quality of peer review. A randomized trial. JAMA, 263, 1371–1376.

Moher, D., Cook, D., Jadad, A., Tugwell, P., Moher, M., Jones, A., et al. (1999). Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses. Health Technology Assessment, 3(12), 1–98.

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., Jadad, A., & Tugwell, P. (1996). Assessing the quality of randomized controlled trials. Current issues and future directions. International Journal of Technology Assessment in Health Care, 12, 195–208.

Moher, D., & Olkin, I. (1995). Meta-analysis of randomized controlled trials. A concern for standards. JAMA, 274, 1962–1964.

Moher, D., Pham, B., Jones, A., Cook, D., Jadad, A., Moher, M., et al. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet, 352, 609–613.

Moher, D., Schulz, K., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA, 285, 1987–1991.

Morley, S., Eccleston, C., & Williams, A. (1999). Systematic review and meta-analysis of randomized controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. Pain, 80, 1–13.

Ortiz, Z., Shea, B., Suarez-Almazor, M., Moher, D., Wells, G., & Tugwell, P. (1998). The efficacy of folic acid and folinic acid in reducing methotrexate gastrointestinal toxicity in rheumatoid arthritis. A meta-analysis of randomized controlled trials. Journal of Rheumatology, 25, 36–43.

Oxman, A. (1994). Checklists for review articles. British Medical Journal, 309, 648–651.

Petersen, M., & White, D. (1989). An information synthesis approach to reviewing literature. In M. Petersen & D. White (Eds.), Health care of the elderly: An information sourcebook (pp. 26–36). Newbury Park, CA: Sage.

Saunders, L., Soomro, G., Buckingham, J., Jamtvedt, G., & Raina, P. (2003). Assessing the methodological quality of nonrandomized intervention studies. Western Journal of Nursing Research, 25, 223–237.

Schulz, K., Chalmers, I., & Altman, D. (2002). The landscape and lexicon of blinding in randomized trials. Annals of Internal Medicine, 136, 254–259.

Schulz, K., Chalmers, I., Hayes, R., & Altman, D. (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment. JAMA, 273, 408–412.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. Clinical Psychology Review, 17, 881–901.

Sindhu, F., Carpenter, L., & Seers, K. (1997). Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. Journal of Advanced Nursing, 25, 1262–1268.

Sterne, J., Juni, P., Schulz, K., Altman, D., Bartlett, C., & Egger, M. (2002). Statistical methods for assessing the influence of study characteristics on treatment effects in ''meta-epidemiological'' research. Statistics in Medicine, 21, 1513–1524.

Tritchler, D. (1999). Modelling study quality in meta-analysis. Statistics in Medicine, 18, 2135–2145.

University of Maryland (2003). Meta-analysis of research studies. Retrieved May 13, 2003, from http://ericae.net/meta.

van der Heijden, G., van der Windt, D., Kleijnen, J., Koes, B., & Bouter, L. (1996). Steroid injections for shoulder disorders: A systematic review of RCTs. British Journal of General Practice, 46, 309–316.

Valentine, J., & Cooper, H. (2003). What Works Clearinghouse Study Design and Implementation. Assessment Device (version 0.6). Washington, DC: U.S. Department of Education. Retrieved from http://www.w-w-c.org.

Vickers, A., & Smith, C. (2000). Incorporating data from dissertations in systematic reviews. International Journal of Technology Assessment in Health Care, 16, 711–713.

West, S., King, V., Carey, T., Lohr, K., McKoy, N., Sutton, S., et al. (2002). Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47 AHRQ (Publication No. 02-E016). [Prepared by Research Triangle Institute–University of North Carolina Evidence-Based Practice Center under Contract No. 290-97-0011.] Rockville, MD: Agency of Healthcare Research and Quality.

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