The article addresses the issue of intercoder reliability in meta-analyses. The current practice of reporting a single, mean intercoder agreement score in meta-analytic research leads to systematic bias and overestimates the true reliability. An alternative approach is recommended in which average intercoder agreement scores or other reliability statistics are calculated within clusters of coded variables. These clusters form a hierarchy in which the correctness of coding decisions at a given level of the hierarchy is contingent on decisions made at higher levels. Two separate studies of intercoder agreement in meta-analysis are presented to assess the validity of the model.

ON THE RELIABILITY OF META-ANALYTIC REVIEWS

The Role of Intercoder Agreement

WILLIAM H. YEATON
University of Michigan

PAUL M. WORTMAN
State University of New York

Meta-analysis, a quantitative method for aggregating data from a set of similar research studies, has become increasingly popular (Bangert-Drowns 1986). There are now over a half-dozen, book-length presentations that discuss quantitative synthesis techniques in the social sciences (e.g., Glass, McGaw, and Smith 1981; Hedges and Olkin 1985; Hunter and Schmidt 1990). In addition, meta-analytic procedures are now routinely discussed in undergraduate methods texts (e.g., Dooley 1990, 308-11; Judd, Smith, and Kidder 1991, 425-49).

AUTHORS' NOTE: *The authors thank Philippe Byosiere, Jack Langenbrunner, Jeannie Storer, Dawn Smith, and Yos Susanto for their assistance in collecting the data for this research. Additional thanks go to David Cordray, Mark Lipsey, and William Stock for their useful comments on earlier drafts. Work on this article was supported by Grant number HS06264 from the Agency for Health Care Policy and Research. Requests for reprints should be sent to Paul Wortman, Department of Psychology, SUNY—Stony Brook, Stony Brook, NY 11794-2500.*

EVALUATION REVIEW, Vol. 17 No. 3, June 1993 292-309

With its increasing popularity, a number of authors have suggested that standards be developed to ensure the quality of meta-analyses (Bullock and Svyantek 1985; Ganong 1987; Sacks et al. 1987). Although there is no consensus on the content of these proposed standards (e.g., Wortman 1987), one of the elements of a well-done meta-analysis consistently acknowledged to be important is the reliability of the data extraction or coding process (Glass, McGaw, and Smith 1981; Matt 1989; Orwin and Cordray 1985). In particular, inaccurate coding of those variables used to calculate an effect size is especially critical because it introduces measurement error that may result in an underestimate of the true effect size (Hedges and Olkin 1985, p. 135).

In their seminal volume, *Meta-Analysis in Social Research*, Glass and his associates (Glass, McGaw, and Smith 1981) urged that "all but the simplest meta-analyses should be subjected to an assessment of reliability (in the rater agreement sense of the word) of the coding procedures" (p. 76). Unfortunately, early meta-analyses often did not include measures reflecting the reliability of the coding process. For example, only one of the meta-analyses in the 1983 *Evaluation Studies Review Annual* (Light 1983), an anthology of the year's best meta-analyses, provided a reliability score. Only 29% of the 21 meta-analytic articles published in a recent 3-year period by the *Psychological Bulletin* reported a measure of intercoder reliability. However, the absence of reliability measures is not unique to psychology. Sacks and his colleagues (Sacks et al. 1987) found that only four of the 86 meta-analyses published in the medical literature supplied clearly measured intercoder agreement.

This article addresses the reliability of the coding process primarily from the perspective of intercoder agreement, because it is the most transparent and widely used of those available. Although percentage agreement can be a biased estimate of reliability when nonoccurrence agreements are included, it is used here primarily for its heuristic value to illustrate an overlooked problem that effects the reliability of all meta-analyses independent of the actual statistic calculated. This approach also allows one to conceptualize the coding of meta-analytic data as a judgmental process leading to either an agreement or a disagreement.

SOME SOURCES OF CODING ERROR

Meta-analyses in the social sciences typically report findings on a wide range of independent and dependent variables, providing a single, overall mean reliability score at best. This practice masks the unreliability of indi-

vidual variables especially those used to calculate measures of effect size that are critical to inference about treatment effectiveness. Fortunately, methodological research addressed a number of important sources of intercoder unreliability (e.g., Horwitz and Yu 1984; Yeaton and Wortman 1984). For example, Kaylor, King, and King (1987) simply coded and reported separate reliability scores on four study characteristics "most subject to coder error" (p. 262). Orwin and Cordray (1985) introduced "confidence judgments" to help demonstrate that coding difficulty as reflected in intercoder agreement is related to "deficient reporting quality." When separate reliability scores were calculated for sets of items at different confidence levels, the results "leave little doubt that confidence and agreement are associated" (Orwin and Cordray 1985, 142). In some instances, the mean percentage agreement rate for high-confidence cases was more than twice that of low-confidence cases (.92 vs. .44).

Although the use of confidence judgments was intended to take into account "the existence of considerable variation in macrolevel reporting quality" (Orwin and Cordray 1985, 137), many sources of unreliability exist. As Orwin and Cordray (1985) acknowledge "good reporting does not guarantee agreement, because coding errors, idiosyncratic interpretations, and so on will sometimes preclude it" (p. 136). Below, we argue that the hierarchical nature of the variables is likely to account for a substantial amount of variability in intercoder reliability scores.

A HIERARCHICAL MODEL OF THE CODING PROCESS

In many cases, the relationship between the variables to be coded in the studies included in a meta-analysis is such that the code given to a particular variable effects the codes for other variables that are related to it. For example, if a coder assigns the wrong value or code to a treatment group, then all of the outcome measures dependent on that code will also be incorrect. In a hierarchical structure of this sort, there will be a dependent relationship between the intercoder reliability scores of variables within different levels of the hierarchy (as will be shown below). This situation presents a challenge for the meta-analyst because existing approaches to intercoder reliability assume the independence of all coding decisions.

In the hierarchical case, the lower or deeper one goes in the hierarchy, the greater the number of judgmental steps implicated in a coding decision. To illustrate the utility of this conceptual model, consider the classic meta-

analysis of Smith and Glass (1977) on the effectiveness of psychotherapy. For example, their Table 1 contains study characteristics or coded variables that produce three hierarchical levels:

Level 1 Condition: Is this treatment or control group data?
Level 2 Condition type: What type of treatment or control group is used?
Level 3 Measure: For that type of condition, what measure is reported?

Each level of the hierarchy consists of a set or cluster of variables. An important level 1 decision might identify a given group as a psychotherapy or a control group. Because all variables should be included in the model, level 1 variables also include noncontingent variables such as "form [and date] of publication," "source of subjects," "diagnosis of client," and the like. All variables *after* level 1 would be contingent on the codes of variables in the preceding levels. Thus in the above example, level 2 variables include 10 general types of therapy (e.g., behavioral therapy) as well as placebo and no-treatment groups that are contingent on the level 1 code of treatment or control conditions. Level 3 consists of four general types of outcome measures (e.g., fear-anxiety reduction) used in the calculation of effect size measures.

It is our general assumption that, in most instances, only three levels are needed to code meta-analytic data. In fact, all of the study variables in the Smith and Glass (1977) meta-analysis could be placed neatly into one of the three levels described above. Moreover, as Landman and Dawes (1982) found in their reanalysis of the Smith and Glass meta-analysis, the use of different measures for different types of therapy or control groups can, indeed, affect study results. However, in those cases where global or summary assessments are involved, a fourth level may be necessary (see below).

In order to compute an effect size, it is first necessary to identify the specific treatment and control groups involved in level 2 as well as the relevant dependent measures or test statistics for these groups (e.g., F, t, etc.) that comprise level 3. Thus the reliability of level 3 is strictly contingent on level 2 decisions. In fact, we argue that what is often characterized as coding difficulty is almost completely reflected by the contingent, hierarchical structure of the variables. That is, incorrect coding of a variable at a given level ensures incorrect coding at all lower levels.

In the Smith and Glass (1977) meta-analysis, for example, if the type of therapy—a level 2 variable—was incorrectly coded as rational-emotive rather than cognitive, then each dependent variable in level 3 associated with cognitive therapy would be coded incorrectly *even* if the correct dependent variable value was used, because it would not represent the correct treatment condition. Alternatively correct coding of a level 3 variable would indicate

that correct coding occurred in *each* of two preceding hierarchical levels. In general, when there is a hierarchical structure to a given variable (e.g., all dependent variables or outcome measures), coding difficulty at one level will be directly reflected in the intercoder reliability scores at all lower levels. Thus coding becomes less accurate with each level of the hierarchy because (as will be demonstrated below) the errors tend to accumulate with each lower level.

Extending the above illustration to correlational research, a level 1 variable must first be correctly identified as an independent or dependent variable or as a moderator. Discriminations within levels 2 and 3 will similarly depend on those made in level 1 (e.g., the particular construct and measure of that construct, respectively). The fact that errors can be made at each of the steps of the hierarchy adds to the variability of intercoder reliability scores reported by different pairs of coders. Fortunately, as the examples discussed below will clearly illustrate, the approach to intercoder reliability we present is applicable in meta-analyses of experimental, quasi-experimental, and correlational studies whether or not the structure of the variables is strictly hierarchical.

Under these circumstances, it is of limited value for those few meta-analyses that do report intercoder reliability to report only an overall measure. According to the hierarchical model, this approach will always *over*estimate the lower reliabilities of individual variables that figure prominently in meta-analysis. The problem of interpreting intercoder reliability is especially critical because the model predicts low reliability on precisely those variables (i.e., means and standard deviations of outcome variables used to calculate an effect size) that directly affect the major conclusions of a meta-analytic study.

The amount of bias or overestimation in such a global estimate of reliability will be greatest when there is a relatively large number of straightforward level 1 variables on which agreement is high and a relatively small number of level 2 and level 3 variables on which agreement is low (e.g., outcomes used in calculating effect size measures). Bias will also be magnified when there is a small number of coding errors that affect key variables in levels 1 and 2 such as treatment and control group membership. Consequently, the validity of many meta-analytic study conclusions, especially those relating specific independent and dependent variables, cannot be adequately assessed because the intercoder reliability reported in meta-analyses are based on both the variables of immediate interest and a mix of other variables within a hierarchical structure. As an illustration noted in the next section, a true reliability of .73 could be reported as .90.

RELATIONSHIP OF THE HIERARCHICAL MODEL TO OTHER RELIABILITY MODELS

As noted above, Orwin and Cordray (1985) present a model that relates "deficient reporting" of individual study characteristics or items to "decreased reliability" in meta-analytic data. They offer support for the hypothesis that deficient microlevel reporting causes decreased reliability by examining interrater reliabilities for 25 study characteristics extracted from a subset of 25 studies used in the Smith, Glass, and Miller (1980) meta-analysis of psychotherapy. Sixteen of these study characteristics were also used as predictor variables in a series of regressions to determine the effect of four reliability adjustments on effect size. Orwin and Cordray's (1985) results revealed a considerable amount of variability in the reliabilities for the coded items and indicated the "relative importance" of these items as predictors of effect size changed markedly when corrected for unreliability."

These results are also consistent with the hierarchical model. In fact, the Smith, Glass, and Miller (1980) regression model employed by Orwin and Cordray (1985) is based on a hierarchical structure of three classes and six subclasses of variables (Orwin and Cordray 1985, 139). More importantly, the hierarchical model predicts considerable variation in reliabilities among variables from level 1 to level 3. The range of results for the studies to be presented below is similar to that reported by Orwin and Cordray in their Table 1. In addition, the hierarchical model utilizes all of the available data whereas Orwin and Cordray were forced to use a reduced regression model due to multicollinearity. The hierarchical model predicts the occurrence of such multicollinearity among variables selected from the same level. It also considerably simplifies the meta-analyst's task by eliminating the need to compute interrater reliabilities for each "item" or variable and by reducing the number of studies needed to estimate coder reliability.

PREDICTING INTERCODER RELIABILITY FROM THE HIERARCHICAL MODEL

The hierarchical structure of coded variables can be represented as a quantitative model that reflects the change in reliability for different levels of the hierarchy. This model allows the reliability at one level of the hierarchy to be used to predict the reliability at other levels. From the preceding discussion, this reliability depends on two factors—the reliability of coding within a given level of the hierarchy and the probability of choosing the correct path to that level of the hierarchy.

From a quantitative perspective, one can represent the estimated, contingent intercoder reliability at level *i* in the hierarchy as

$$r'_i = (r'_{i-1})(p_{i-1}) \; (i > 1), \qquad [1]$$

where r'_i is the estimated intercoder reliability for level i and p_{i-1} is the probability of making a correct coding decision at the $(i - 1)$ level of the hierarchy (i.e., choosing the correct path or branch in the tree from one level to the next). Briefly, to predict the intercoder reliability at a given level *i*, multiply the estimated intercoder reliability at the previous level by the probability of making a correct coding decision (or choosing the correct path) at the previous level.

Here we assume that:

$$r'_1 = r_1,$$

the estimated reliability at level 1 is equal to the observed reliability at that level; and that:

$$p_i = p_{i-1} \; (i > 1),$$

the probability of choosing the correct path is the same for all levels of the hierarchy. This assumption (and the next) is consistent with current approaches to assessing reliability that assume a uniform reliability for all coding decisions (but see above discussion of Orwin and Cordray 1985) and with the notion that unreliability is due to random error because a well-developed coding form and intensive coder training eliminates bias. If we further assume that the probability p_{i-1} is no different than any other coding decision, then $p_{i-1} = r_1$ for $i > 1$. That is the probability of making a correct, contingent coding decision is equal to the observed, average reliability of all coding decisions at level 1. Thus substituting in Equation 1:

$$r'_i = (r'_{i-1})(r_1) \quad (i > 1).$$

$$\text{For } i = 2,\ r'_2 = (r'_1)(r_1) = (r_1)(r_1) = r_1^2, \text{ because } r'_1 = r_1.$$

$$\text{For } i = 3,\ r'_3 = (r'_2)(r_1) = (r_1^2)(r_1) = r_1^3, \text{ because } r'_2 = r_1^2.$$

Following this inductive line of argument, for any $i > 1$,

$$r'_i = r_1^i. \qquad [2]$$

That is, the reliability at any given level *i* is simply the *i*th power of the level 1 reliability. Thus for example, if the observed intercoder reliability is .90 at level 1, then the estimated or predicted reliabilities for levels 2 and 3, respectively, will be:

$r'_2 = .81$ at level 2 ($.90 \times .90$)

$r'_3 = .73$ at level 3 ($.90 \times .90 \times .90$ or $.81 \times .90$).

The logic and algebra of the model developed here are identicial to a similar problem in the physical sciences involving the reliability of electrical circuits. The formula for determining the reliability at any given level of the hierarchy of the circuit can be viewed as a special case of the reliability (R) of a system of parallel elements in a circuit (i.e., items or variables) configured in a series of units (i.e., levels). The general equation for the reliability of such an electrical system is:

$$R = [1 - (1 - p)^m]^n \quad [3]$$

where m is the number of parallel elements in the circuit and n is the number of levels in the system (hierarchy). In the approach taken in this article, $m = 1$ because the unit is considered to act as a single element with reliability $p = r$. These substitutions yield $R = r^n$, which is identical to equation 2. Substituting $n = 1$, $n = 2$, and $n = 3$ into the general equation yields precisely the same reliability equations as derived above from our quantitative model (Von Alven 1964, 202-4).

The quantitative relationship between variables in the model of reliability presented in equation 2 is based on the hierarchical framework. It is precisely this contingent relationship between variables within levels that accounts for the changing intercoder reliabilities. As one goes lower and deeper in the hierarchy, intercoder reliability decreases. Fortunately, it is possible to assess this hierarchical model by comparing the predicted reliabilities for levels 2 and 3 with those actually obtained.

EVALUATING THE HIERARCHICAL MODEL

To test this quantitative model of intercoder reliability, we conducted two formal studies. The first meta-analytic study involved correlational data to examine the relationship between occupational stressors and strains (Susanto, Yeaton, and Wortman 1990). The second consisted of quasi-experimental data to assess mortality and morbidity resulting from carotid endarterectomy, a widely performed surgical procedure (Dyken and Pokras 1984; Langenbrunner 1990).

STUDY 1: THE RELATIONSHIP BETWEEN OCCUPATIONAL STRESSORS AND STRAINS

The first study is drawn from our meta-analysis of occupational stressors and strains, a large literature replete with independent and dependent variables of diverse complexity. Occupational stressors have been implicated in numerous physical and mental problems (Holt 1982). To better understand the nature and strength of the relationship between occupational stressors and strains in their various forms, a meta-analysis of this literature was conducted (Susanto, Yeaton, and Wortman 1990) using statistical techniques developed by Hunter and his colleagues (Hunter, Schmidt, and Jackson 1982).

We were not only interested in a general classification of variables as stressors or strains (i.e., level 1 variables) but also in classifying these stressors or strains into more specific types (level 2 and level 3 variables). It is a relatively simple conceptual task to distinguish a stressor from a strain with a high degree of reliability. However, it is a much more difficult task to discriminate reliably either similar classes of stressors (e.g., role stressors, environmental stressors) or strains (e.g., psychological, physical). And, it is more difficult yet to identify reliably the precise kind of stressor (e.g., role ambiguity) or strain (e.g., job satisfaction).

In this meta-analysis the average correlation between stressors and strains in health care workers was .20 although there was considerable variability in the magnitude of this correlation depending on the context in which it was examined. These contexts included aspects of the study (e.g., year of publication), occupation type (e.g., nurse or other health professional), study design (cross-sectional or longitudinal), as well as the different types of stressors or strains described above. Using the hierarchical approach, we were able to represent the varying degree of relationship between occupational stressors and strains and, in some instances, to identify potential reasons for the varying magnitude of the relationships.

Method

Sample. To determine average reliability within these three levels, a sample of 10 studies was randomly selected from the 50 studies used in the meta-analysis. Two advanced graduate students at the University of Michigan independently coded the variables from these studies. These coded variables provided the basic data for the reliability analyses. Each coded variable was placed into one of three conceptual classes described above: Level 1

variables were classified into the categories stressors, strains, and moderators or as data necessary to calculate a measure of correlational effect size (viz., sample size, correlation coefficient, reliability of the independent and dependent variables); level 2, the general type or category of stressor, strain, or moderator; and level 3, the specific type of stressor, strain, or moderator. The latter information is essential to calculating effect sizes based on level 1 data for the relationship between *specific* stressors and strains. In other words, the effect size variables of level 1 are carried to levels 2 and 3 when addressing questions related to general or specific stressors, strains, and moderators, thereby establishing the contingent relationship between levels.

Intercoder reliability. Three contingent, intercoder reliability scores were calculated, one from each of the three levels of variables, including an overall classification as a stressor, strain, or moderator (level 1), an assignment to a particular subcategory (level 2), and designation of a specific variable type within the subcategory (level 3). Intercoder agreement scores were used as a measure of reliability in the meta-analysis and calculated using the following formula:

$$\text{agreement} = (\text{Number of agreements})/(\text{Number of agreements} + \text{Number of disagreements}) \quad [4]$$

An agreement was scored when both coders reported the same response to a given variable. A disagreement was scored when the responses of the two coders differed.

The problem of calculating reliability in a hierarchical context is independent of the specific statistic used. Although an analogous model might be developed for indexes of correlational agreement, the approach illustrated in this article provides a useful heuristic for addressing problems with coding variables in a hierarchical structure.

Results

The intercoder agreement scores for Levels 1, 2, and 3 are presented in Table 1. The overall reliability across levels was 86.3%. The average intercoder reliability was 90.0% in level 1. In level 2 agreement was 83.3% and in level 3 it was 78.1%. The decreasing trend in average intercoder agreement across these three levels is consistent with the hierarchical model of intercoder reliability and the expected impact of contingent coding decisions in diminishing these averages.

Assuming $r_1 = .900$, the result obtained for level 1, the model predicts intercoder reliabilities of .810 and .729, respectively, for levels 2 and 3. These

TABLE 1: Observed and Predicted Intercoder Reliabilities for the Occupational Stress Meta-Analysis

Level	*N of Coded Variables*	*Observed Reliability*	*Predicted Reliability*	*95% Confidence Interval*
1	630	.900	—	(.884, .914)
2	210	.833	.810	(.787, .870)
3	210	.781	.729	(.722, .829)

predicted reliabilities are very close to the results obtained. The 95% confidence interval (Edwards 1976, 86-89) for each level does not include the observed results at other levels, but does include the predicted results for that level. The lowest reliability score for a single variable computed across studies was .75 (not shown in Table 1).

STUDY 2: ASSESSMENT OF CAROTID ENDARTERECTOMY

The second study of intercoder reliability was conducted as part of a larger investigation into the application of meta-analytic methods to the scientific literature on carotid endarterectomy (CE). CE is a widely performed, but controversial, medical procedure to remove plaque from the walls of the carotid arteries (Warlow 1984). Its primary objective is to increase blood supply to the brain and to prevent stroke. The large majority of the studies are either quasi-experimental or uncontrolled.

Method

Sample. A random sample of six studies was drawn from 55 published studies. These studies were selected and coded independently by two highly trained students at the University of Michigan. As part of that training, each coder was asked to read and code 5 to 10 articles that had been previously read and coded with complete agreement by both authors. These articles were chosen to represent the full range of coding decisions on all variables in the database. Extensive feedback was provided, particularly on incorrectly coded variables. Additional studies were coded until there were no errors on these problematic variables. Coders were strongly encouraged to consult with senior staff when anomalous coding problems arose. Data were coded using a three-page coding form containing 183 separate variables. To ensure macro-level reporting quality, the coding sheet was reviewed by an international

TABLE 2: Observed and Predicted Intercoder Reliabilities for the Carotid Endarterectomy Meta-Analysis

Level	*N of Coded Variables*	*Observed Reliability*	*Predicted Reliability*	*95% Conference Interval*
1	104	.923	—	(.888, .947)
2	141	.865	.852	(.816, .901)
3	70	.771	.786	(.655, .852)

panel of six experts in CE surgical research and revised to incorporate suggested changes.

Intercoder reliability. Intercoder agreement scores were calculated for three levels of variables using all of the coded information. Level 1 variables included basic study data such as author, year of publication, type of sampling, and surgical or control group. Level 2 variables were primarily composed of a variety of subcategories contingent on the level 1 variable, surgery or control, such as the specific type of surgical indication (e.g., asymptomatic). Level 3 variables included specific mortality and morbidity outcomes (e.g., cerebrovascular accidents, myocardial infarctions, and deaths) as well as the time period in which they occurred (e.g., greater than 30 days). The logic for coding these contingent data was as follows: Level 1 required coders to decide whether patients were either surgical or control group members with either similar or mixed indications. Once this decision was made, the coders then had to identify the specific type of indication in level 2. Finally, they had to code the specific outcomes in level 3 for each of these subgroups of patients for the various time periods.

Results

The overall average agreement between coders for the six studies was 86.3%. The results for each level were presented in Table 2. For level 1, intercoder agreement was 92.3%. The agreement scores for levels 2 and 3 were 86.5% and 77.1%, respectively. The 95% confidence intervals for these results do not include the observed results for the other levels. Using equation 2 and the observed level 1 intercoder reliability, r_1 of 92.3% to predict reliability for levels 2 and 3 yielded agreement scores of 85.2% and 78.6%. Again, the results predicted by the model are quite close to those actually obtained and, further, are within the 95% confidence intervals for the observed results (see Table 2).

The results in Table 2 are based on all of the meta-analytic variables as is commonly done in computing intercoder reliability. However, not all of these

data are contingent through all three levels. Some variables such as publication date are noncontingent, level 1 variables. Some variables such as "hospital(s)" are contingent for only two levels (i.e., number of hospitals followed by specific "type"). As noted above, only the outcome-based variables are contingent for all three levels. The intercoder agreement scores for these latter, strictly contingent variables (described above) were 92.3%, 88.8%, and 77.1%, respectively, for levels 1, 2, and 3. The predicted reliabilities for levels 2 and 3, using the observed level 1 result and equation 2, are 85.2% and 78.6%, respectively. These predicted results are nearly identical to those predicted using the entire database as well as the observed results for the strictly contingent variables.

Finally, a random sample of 12 studies previously coded by a physician untrained in research methods, but with extensive experience in data coding, yielded level 2 and 3 intercoder agreement of 83.8% and 72.5%, respectively, based on agreement scores with another highly trained coder. (It was not possible to compute a valid level 1 code because the final coding scheme containing these variables was not developed until after this preliminary coding.) Again, these reliabilities are close to both the observed and predicted results reported above.

RECONCEPTUALIZING RELIABILITY IN META-ANALYTIC RESEARCH

These two studies provide support for several conclusions: (1) the mean reliability estimates reported in meta-analysis can mask low intercoder reliability for particular, but important, classes of variables; (2) the intercoder reliability of variables most critical in calculating an effect size may be inaccurate; (3) a hierarchical model reflecting the contingent nature of the variables to be coded allows one to predict the magnitude of the reliabilities within levels of the hierarchy; (4) the hierarchical model holds irrespective of whether variables are strictly hierarchical.

PRECEDENTS FOR THE HIERARCHICAL STRUCTURE APPROACH

We have argued that reliability should be reported for the important variables in a meta-analysis using a hierarchical structure in which the reliabilities of clusters of variables within levels are independent. We do not claim that the hierarchical model is the only model that explains reliability results in meta-analytic reviews. However, it does pass one important crite-

rion for assessing theories, namely, it is parsimonious. The model only requires one additional construct—namely, the hierarchical structure of the variables to be coded. Moreover, it has face validity in that it is consistent with both theories of how humans organize information in memory (Mandler 1967) and their supporting empirical evidence (Wortman and Greenberg 1971).

The hierarchical model is also applicable to a wide range of circumstances in which meta-analyses have been conducted. In fact, the model passes a second important criterion for assessing theories—that of the "generality" of the phenomena that are consistent with it (Dooley 1990, p. 73). For example, in their seminal study of coder reliability in meta-analysis, Stock et al. (1982) found that two variables, "number of subsamples" and "quality of study," initially had unacceptably low intercoder agreement (i.e., less than 80%). Despite changes in the coding form and additional training, the mean intercoder agreement was only 66% for the nine "dimensions" of study quality. Most of these dimensions (e.g., "appropriateness of statistical analysis," "reasonableness of conclusions," etc.) can be viewed as level 4 variables, because they are judgments of the quality of level 3 variables. Such global assessments are inherently unreliable because they go one step beyond that necessary to code meta-analytic data, thereby adding a decision contingent on level 3 coding. Interestingly, if one uses the 82% result in level 2 to predict the reliability for level 4 (note that $r_4 = r'_2{}^2$ from equation 2), then one obtains 67% (.82 × .82), which is nearly identical to the 66% value for the level 4 "quality of study" cluster of variables.

Chalmers and his associates (1987) have noted the "replicate variability" in the results of a number of meta-analyses conducted by independent investigators. They located 57 different meta-analyses of 20 different interventions. Only half of these 20 sets contained replicate meta-analyses that agreed statistically with one another (i.e., in direction and significance level). Another five sets were in close agreement (i.e., in direction but slightly different significance level). Thus only 75% of the replicated meta-analyses were in general agreement (i.e., 15 out of 20). This finding is consistent with the level 3 results for the two studies reported above.

In another recently published study involving meta-analysis, Matt (1989) reported the "coder agreement" among two highly trained coders in extracting effect sizes from a sample of psychotherapy studies. Based on equation 4 in this article and using the larger number of coded "observations" (from the more experienced, senior coder) as the denominator (i.e., 172 coded effect sizes), it was possible to calculate that the intercoder agreement (i.e., based on 126 effect sizes that both coders agreed on) was 73% (Matt 1989, 110; also, see Orwin and Cordray 1985, 138). Matt's focus was on various decision

rules for selecting the appropriate effect sizes to be included in the meta-analysis—clearly a level 3 variable because effect size calculations require the extraction of means and standard deviations of dependent variables. In this instance, the intercoder agreement was nearly identical to that predicted by the hierarchical model in the illustration provided for highly reliable coders ($r_1 = .90$) and comparable to those reported above for level 3 in studies 1 and 2. In fact, an "average overall [agreement] rate of .90" for these coders was reported using level 1 variables (see Orwin and Cordray 1985, 138-139).

Although nonhierarchical models or modifications of the hierarchical approach may be necessary in some circumstances, the data and the examples presented above suggest that the hierarchical model is quite robust. Given the trend in meta-analytic reviews to stratify outcomes into successively more homogeneous subsets of persons, interventions, and study conditions so that generalizability can be better assessed (Light 1983), researchers will necessarily confront coding decisions in which the relationship between variables is hierarchical.

LIMITATIONS

The hierarchical model used to predict reliability at different levels is based solely on occurrence agreements. When both coders did not provide a score for a given variable (a nonoccurrence agreement), no agreement was scored. Nonoccurrence agreement provides an alternative operational definition of Orwin and Cordray's (1985) "deficient reporting quality" at the microlevel (i.e., "completeness"). Calculations of reliability in which both occurrence and nonoccurrence agreement were used within each level did not fit the predictions of our model.

In addition, the approach to reliability recommended in this article addresses only one of the four stages in which questions of agreement can occur within a meta-analysis, namely the accurate coding of variables in already identified studies. As Cooper (1982) has discussed, unreliability can also occur within other stages of meta-analytic reviews, including the identification of articles and the analysis and interpretation of results. Despite these possible sources of disagreement, recent findings suggest that the degree of consistency between meta-analysis of the same topic is "encouraging" (Chalmers et al. 1987). In fact, Chalmers and his colleagues (1987) speculate that the main reason for different conclusions among replicated meta-analyses "most probably lies in the fact that authors chose different primary endpoints" (p. 739; i.e., different level 3 outcome variables).

RECOMMENDATIONS FOR IMPROVING INTERCODER RELIABILITY

High reliability is usually a prerequisite for publication and those actually reported in meta-analyses have generally been high. However, the degree to which level 1 variables are overrepresented in these high reliabilities is unknown, but certainly predictable given publication contingencies. One way of dealing with the low reliabilities found for level 3 variables is to increase the reliability of level 1 coding. Using the model described earlier, an intercoder reliability of .93 at level 1 will result in a contingent reliability score of .80 for level 3. This can be accomplished by extensively training coders in discriminations of the type reflected in level 1 variables.

Another way to increase the reliability in meta-analysis is to restrict analyses to level 1 and level 2 variables that are drawn from study-level data (i.e., are not a subgroup of the larger study). By limiting the number of levels introduced, the systematic increase of unreliability at lower levels will necessarily be reduced, because this decreases the number of steps on which decisions for specific measures are made. Moreover, by employing homogeneous categories of interventions and measures of their effects, it also avoids the "apples and oranges" problem (Glass, McGaw, and Smith 1981) in which conceptually different treatment and outcomes are combined. This practice would have the additional advantage of enhancing the external validity of the findings (Wortman 1983).

Finally, meta-analysts should report their own evaluation of the contingent nature of the variables coded and consider the hierarchical approach presented in this article. Minimally, they should indicate the number of variables in each level, the mean reliability within each level, and the frequency of judgments on which these mean reliabilities are based. With this additional information, the adequacy of average reliability within levels can be fairly judged. Without such information it will not be possible to assess the validity of meta-analytic results. As the meta-analytic literature now stands, we conclude that the reliability of study results are inaccurate and, hence, that the validity of conclusions is more tenuous than previously assumed.

REFERENCES

Bangert-Drowns, R. L. 1986. Review of developments in meta-analytic method. *Psychological Bulletin* 99:388-99.

Bullock, R. J., and D. J. Svyantek. 1985. Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology* 70:108-15.

Chalmers, T. C., J. Berrier, H. S. Sacks, H. Levin, D. Reitman, and R. Nagalingham. 1987. Meta-analysis of clinical trials as a scientific discipline. II. Replicate variability and comparison of studies that agree and disagree. *Statistics in Medicine* 6:733-44.

Cooper, H. M. 1982. Scientific guidelines for conducting integrative research reviews. *Review of Educational Research* 52:291-302.

Dooley, D. 1990. *Social research methods*. 2d ed. Englewood Cliffs, NJ: Prentice-Hall.

Dyken, M. L., and R. Pokras. 1984. The performance of endarterectomy for disease of the extracranial arteries of the head. *Stroke* 15:948-50.

Edwards, A. L. 1976. *An introduction to linear regression and correlation*. San Francisco: Freeman.

Ganong, L. H. 1987. Integrative reviews of nursing research. *Research in Nursing and Health* 10:1-11.

Glass, G. V., B. McGaw, and M. L. Smith. 1981. *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L. V., and I. Olkin. 1985. *Statistical methods for meta-analysis*. New York: Academic Press.

Holt, R. R. 1982. Occupational stress. In *Handbook of stress*, edited by L. Goldberg and S. Breznitz. New York: Free Press.

Horwitz, R. I., and E. C. Yu. 1984. Assessing the reliability of epidemiologic data obtained from medical records. *Journal of Chronic Disease* 37:825-31.

Hunter, J. E., and F. L. Schmidt. 1990. *Methods of meta-analysis*. Beverly Hills, CA: Sage.

Hunter, J. E., F. L. Schmidt, and G. B. Jackson. 1982. *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Judd, C. M., E. R. Smith, and L. H. Kidder 1991. *Research methods in social relations*. 6th ed. Fort Worth, TX: Holt, Rinehart & Winston.

Kaylor, J. A., D. W. King, and L. A. King. 1987. Psychological effects of military service in Vietnam: A meta-analysis. *Psychological Bulletin* 102:257-71.

Landman, J., and R. Dawes. 1982. Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist* 37:504-16.

Langenbrunner, J. 1990. *Quantitative synthesis methods: Scientific validity and utility for policy. A case study of carotid endarterectomy*. Unpublished doctoral dissertation, University of Michigan.

Light, R. J. (ed.) 1983. *Evaluation studies review annual*, Vol. 8. Beverly Hills, CA: Sage.

Mandler, G. 1967. Organization and memory. In *The psychology of learning and motivation: Advances in research and theory*, Vol. 1, edited by K. W. Spence and J. T. Spence, 327-72. New York: Academic Press.

Matt, G. E. 1989. Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin* 105:106-15.

Orwin, R. G., and D. S. Cordray. 1985. Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin* 97:134-47.

Sacks, H. S., J. Berrier, D. Reitman, V. A. Ancona-Berk, and T. C. Chalmers. 1987. Meta-analyses of randomized controlled trials. *New England Journal of Medicine*, 316:450-55.

Smith, M. L., and G. V. Glass. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist* 32:752-60.

Smith, M. L., G. V. Glass, and T. I. Miller. 1980. *Benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Stock, W. A., M. A. Okun, M. J. Haring, W. Miller, C. Kinney, and R. W. Ceurvorst. 1982. Rigor in data synthesis: A case study in reliability in meta-analysis. *Educational Researcher* 11:10-14, 20.

Susanto, Y. E., W. H. Yeaton, and P. M. Wortman 1990. *Multilevel-multifacet meta-analysis: Occupational stress in health care workers*. Unpublished manuscript.

Von Alven, W. H. 1964. *Reliability engineering*. Englewood Cliffs, NJ: Prentice Hall.

Warlow, C. 1984. Carotid endarterectomy: Does it work? *Stroke* 15:1068-76.

Wortman, P. M. 1983. Evaluation research: A methodological perspective. *Annual Review of Psychology* 34:223-60.

———. 1987. Meta-analysis. *New England Journal of Medicine* 317:575-76.

Wortman, P. M., and L. D. Greenberg. 1971. Coding, recoding, and decoding of hierarchical information in long-term memory. *Journal of Verbal Learning and Verbal Behavior* 10: 234-43.

Yeaton, W. H., and P. M. Wortman. 1984. Evaluation issues in medical research systhesis. *New Directions for Program Evaluation* 24:43-56.

William H. Yeaton teaches evaluation research methods at the University of Michigan—Ann Arbor. He has conducted extensive research on meta-analysis methods for assessing medical technology.

Paul M. Wortman is Professor of Psychology in the Department of Psychology at SUNY—Stony Brook. He is the author of "Judging Research Quality," which will be published in the forthcoming volume, The Handbook of Research Synthesis, *and he is interested in the evaluation of innovative therapeutic interventions.*