Psychological Methods
2009, Vol. 14, No. 2, 165–176

1082-989X/09/$12.00 DOI: 10.1037/a0015565

The Relative Benefits of Meta-Analysis Conducted With Individual Participant Data Versus Aggregated Data

Harris Cooper and Erika A. Patall
Duke University

The authors describe the relative benefits of conducting meta-analyses with (a) individual participant data (IPD) gathered from the constituent studies and (b) aggregated data (AD), or the group-level statistics (in particular, effect sizes) that appear in reports of a study's results. Given that both IPD and AD are equally available, meta-analysis of IPD is superior to meta-analysis of AD. IPD meta-analysis permits synthesists to perform subgroup analyses not conducted by the original collectors of the data, to check the data and analyses in the original studies, to add new information to the data sets, and to use different statistical methods. However, the cost of IPD meta-analysis and the lack of available IPD data sets suggest that the best strategy currently available is to use both approaches in a complementary fashion such that the first step in conducting an IPD meta-analysis would be to conduct an AD meta-analysis. Regardless of whether a meta-analysis is conducted with IPD or AD, synthesists must remain vigilant in how they interpret their results. They must avoid ecological fallacies, Simpson's paradox, and interpretation of synthesis-generated evidence as supporting causal inferences.

Keywords: meta-analysis, aggregated data, individual participant data, Simpson's paradox, ecological fallacy

> Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in the abbreviated reports of research in journals and other printed sources. . . . Meta-analysis needs to be replaced by archives of raw data that permit the construction of complex data landscapes that depict the relationships among independent, dependent and mediating variables. (Glass, 2000)

These are strong words from the person who coined the term *meta-analysis* and is responsible for much of its early dissemination throughout the social sciences. Glass coined the term in 1976 to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Whereas Glass gave us the term, Olkin (1990) pointed out that ways to estimate effect sizes and the use of effect sizes to aid in policy decisions have existed since the turn of the century. For example, Pearson (1904) took the average of correlations from five separate samples to help decide whether a new inoculation for enteric fever should be used to reduce mortality (see Shadish & Haddock, 2009).

The use of meta-analysis in the social sciences, as well as in other fields, however, was rare before the 1970s. Late in that decade, several applications of meta-analytic techniques captured the imagination of behavioral scientists. Included among these were Smith and Glass's (1977) meta-analysis of psychotherapy research, which transformed results from evaluations of therapy effects into standardized mean differences; Schmidt and Hunter's (1977) combined analyses of validity coefficients to assess the validity of employment tests; Rosenthal and Rubin's (1978) combinations of probability tests to integrate research on interpersonal expectancy effects, and; Glass and Smith's (1978) synthesis of correlations between class size and achievement.

Two Forms of Meta-Analysis

Today, meta-analysis takes two forms as a method for combining quantitative results from numerous studies. First, meta-analyses can be conducted by relying on summary results of studies available in published or unpublished reports. That is, aggregated data (AD) meta-analysis is the statistical synthesis of the data from separate but similar studies conducted with the summary statistics presented in reports. In a meta-analysis of AD, a synthesist (a) system-

Harris Cooper and Erika A. Patall, Department of Psychology and Neuroscience, Duke University.

Thanks are extended to Jesse Berlin, Mike Clark, Ingram Olkin, and Alex J. Sutton for comments on the manuscript.

Correspondence concerning this article should be addressed to Harris Cooper, P.O. Box 90739, Department of Psychology and Neuroscience, Duke University, Durham, NC 27708-0739. E-mail: cooperh@duke.edu

atically attempts to collect all published and unpublished reports addressing a topic, (b) extracts effect sizes based on summary statistics presented in reports, and (c) statistically combines effects, in order to obtain an estimate of the average effect size and the associated confidence interval (Cooper, 2009). In addition, many synthesists examine sample and study features that might influence study outcomes.

Alternatively, meta-analyses can be conducted by obtaining and cumulating individual participant-level data (IPD).[1] Unlike meta-analyses based on AD, IPD meta-analysis involves the central collection, checking, and re-analysis of the raw data from each study in order to obtain combined results. Like meta-analysts of AD, IPD meta-analysts seek to collect every study addressing a topic, both published and unpublished. However, rather than the reports of aggregate data analysis, IPD are obtained. If the outcomes across studies are measured identically (an issue discussed below), meta-analysts can pool and re-analyze the raw data using traditional inferential statistics or more sophisticated techniques.[2]

We believe that Glass's (2000) call for replacing meta-analysis of AD with the integration across studies of IPD should be taken less as an indictment of meta-analysis of AD than as a comparative judgment contrasting the limitations of cumulating group statistics relative to cumulating the primary data on which those statistics are based. Clearly, there is no contest here. Having the IPD allows the synthesist to replicate the results of any meta-analysis of AD, as well as to improve it in numerous ways. Such methods and opportunities are outlined below.

Glass (2000) also recognized that his vision for the future of data synthesis was based on the archiving of primary research data sets in ways that are highly accessible to future users. He noted that the development of the Internet demolished most technological barriers inhibiting data sharing. But two problems remain. First, data storage on the Internet is underway and is simple to accomplish for new data collections, but what is to become of older data sets? Some databases collected in the past could certainly be uploaded retroactively, but others have been lost or destroyed, though their group statistics live on in available reports.

Additionally, uploading even the available data sets might require resources and incentives unavailable to the primary researchers. Second, whereas the technological barriers to data sharing are gone, the human barriers remain. Not the least of these is researchers' willingness to share data. Further, if research data sets were freely available on the Internet, would this violate ethical principals of privacy and informed consent? That is, even when data have been rendered "anonymous," just a small amount of demographic and background information on a participant may be enough to identify individuals. Below, we briefly discuss these issues more fully.

In sum, then, we might describe the science of research synthesis as being in a second stage of transition. The first transition is from the narrative research review—in which opaque rules of cognitive algebra are used to synthesize the results of studies—to meta-analysis of AD. The second stage involves the transition from meta-analysis of AD to the accumulation of IPD. Just as we can still find today examples of narrative research review (some appropriate, some not; see Cooper, 2003), we can expect to continue seeing the meta-analysis of group statistics in the future. Each transition involves not a complete overturning of the old way but a gradual shift in emphases. The speed with which the new techniques are adopted will be a function of (a) how quickly the remaining barriers to cumulating IPD can be lowered and (b) what incentives are available for using them.

We attempt to answer several questions: What are the relative benefits of meta-analysis of IPD and meta-analysis of AD? What are the intersections and points of divergence between the two synthesis techniques? Can the two approaches be used in a complementary fashion? What should an applied research synthesist be expected to do? In the following sections, we attempt to address these questions. First, we discuss the relative benefits of conducting an IPD meta-analysis and how IPD meta-analysis may be less susceptible to errors in interpretation. Then, we discuss the relative benefits of conducting an AD meta-analysis, especially the ability to include more studies and its lower cost. Next, we discuss issues needing consideration regardless of which meta-analytic approach is taken. These issues include the model of error assumed to underlie the generation of study outcomes and limitations on the type of evidence generated by research synthesis. Finally, in the context of the current state of the field, we make recommendations for what may be the best practice for applied researchers.

---

[1] Various labels have been used to refer to what we call individual participant data (IPD) meta-analysis. Such terms used include *mega-analysis* or *integrative data analysis.* We choose to use *IPD meta-analysis* to remain consist with the literature published on meta-analysis within the social sciences and medical sciences.

[2] Issues related to those of IPD and AD meta-analyses are also presented in the literature on the synthesis of single-subject designs. Single-subject designs utilize a series of observations for a single person. These studies have been synthesized by extracting raw participant data from reports (often graphs) and combining them through traditional parametric methods, meta-analytic methods, or more sophisticated techniques, such as hierarchical linear modeling, though what the most appropriate synthesis method is for this design remains unresolved. This article limits discussion to the synthesis of group designs, as the issues relevant to single-subject designs are different and deserve separate discussion (see Shadish & Rindskopf, 2007, for a review of the methods and issues in doing quantitative synthesis of single-subject designs).

The Benefits of IPD Meta-Analysis

Whereas meta-analyses of IPD rarely have been conducted in the social sciences, much work on the topic has been done in the medical sciences. Consequently, social scientists can look to the medical literature to find the most thoughtful analytic work on the relative benefits of cumulating IPD. In the medical literature, IPD stands for "individual patient data" and AD stands for "aggregated data," or group-level statistics. We suggest that the simple substitution of "participant" for "patient" be used by social scientists. In this way, we highlight the parallels between the issues involved in cumulating social and medical science data as well as facilitate cross-disciplinary communication by standardizing language.

Comparisons of IPD and AD Meta-Analysis in Medical Research

The earliest comparison of IPD meta-analysis and AD meta-analysis in the medical literature appears to be that reported by Stewart and Parmar (1993). This study focused on the effectiveness of two drug treatments for ovarian cancer. The researchers compared (a) the results of an AD meta-analysis conducted with effect sizes from 8 published clinical trials and (b) the results of an IPD meta-analysis of 11 studies: the 8 published studies plus 3 additional data sets, 2 of which were unpublished. Thus, this was not a comparison between the two techniques that held constant the data sets analyzed by each. Moreover, social scientists would be surprised to find circumstances under which a search for evidence revealed available sets of IPD that were not accompanied by reports of group statistics. This is an issue we return to below. Regardless, the authors concluded that the two approaches to research synthesis yielded different estimates of the efficacy of the treatment and judged the results of the IPD meta-analysis to be less biased, that is, less favorable to the treatment.

In 1995, Jeng, Scott, and Burmeister conducted a comparison of AD meta-analysis and IPD meta-analysis. However, similar to Stewart and Parmar (1993), these authors varied not only the meta-analytic technique but also the literature included in the AD and IPD meta-analyses. Studying paternal cell immunization effects on recurrent miscarriage, they added individual patient data missing from the AD meta-analysis data sets and included four additional sets of patient-level data. Again, the authors found that the effect size estimate in the AD meta-analysis was larger than the estimate in the IPD meta-analysis.

Perhaps the seminal event in the growth of IPD meta-analysis in medicine occurred in 1994, when the Cochrane Collaboration convened a workshop with nearly 40 researchers who had been involved in planning and conducting IPD meta-analyses. The group discussed and prepared an article on the methodological and practical aspects of carrying out IPD meta-analyses (Stewart & Clarke, 1995). This group identified several benefits of IPD meta-analyses. First, IPD meta-analyses can be used to perform within-study subgroup analyses that might not have been conducted by the initial data collectors. For example, the IPD meta-analysis might include an analysis examining whether the sex of participants was associated with the effect of a treatment, even though the original data analysis did not include such an analysis (we present some cautions about such moderator analyses below). Second, the researchers conducting the IPD meta-analysis can check the data in the original studies, perhaps finding errors, looking for statistical outliers, and so forth. Third, the researchers conducting the IPD meta-analysis can ensure that the original analyses were conducted properly and can conduct standardized analyses across all data sets. For example, the synthesist may wish to perform identical missing or ambiguous data analyses on each data set or perform a standard stratification. Further, researchers can perform more complex analyses that might test questions not possible with AD. For example, time-to-event or survival analyses can be conducted with IPD meta-analysis in a straightforward manner but can only be done with AD meta-analysis by making a series of assumptions. Likewise, many studies measure a multitude of variables within a single design. Multivariate approaches such as multivariate analysis of variance (MANOVA), multiple regression, structural equation modeling (SEM), or multilevel modeling can be conducted in the context of IPD. However, with the exception of meta-regression, techniques parallel to such sophisticated multivariate analysis approaches are much more complex or often impossible in the context of AD meta-analysis. Nevertheless, it should be noted that a recent review of the methods used in IPD meta-analysis suggested that, in practice, a two-stage analysis method that mimics those of AD meta-analysis—a method in which the raw data in each study are first converted to a standardized effect size and then are combined—was the most common method of analyzing IPD, rather than more flexible one-stage models (Simmonds et al., 2005). Finally, those conducting the IPD meta-analysis can add new information to the data sets, perhaps by adding data about participants or study methodology, through follow-up queries to the participants or the researchers.

The primary goal of the Cochrane Collaboration workshop, however, was not to compare the two approaches but rather to provide guidance on conducting IPD meta-analyses and to assist groups of investigators in forming IPD meta-analysis collaborations. Interestingly, the workshop participants suggested that during the first stage of an IPD meta-analysis, the initiators of the project should conduct an AD meta-analysis as part of the background research in order to provide a rationale for researchers to share their

data and assess whether it would be worthwhile to allocate resources to an IPD meta-analysis on a particular topic.

In contrast to earlier investigations, in which dissimilar data sets were used for the IPD and AD meta-analyses, fairer tests of the difference in effect size estimates and their precision were soon to follow. Steinberg et al. (1997) studied a literature examining the relation between use of oral contraceptives and the risk of ovarian cancer. These authors found that when the pool of studies used in the two types of analyses and the measures extracted from each were essentially identical, the results of the AD meta-analysis "were strikingly similar" (Steinberg, 1997, p. 919) to those of the IPD meta-analysis.

Olkin and Sampson (1998) asked the question, "If one uses summary data and individual patient data from the same set of studies, under what models will both approaches yield identical results?" (p. 317). For the sake of simplicity, these authors examined a fixed-effect model, using data points that were independent within and across studies, were identically distributed, and had common variance and homogeneous treatment differences across studies. The authors' concern was with the estimate of the overall effects and not with tests of moderators of treatment effects. Olkin and Sampson showed mathematically that the results of a two-way analysis of variance (Treatment $\times$ Study, with no interaction term) and an AD meta-analysis of the same data yielded the same estimates for the treatment main effect in both the balanced and the unbalanced cases, that is, regardless of whether there was a common or proportionate number of observations for each treatment–study combination. Mathew and Nordström (1999) then relaxed some of Olkin and Sampson's assumptions and demonstrated that the estimate of the treatment main effect also was the same for IPD meta-analysis and AD meta-analysis even when observations were not independent within studies and the observations had different covariation matrices across studies.

The situation is different for the study of interaction effects, or in meta-analysis parlance, for tests of moderators of overall effect sizes. Lambert, Sutton, Abrams, and Jones (2002) used computer simulations to examine the differences between AD and IPD meta-analysis. They created 27 meta-analytic scenarios. First, they varied the overall effect size in studies by defining the outcome of interest as an event that had the exact same odds of happening (20%) under a standard and a new treatment for low-risk patients. For high-risk patients they set the standard treatment event rate at 40% and the new treatment event rate at 20%, 30%, and 35%. This resulted in measures of effect expressed as odds ratios of 0.38, 0.64, and 0.81 for three different sets of studies. The number of studies (5, 10, and 20) and the number of patients in each study (200, 500, and 1,000, with the size of each study within a meta-analysis assumed to be equal) was also varied in the simulated meta-analyses. Next, the authors ran 1,000 repetitions of each of the 27 scenarios, with each patient randomly assigned a probability of .3, .4, .5, .6, or .7 of being in the high-risk group for each study and the probabilities repeated 1, 2, or 4 times, depending on whether the meta-analysis included 5, 10, or 20 studies. This created variation in the proportion of high-risk patients in each study. Then, the authors compared the results of 27,000 fixed-effect meta-regressions (AD meta-analysis) with the results of unconditional logistic regressions using the same data and including predictors for the study, treatment, risk group, and treatment by risk group interaction.

Lambert et al. (2002) found that "although not biased, there is far greater variation for the meta-regression estimates than for the IPD analysis estimates" (p. 91). The implication of this finding is that AD meta-analysis had far less power than IPD meta-analysis to reject the null hypothesis of no treatment difference. Only under circumstances involving the largest effect sizes (odds ration [OR] = 0.81) with the largest samples (n = 1,000) and the most studies (k = 20) did the AD meta-analysis likelihood of rejecting the null hypothesis of no influence of the moderator (a hypothesis known to be false) exceed 85%. The authors suggested that although their results were tested under restricted parameters, there was no reason to believe that varying these (e.g., random effects, unequal sample sizes of studies within a meta-analysis) would yield different results. Thus, although estimates produced by both meta-analytic techniques were similar, a difference in power was found in this case, in which the data available for the IPD meta-analysis and AD meta-analysis came from an identical set of studies. However, the power of meta-regression, in comparison with IPD methods, to detect interactions has since been found to depend on the heterogeneity in the covariate distributions across studies (Schmid, Stark, Berlin, Landais, & Lau, 2004; Simmonds & Higgins, 2007). Whereas meta-regression can be a powerful method when the covariate distributions are very different across studies, interaction questions may be better addressed with IPD methods when the effect of a covariate is apparent primarily within studies.

The Ecological Fallacy and Simpson's Paradox

A lack of statistical power is not the only concern when moderators of overall effects are examined with AD meta-analysis (and, as we see in our review, with IPD meta-analysis as well). In particular, the magnitude and even the direction of relationships between variables found by comparing group-level statistics may not be indicative of the same relationships tested with IPD. This concern manifests itself in two ways. First, there is the phenomenon often referred to as the "ecological fallacy," introduced to the discipline of sociology by Robinson in 1950. Robinson demonstrated the ecological fallacy using data on the relationship between race and literacy. He began by calculating the correlation between the percentage of African American

residents and the percentage of illiterate residents in the nine geographic divisions of the 1930 U.S. Census. The correlation was $r = .95$. Next, Robinson calculated the same correlation using data for nearly 100,000 individuals, about 9,000 of whom were African American. The correlation was $r = .20$.[3]

Even more dramatically, Robinson (1950) then showed how the sign of the correlation could reverse depending on the ecological unit. He found that the correlation between whether an individual was born outside of the United States and whether he or she was illiterate (in the 1930 U.S. Census) was $r = .118$. Individuals who were not born in the United States were more likely to be illiterate. However, Robinson also found that the correlation between the percentage of foreign-born individuals in a geographic region was negatively correlated with the region's illiteracy rate ($r = -.62$). Berlin, Santanna, Schmid, Szczech, and Feldman (2002) discussed the ecological fallacy in the context of medical research.

Related to the ecological fallacy is another phenomenon called "Simpson's paradox" (Simpson, 1951). Here, the result of an aggregate analysis of IPD that ignores subgroup memberships is compared with results within meaningful subgroups, and the results are found to differ. Perhaps the best known example of Simpson's paradox involves a lawsuit brought against the University of California at Berkeley alleging bias against women in the selection of graduate school applicants (Bickel, Hammel, & O'Connell, 1975). The aggregate analysis of IPD revealed that 44% of male applicants and 35% of female applicants were admitted to graduate school. However, when Bickel et al. examined the data by the 85 individual departments (or admitting units), the data showed few units departing significantly from expected frequencies of admissions and about the same number of units admitting higher percentages of women than men. The aggregate appearance of bias was caused by women tending to apply to units with low rates of admission.

In the case of the ecological fallacy, the researcher is comparing findings (effects, relationships) based on IPD that ignore group membership with findings based on data aggregated to the group level. In Simpson's paradox, IPD findings ignoring group membership are compared with findings within multiple meaningful groups (and ignoring any group-level differences). Taken together, the ecological fallacy and Simpson's paradox illustrate that there is no a priori reason to expect that the results of group-level analyses and IPD analyses should be similar in either magnitude or direction. Different causal processes may be at work influencing (a) meaningful groupings of individuals and (b) individuals within those meaningful groups. The issue of which analysis is "correct," therefore, depends on the level at which the researcher wishes to make the process description.

Let us summarize these concerns with a hypothetical example. Suppose a researcher is interested in the relationship between teachers' expectations for student performance and how frequently students are called on in their class to speak before the group. The researchers could find that an analysis conducted with IPD that ignores classroom entirely reveals no relationship between the two variables. However, if teachers' average expectations for students in their class are correlated with the total frequency of public speaking in the class, the researchers could find, with the same data, that higher average expectations for a class are associated with less frequent opportunity for students to speak to the group as a whole. The causal process explaining this negative correlation might be that a higher expectation for the group as a whole leads the teacher to use more small-group, independent learning opportunities. This illustrates the ecological fallacy.

Likewise, if the researchers also examined the relationship *within* each class separately, they might find that in all or most classes, teachers more frequently called on those students for whom they held higher expectations. Here, the underlying causal mechanism might be that teachers felt group instruction would proceed more smoothly if answers given in public were more likely to be correct. Within classes, then, the relationship between expectations for individual students and their opportunities to speak to the group could be positive. This illustrates Simpson's paradox. In sum, then, these researchers could *correctly* conclude that "teachers with higher expectations for their class also provide less time for public speaking." Likewise, the researchers could *correctly* conclude that "within classes, students for whom teachers hold higher expectations are also more likely to speak in public." We see, then, that problems in drawing inferences from data can occur when (a) IPD is analyzed ignoring meaningful grouping of data and (b) when data representing between-group or within-group analyses is erroneously used to make inferences at the other level.

In meta-analysis, the study is almost always a meaningful grouping. Studies differ on characteristics such as the populations from which participants were drawn and the context in which the studies were conducted (e.g., different universities, clinics, schools) to name just two of many potentially important variations. Rarely, if ever, do all of the studies in a meta-analysis randomly sample participants from the same population and have no variation added to the data because of the studies' methodologies. Thus, research-

[3] It should be noted that an alternative reason for why the individual-level correlation is small compared with the regional-level correlation in Robinson's example is because base rates of African Americans compared with Caucasian Americans are grossly unequal.

ers conducting AD meta-analyses can make inferential errors if they conclude that study-level variables found to be associated with effect sizes are also descriptive of relationships at the level of individual participants. Similarly, researchers conducting IPD meta-analyses can make inferential errors if they do not include, as a categorical variable in their analyses, the study from which the IPD was obtained and examine whether interactions involving the study variable are important. (A recent review of current practices for IPD meta-analyses in health care suggested that most IPD meta-analysts do not make this mistake in that almost all reviewed syntheses accounted for the study in the analyses; Simmonds et al., 2005). An advantage of properly analyzed IPD meta-analyses, then, is the ability to test for moderators of effects sizes at both levels of analysis.[4]

Benefits of AD Meta-Analysis

Two benefits of AD meta-analysis that appear repeatedly in the medical literature are that this type of research synthesis can be done quickly and inexpensively relative to IPD meta-analysis. The Cochrane Collaboration working group (Stewart & Clark, 1995) wrote the following:

> It is perhaps not generally appreciated just how much time and effort is involved in performing an IPD meta-analysis. It is not something to be undertaken lightly, and since clinical, scientific, statistical, computing, and data management skills are required, it is generally not something to be undertaken by a single individual. Of necessity, projects usually take a few years from initiation to first publication. (p. 2060)

IPD meta-analyses require much more staff for data collection, entry, and cleaning than do AD meta-analyses. In fact, working with a set of a dozen studies, Steinberg et al. (1997) estimated that the cost of the IPD meta-analysis ($259,000) was about five times that of the AD meta-analysis.[5] Further, we can expect that the difference in resources needed to conduct an IPD or an AD meta-analysis in the same topic area will increase as the amount of research associated with the topic increases. Locating each data set, contacting its authors, obtaining the data, and making its coding and data matrix commensurate with other data sets will be far more labor-intensive and time-consuming than will obtaining and coding the associated research report.[6]

A benefit of IPD meta-analysis over AD meta-analysis often cited in comparisons carried out in medicine is the capability of IPD meta-analyses to include unpublished data sets (e.g., Jeng et al, 1995; Stewart & Parmar, 1993). However, in the social sciences, it is much rarer than in medicine to have unpublished data sets for which reports of group statistics are not also available. Rather, social scientists are more likely to find reports, both published and unpublished, of group statistics for which IPD is unavailable. For example, Wicherts, Borsboom, Kats, and Molenaar (2006) attempted to obtain the 249 data sets used in 141 empirical articles published in journals of the American Psychological Association. Six months of effort resulted in the acquisition of data from 64 studies, or about 26% of the possible data sets. Further these were data sets collected relatively recently, meaning that they were likely still intact and not lost in time. Thus, experience suggests that in most topic areas, social scientists conducting parallel AD and IPD meta-analyses might uncover a few unpublished data sets but are more likely to uncover reports of group-level statistics for which no IPD is available.

Thus, we should list among the benefits of AD meta-analysis the likelihood that group-level statistics, especially effect sizes, may be available from studies for inclusion in AD meta-analysis when the participant-level data from these same studies are unavailable for inclusion in IPD meta-analysis. This difference in evidence available for IPD meta-analysis and AD meta-analysis raises two complex issues. The first concerns what forms of biases might be manifest in data sets available for analysis at the individual participant level, relative to those biases manifest in data available for AD meta-analysis and relative to all data sets on a topic (if they could be obtained). We know of a long list of potential availability biases when the results of published data are compared with those of unpublished data. Most notably, results of published data are generally more likely to reject the null hypothesis than unpublished data testing the same notion (Greenwald, 1975) and are more likely to exhibit larger effect sizes (Lipsey & Wilson, 1993). Considering how available sets of IPD might differ from all data sets on a topic is more speculative. This conjecture becomes even more complex if we assume that as the barriers for data sharing come down, the differences be-

[4] It is possible for synthesists performing AD meta-analyses to conduct such multilevel analyses, if the required data are available. This becomes feasible when the synthesist has available the individual effect sizes of interest (e.g., the correlation between the sex of individual participants and the frequency of each individual's speaking in public) and each study's cumulative value on the same variables at the group level (e.g., each study's percentage of male and female participants and total frequency of public speaking). (See Goldstein, Yang, Omar, Turner, and Thompson, 2000, and Turner, Omar, Yang, Goldstein, and Thompson, 2000, for an example of multilevel analyses with IPD and AD.)

[5] I. Olkin (personal communication, December 12, 2007) reported that the synthesists did not finish when the money ran out and that they continued to work on the project. Thus, he suggests that it may be more accurate to use the figure of IPD meta-analyses costing about eight times the cost for AD meta-analyses.

[6] Speed is a good thing, of course, as long as the press for rapid results does not interfere with the researchers' engaging in the thoughtful analysis of their data and contemplation of their findings.

tween available and unavailable data sets will diminish and eventually may disappear.

Looking at the state of things today, therefore, we can plausibly assert that available data sets are likely to contain data that were more recently collected than all data that have been collected on a hypothesis. This may not be a bad thing if we believe that the effect of interest has changed over time, or that research methods have improved, and we are interested in how the phenomenon of interest operates today. Less assuredly, we might suggest that the availability of data could be positively related to the quality of the data. We see below that data sets may be more readily available from funded than from unfunded research, so if funding is related to competency (an assumption that might be more or less valid for different topic areas), then available data sets might lead to more trustworthy inferences. This seems like a good thing, assuming that in the area of interest, it is more difficult to do good research inexpensively. Nevertheless, in the case of the AD meta-analysis that has a more complete set of relevant studies than that of a parallel IPD meta-analysis, a meta-analysis should be conducted on the entire set of AD, and moderator analyses should be conducted to assess whether temporal trends in study results exists. Again, this will help assess the trustworthiness of an IPD meta-analysis for which only the most recent data is available.

A second implication of the likelihood that social scientists may have available more—often many more—group-level findings for use in an AD meta-analysis than they will have IPD for use in a meta-analysis brings us back to the issue of statistical power to detect effects. We saw above that when the same studies are available for IPD and AD meta-analysis, the power of IPD meta-analysis exceeded that of AD meta-analysis (Lambert et al., 2002). However, at least for overall effects, at some point, if many more studies are available for inclusion in an AD meta-analysis because the related IPD is not available, the relative power of the two techniques to detect a known effect will reverse. In sum, then, the likelihood that group-level findings will be available for more studies than will IPD suggests that AD meta-analyses may exhibit fewer forms of availability bias and potentially have greater power to test hypotheses.[7]

Another advantage of AD meta-analysis is that it is easy to combine studies even when measurement scales used in studies are not commensurate in analyses because an initial step in an AD meta-analysis is to standardize the measures of effect found in separate studies into a common metric (Borenstein, 2009). However, the process of combining studies with different forms of measurement is often a multiple-stage process in an IPD meta-analysis. So, for example, a researcher interested in the impact of a clinical intervention on depression will likely uncover studies that use several different scales to measures depression. Placing these raw scores into a single multiple regression, for example, would produce results wholly uninterpretable. When outcomes have been measured identically across all studies, it is feasible to conduct meta-analytic analyses on IPD with a one-stage process in which all IPD for all studies are combined to perform a single analysis.[8] However, to date, it has been challenging to include different measures in different units of an outcome in the same IPD meta-analysis with a one-stage process. This problem for IPD meta-analysis may be overcome if a two-stage process is used, in which IPD analyses within each study (or each subgroup within a study) are conducted first and then summary effect size statistics for each study can be computed and AD methods used. Further, Bauer and Hussong (2009) has proposed an analytic technique to circumvent this problem.

It is worth noting that incommensurate scaling is less problematic in the medical literature because the outcome of interest in a large proportion of medical meta-analyses conducted with IPD is a dichotomous event rate, for example, mortality. Thus, although researchers might express effects using different metrics (e.g., rate ratios, odds ratios), it is easy to convert these to a similar metric with the IPD.[9] Whereas such measures do appear in social science data (e.g., graduation rates, recidivism rates), they often represent only a portion of the measures taken in a study and meant to tap the same underlying construct (e.g., measuring the construct of achievement by gathering data on students'

[7] Reading the medical literature suggests an advantage of IPD meta-analysis over AD meta-analysis in its capability to include unpublished data sets and thus reduce publication bias. However, in social science, it is standard practice for preparers of AD meta-analyses to conduct thorough searches for grey literature (cf. Rothstein & Hopewell, 2009). Because AD meta-analyses including only published studies are rare in social science, the suggested benefit of including unpublished data in IPD meta-analyses likely does not obtain; if a social science researcher performing an AD meta-analysis uncovers unpublished IPD, they would calculate commensurate group statistics so that the study could be included. As in the medical field, the results of the IPD meta-analysis would then be tested for robustness against reporting bias (Rothstein, Sutton, & Borenstein, 2005). Thus, we suggest that the capability to include unpublished evidence does not differentiate the IPD and group-level statistics approaches to meta-analysis.

[8] In practice, IPD syntheses have sometimes used a "mega-trial" analysis, in which all IPD is combined for a single analysis and distinctions between studies are ignored and data are analyzed as if they came from a single trial. However, it is more appropriate to use a stratified analysis where trial identities are included in the model, such as in multilevel modeling. Random effects multi-level models synthesizing IPD in a one-stage process have been developed but appear to be used less frequently in practice than two-stage process analyses.

[9] For a discussion of IPD methods used to deal with ordinal and continuous outcomes within the medical literature, see Whitehead et al. (2001) and Higgins, Whitehead, Turner, Omar, and Thompson (2001).

graduation, cumulative grade point averages, and scores on standardized achievement test). Still, even among medical IPD syntheses, in which incommensurate scaling is less problematic, two-stage analyses are by far the most common approach (Simmonds et al., 2005).

Intersections and Points of Divergence Between the Two Techniques

Table 1 provides a summary of the relative merits of IPD meta-analysis and AD meta-analysis. We have seen that each has some clear benefits (e.g., data cleaning and supplementation for IPD meta-analysis, cost and speed for AD meta-analysis). Further, each procedure has some benefits depending on characteristics of the particular literature. For example, AD meta-analysis may have greater relative power to detect effects and greater subsequent resistance to publication bias if IPD is unavailable for studies for which AD is available. However, Table 1 illustrates that when both IPD and AD are available for all included studies, the benefits of IPD meta-analysis outweigh those of AD meta-analysis. It is important as well to briefly discuss two additional issues that synthesists need to consider regardless of whether they are conducting an IPD meta-analysis or an AD meta-analysis.

Fixed Versus Random Effects

One important aspect of averaging effect sizes and estimating their dispersion involves the decision about whether a fixed-effect or random-effect model underlies the generation of study outcomes. In a fixed-effect model, each effect size's variance is assumed to reflect sampling error of participants only, that is, error solely due to participant differences. Hedges and Vevea (1998) stated that fixed-effect models are most appropriate when the goal of the synthesists is "to make inferences only about the effect size parameters in the set of studies that are observed (or a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects)" (p. 3).

However, studies also can be viewed as a random sample drawn from a population of studies. For example, as we noted above, studies that look at the impact of a treatment for depression might vary in numerous meaningful ways, including setting, treatment frequency, and the expertise of the treatment deliverer, to name a few. In addition to being potential moderators of the treatment effect, this variation may suggest that it is most appropriate to consider treatments represented in the synthesis as "randomly" sampled from all possible treatment manifestations. That is, in a random-effects analysis, study-level variance is assumed to be present as an additional source of random influence that must be added to variation because of the sampling of participants. Thus, when their assumptions are violated, fixed-effect models may seriously underestimate error variance and random-effects models may seriously overestimate error variance (Overton, 1998).

Regardless, researchers conducting most of the comparisons of IPD and AD meta-analyses we have observed in the medical literature have chosen to do the comparison using a fixed-effect model. Further, a review of the methods used in IPD meta-analysis suggests that random effects analyses are used infrequently in such meta-analyses (Simmonds et al., 2005). This should not be taken by potential synthesists to mean that the issues associated with a choice of model are different when IPD or group-level statistics are to be combined. They are not. In both instances, the synthesist must choose the error model most appropriate to their topic and the associated studies.

Table 1
Benefits of Meta-Analysis on Individual Participant Data and Aggregated Data Compared With One Another

Meta-analysis of individual participant data
Can perform subgroup analyses that were not conducted by the initial data collectors
Can check data in the original studies
Can ensure that the original analyses were conducted properly and standardize analyses across studies
Can perform more complex analyses more easily
Can add new information to the data sets
Can test variables that moderate effect sizes with greater power, assuming all individual participant data sets are available
Can test for both between-study and within-study moderators
Meta-analysis of aggregated data
Can be conducted for less cost, in both money and time
Can be carried out faster
Can include group-level statistics for which individual participant data are not available
May diminish bias if study results are associated with availability of individual participant data
May increase power to detect effects if many studies are available without individual participant data

Study-Generated Versus Synthesis-Generated Evidence

Second, regardless of whether IPD or AD meta-analysis is the technique of choice, it is important to bear in mind a limitation of the type of evidence generated by research synthesis (Cooper, 2009). Research syntheses can contain two different sources of evidence about the research problem or hypothesis. The first type is called *study-generated evidence*. Such evidence is present when a single study contains results that directly test the relation being considered. Research syntheses also contain evidence that does not come from individual studies but rather from the variations in procedures across studies. This type of evidence, called

synthesis-generated evidence, is present when the results of studies conducted with different procedures testing the same hypothesis are compared with one another.

There is a crucial difference between study-generated and synthesis-generated evidence; only study-generated evidence based on experimental research allows synthesists to make statements concerning causal relationships. For example, suppose that the research synthesists are interested in the relative effects of two treatments of depression. Suppose further that the literature search uncovered 10 studies that used both types of treatment approaches and randomly assigned participants to one treatment or the other. These studies provide study-generated evidence regarding the effect of treatment. Here, if outcomes revealed more positive effects for Treatment A than for Treatment B, a conclusion that the type of treatment caused the difference would be warranted.

Suppose instead that no study manipulated both treatments but five studies experimentally manipulated Treatment A and compared it with a no-treatment control, whereas five different studies compared Treatment B with no treatment. Let us say that outcomes revealed positive results across both types of treatment, but again the effect was more positive for Treatment A compared with Treatment B. As in the previous case, treatment overall, or both Treatment A and Treatment B separately, can be inferred to have a causal relationship with the outcome, as this evidence was directly generated by the individual studies. However, when the effect of treatment is compared between the two sets of studies, it can only be inferred that an *association* exists between type of treatment and outcome. It cannot be inferred that Treatment A *causes* larger treatment effects than Treatment B. Here, the relationship between the type of treatment and the outcome was generated by the synthesis rather than by the studies themselves.

When the effect sizes from groups of studies are compared within a research synthesis—regardless of whether they were derived from simple correlational analyses or controlled experiments with random assignment—the synthesists can only establish an association between a moderator variable—a characteristic of the studies—and the outcomes of studies. They cannot establish a causal connection. This is because it is the ability to use random assignment of participants to conditions within experiments that allows primary researchers to assume that third variables are represented equally in the different conditions. At the synthesis level, experiments are never randomly assigned to study characteristics. To continue this example, it might be the case that the set of studies testing Treatment A (versus control) also used different types of participants or different outcome measures than experiments testing Treatment B, to name just two additional ways the sets of studies might differ. Therefore, it is impossible to determine whether it was the variation in treatment, the participants, or the measures that "caused" the difference in effect sizes. When a synthesis finds an association between a characteristic of a study and the outcome, results should be used to guide further research that follows more controlled research designs meant to test causal hypotheses.

Thus, regardless of whether IPD meta-analysis or AD meta-analysis is used, when study characteristics are found to be associated with study outcomes, the synthesists should report the finding as an association, regardless of whether the included studies tested the causal effects of a manipulated variable. Put in the language of analysis, in AD meta-analysis, causal inferences can never be drawn from the moderator analyses linking features of studies to effect sizes (or other study features for that matter). In IPD meta-analysis, interactions involving variables measured at the study level cannot be inferred to test causal relationships. Again, to continue the example involving treatment of depression, if the IPD meta-analysis includes one set of studies testing behavioral treatments and another set testing cognitive treatments, a significant Treatment-Versus-Control $\times$ Treatment-Type interaction cannot be used to conclude that the treatment type causes variation in treatment effects, because treatment type was not randomly assigned to studies (and therefore might be confounded with other characteristics of studies). Thus, in both AD and IPD meta-analysis, causal inferences can be made concerning only those variables generated by experimental manipulations.[10]

Recommendations for Applied Researchers: How AD and IPD Meta-Analysis Can Be Used in a Complementary Fashion

We mentioned above that the Cochrane Collaboration working group (Stewart & Clarke, 1995) suggested that one of the first steps in conducting an IPD meta-analysis is to conduct an AD meta-analysis. Likewise, given that IPD will be available for fewer studies than will AD (Simmonds et al., 2005) and, consequently, results from such IPD meta-analyses will likely be biased if unavailability is related to study results (Stewart & Tierney, 2002), Riley, Simmonds, and Look (2007) suggested that it is important to supplement available IPD with AD for studies for which IPD is not available. We agree. In fact, in the social sciences, conducting an AD meta-analysis

[10] Finally, and briefly, the above point about when causal inferences are legitimate, as based on both IPD and AD meta-analysis, raises the more general issues of correlated predictors. Regardless of whether IPD meta-analysis or AD meta-analysis is used, research synthesists will find that characteristics of studies are confounded. This is a feature of the data, not the analysis method, and will require the synthesists to include cautions regarding multicollinearity when interpreting their findings.

based on data and statistics from the reports of studies for which IPD is and is not available should be more than a first step; it should be a parallel activity that is used to help interpret the results of an IPD meta-analysis. Currently, the number of social science studies for which synthesists will have group-level statistics likely will exceed, sometimes far exceed, that for which IPD is available.[11] This difference in available data cannot be ignored. Thus, it seems prudent for the synthesists to conduct an AD meta-analysis that includes whether or not IPD was available for the study as a study-level moderating variable. In this way, the synthesists could shed light on whether the availability of IPD was systematically related to the outcomes of studies. This AD meta-analysis should also examine whether study characteristics other than the outcome (e.g., participant characteristics, location, sample size, funding) also were associated with IPD availability. In doing so, the synthesists garner some information about whether the sample of studies with IPD differs systematically from all available studies (which, of course, may still be a selective sample of all studies). If the results of such an AD meta-analysis suggest little or no difference, then synthesists conducting an IPD meta-analysis can proceed with greater confidence in the representativeness of the findings (but in light of possible diminished power). If the AD meta-analysis suggests that systematic differences exist between studies with and without IPD available, then an added caution should be contained in the interpretation of the IPD meta-analysis. Alternatively, state-of-the-art synthesis methods recently have been developed to efficiently combine data when IPD is available for only a portion of included studies (Sutton, Kendrick, & Coupland, 2008). This model provides a promising alternative to choosing between conducting either an IPD or an AD meta-analysis.

In addition, with regard to the issue of combining results that use different outcome measures, synthesists might find it valuable to conduct IPD meta-analyses and AD meta-analyses on separate measures. For example, if synthesists are conducting a meta-analysis on evidence regarding an educational intervention, the synthesists might conduct separate IPD meta-analyses on graduation rates, cumulative grade point averages, and standardized test scores. The standardized effect sizes generated from the IPD could then be compared with those from an AD meta-analysis.

Finally, given the substantial resources that need to be invested when conducting an IPD meta-analysis, an important question is when it is worthwhile to invest in obtaining IPD. Future research should continue to explore methods and benchmarks that can be used to assess when IPD meta-analysis will be most beneficial.

The Future of AD Meta-Analysis and IPD Meta-Analysis: Issues in Data Sharing

As we mentioned in our introduction, when data sharing becomes the norm in the social sciences, many of the benefits of AD meta-analysis will disappear. Of course, there are many barriers to data sharing (see Sieber, 1991), and many still need to be lowered. The issues of authorship and cost in conducting IPD meta-analyses stand out, and these are likely to persist for the foreseeable future, suggesting that AD meta-analysis and the hybrid model outlined above are far from extinct. Still, the incentives for data sharing are increasing while the barriers are coming down. Advances in data storage and ease of data transfer are barriers that largely have been removed. A recent incentive is the development and heightened enforcement of policies requiring or encouraging sharing of data collected with federal funding (National Institutes of Health, 2003). Further, even if research data sets could be made freely available to researchers who would like to conduct IPD meta-analyses, it remains an open question whether an individual's agreement to participate in the original study also implies consent to have data included in a secondary analysis. Still, even this issue may be addressed simply by making data sets available to researchers only under the same rules of confidentiality that applied when the data was first collected.

Of course, the principal incentive for data sharing involves improving the precision and utility of social science, both for the advancement of understanding human behavior and for improving the human condition. Looking at the current state of research synthesis relative to three decades ago, when Glass (1976) coined the term *meta-analysis,* or even less than a decade ago (Glass, 2000), when he called for its retirement, suggests that the imperative is inexorable.

[11] The question "How many studies are needed before conducting a meta-analysis is warranted?" has vexed researchers since the term was coined. Because the answer to this question will vary as a function of the topic area and especially of the heterogeneity of study methods, the only specific answer to the general question that has gained any supporters is "2." Valentine, Pigott, and Rothstein (in press) discuss the issue in the broader context of power analysis.

References

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101–125.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regression for the investigation of treatment effect

modifiers: Ecological bias rears its ugly head. *Statistics in Medicine, 21,* 371–387.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187,* 398–404.

Borenstein, M. (2009). Effect sizes for studies with continuous outcome data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 221–236). New York: Russell Sage Foundation.

Cooper, H. (2003). Editorial. *Psychological Bulletin, 129,* 3–9.

Cooper, H. M. (2009). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher, 5,* 3–8.

Glass, G. V. (2000, March). *The future of meta-analysis.* Paper presented at the University of California, Berkeley–Stanford University Colloquium on Meta-Analysis, Department of Psychology, University of California, Berkeley.

Glass, G. V., & Smith, M. L. (1978). *Meta-analysis of research on the relationship of class size and achievement.* San Francisco: Far West Laboratory for Educational Research and Development.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics, 49,* 399–412.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20,* 2219–2241.

Jeng, G. T., Scott, J. R., & Burmeister, L. F. (1995). A comparison of meta-analytic results using literature vs. individual patient data: Paternal cell immunization for recurrent miscarriage. *Journal of the American Medical Association, 274,* 830–836.

Lambert, P. C., Sutton, A. J., Abrams. K. R. & Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology, 55,* 86–94.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Mathew, T., & Nordström, K. (1999). On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics, 55,* 1221–1223.

National Institutes of Health. (2003). *Final NIH statement on sharing research data.* Rockville, MD: Author. Retrieved December 10, 2007 from http://grants.nih.gov/grants/guide/notice-files/not-od-03–032.html

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3–10). New York: Russell Sage Foundation.

Olkin, I., & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics, 54,* 317–322.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3,* 1243–1246.

Riley, R. D. Simmonds, M. C., & Look, M. P. (2007). Evidence synthesis combining individual patient data and aggregate data: A systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology, 60,* 431–439.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3,* 377–386.

Rothstein, H. R., & Hopewell, S. (2009). The grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 103–125). New York: Russell Sage Foundation.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* Chichester, United Kingdom: Wiley.

Schmid, C. H., Stark, P. C., Berlin, J. A., Landais, P., & Lau, J. (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology, 57,* 683–697.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 257–278). New York: Russell Sage Foundation.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95–109.

Sieber, J. E. (Ed.). (1991). *Sharing social science data: Advantages and challenges.* Newbury Park, CA: Sage.

Simmonds, M. C., & Higgins, J. P. T. (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine, 26,* 2982–2999.

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials, 2,* 209–217.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13,* 238–241.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Steinberg, K. K., Smith, S. J., Stroup, D. F., Olkin, I., Lee, N. C., Williamson, G. D., & Thacker, S. B. (1997). Comparison of effect estimates from a meta-analysis of summary data from published studies and from meta-analysis using individual patient data for ovarian cancer studies. *American Journal of Epidemiology, 145,* 917–925.

Stewart, L. A., & Clarke, M. J. (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine, 14,* 2057–2079.

Stewart, L. A., & Parmar, M. K. B. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *Lancet, 341,* 418–422.

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluations of Health Professions, 25,* 76–97.

Sutton, A. J., Kendrick, D. & Coupland, C. A. C. (2008). Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine, 27,* 651–669.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine, 19,* 3417–3432.

Valentine, J. C., Pigott, T. & Rothstein, H. R. (in press). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics.*

Whitehead, A., Omar, R. Z., Higgins, J. P. T., Savaluny, E., Turner, R. M., & Thompson, S. G. (2001). Meta-analysis of ordinal outcomes using individual patient data. *Statistics in Medicine, 20,* 2243–2260.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726–728.

Received June 26, 2008
Revision received January 22, 2009
Accepted January 29, 2009 ■