Psychological Bulletin
1999, Vol. 125, No. 6, 669-676

0033-2909/99/$3.00

Understanding the Effects of Extrinsic Rewards on Intrinsic Motivation—Uses and Abuses of Meta-Analysis: Comment on Deci, Koestner, and Ryan (1999)

Mark R. Lepper, Jennifer Henderlong, and Isabelle Gingras
Stanford University

Recently, 3 different meta-analytic reviews of the literature concerning the effects of extrinsic rewards on intrinsic motivation have appeared, including that by Deci, Koestner, and Ryan (1999) in this issue. Interestingly, despite their common focus, these reviews have offered dramatically opposed bottom-line conclusions about the meaning and implications of this literature. In this comment, the authors examine differences among these 3 reviews and conclude that the findings of this literature have been more accurately captured by the reviews of Deci et al. and Tang and Hall (1995) than by that of Cameron and Pierce (1994). More broadly, the authors also suggest that there may be significant short- and long-term costs to the unthinking or automatic use of meta-analysis with theoretically derived, procedurally diverse, and empirically complex literatures like that concerning extrinsic rewards and intrinsic motivation.

> Although research syntheses can facilitate the development of understanding in a research area by channeling subsequent research to resolve the uncertainties that emerge, the impact of syntheses has not been uniformly beneficial. They have sometimes distorted understanding of a phenomenon and discouraged further research. (Eagly & Wood, 1994, p. 487)

How should research literatures in psychology be reviewed? For more than a century, psychologists have sought to summarize the findings of studies in different research areas. In the process, although with varying levels of formality and precision, such reviewers have traditionally sought to accomplish four basic goals: (a) to describe the conclusions to be drawn from a given literature, (b) to identify the conditions under which particular effects do and do not occur, (c) to examine the extent to which the accumulated findings in a research area do or do not provide support for various theoretical positions, and (d) to suggest new directions for further research. In short, they have sought to address the age-old questions what, when, why, and whither.

Within the context of such traditional reviews, statistical techniques have long been available for providing more quantitative summaries of experiments that have been repeated, or replicated with small variations, on a number of occasions—methods for combining significance levels across experiments, for example, and for calculating indices of statistical effect size (see, e.g., Cochran, 1937; Mosteller & Bush, 1954; Pearson, 1904; Tippett, 1931). The use of such techniques can have many clear advantages. Certainly, continuous measures of effect size can convey more information than any arbitrary dichotomous identification of effects as merely statistically significant or nonsignificant. Meta-analytic procedures may also help researchers to recognize real but perhaps small effects that emerge consistently, though often nonsignificantly, across a series of experiments. Further, because well-controlled unpublished studies, such as doctoral dissertations, are often included in meta-analyses, such summaries may help researchers to evaluate more effectively the importance of possible "file drawer problems"—illusions of real effects produced by the selective publication of a small number of statistically significant findings, drawn from a much larger pool of failures-to-replicate.

Within the past 15 years, however, discussions of the potential advantages of such statistical procedures have been increasingly generalized to the larger claim that meta-analytic reviews represent an inherently better method of summarizing and evaluating the results of research on a given problem. Although many specific arguments have been offered, the two most fundamental and most general claims are that meta-analytic reviews are superior because they are (a) inherently more precise and (b) intrinsically more objective than traditional nonquantitative or "narrative" reviews (Cooper & Hedges, 1994; Hunt, 1997; Matt & Cook, 1994; Rosenthal, 1991; Wolf, 1986). As a result, their use has proliferated.

In the literature concerning the effects of extrinsic rewards on intrinsic motivation, in particular, three major meta-analyses have appeared within the last 5 years. Strikingly, despite assertions of procedural rigor and objectivity, these three meta-analytic reviews have produced bottom-line conclusions sharply at variance with one another. On the one hand, Cameron and her associates sought to examine the evidence for a general main effect of extrinsic rewards and concluded that "Our overall findings suggest that there is no detrimental effect [of extrinsic rewards] on intrinsic motivation" (Cameron & Pierce, 1994, p. 394), later characterizing the proposition that extrinsic rewards may undermine intrinsic

Mark R. Lepper, Jennifer Henderlong, and Isabelle Gingras, Department of Psychology, Stanford University.

Preparation of this article was supported, in part, by Research Grant MH-44321 from the National Institute of Mental Health and fellowships from the U.S. National Science Foundation, and the Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (Fonds FCAR-Quebec).

Correspondence concerning this article should be addressed to Mark R. Lepper, Department of Psychology, Jordan Hall, Building 420, Stanford University, Stanford, California 94305-2130. Electronic mail may be sent to lepper@psych.stanford.edu.

motivation as a "myth" (Eisenberger & Cameron, 1996).[1] On the other hand, Tang and Hall (1995) examined this same literature in terms of the theoretical models that generated it and summarized the results of their review as showing that "the overjustification effect has been consistently demonstrated in situations when it should be expected to occur" (p. 379). Similarly, in the most recent meta-analysis—which responded directly to Cameron and her colleagues' negative claims about this literature—Deci, Koestner, and Ryan (1999) concluded that "in general, tangible rewards had a significant negative effect on intrinsic motivation for interesting tasks" (p. 653). In the present article, we examine some of the differences in procedures and conclusions among these three meta-analyses.

More generally and more importantly, however, we also seek to use the controversy produced by the appearance of these three recent, and seemingly contradictory, meta-analyses of the literature on the effects of extrinsic rewards on intrinsic motivation as an illustrative case study to argue against any unthinking or ideologically generalized preference for meta-analysis over other reviewing techniques. There are, of course, some conditions under which meta-analysis may indeed be the obvious reviewing technique of choice. For example, in the medical field, meta-analytic techniques may be especially appropriate when real effects might be obscured by necessarily small sample sizes, when research questions can be sharply defined, and when the treatments, dosages, and outcome measures are comparable across studies. However, there are other conditions—particularly in literatures like the present case that feature complex findings obtained using diverse procedures designed to address questions of primarily theoretical, rather than practical, interest—under which the use of meta-analysis may prove misleading. Under such circumstances, as many recent critics have noted (see, e.g., Hennessey & Amabile, 1998; Kohn, 1996; Lepper, 1998; Lepper, Keavney, & Drake, 1996; Ryan & Deci, 1996; Sansone & Harackiewicz, 1998), the use of meta-analysis may produce only illusions of greater precision and objectivity.

The Social Psychology of Psychological Research

To understand our concerns about both the three specific recent meta-analyses of interest and the application of meta-analytic procedures to this literature in general requires a recognition of, indeed a focus upon, the social and historical context in which research is undertaken. As many authors have long noted, science is an intensely social affair. Thus, the interest value, the publishability, and the ultimate significance of any study are determined not simply by the inherent characteristics of the study itself but by the meaning of that study within some particular historical, social, and theoretical context (see, e.g., Kaplan, 1964; Kuhn, 1962; Mahoney, 1976; Merton, 1973; Miller, 1999).

The Social/Historical Context of Intrinsic Motivation Research

In the particular case of the literature on extrinsic rewards and intrinsic motivation, the historical and social context of this field is relatively clear. Prior to 1970, expectations about the effects of extrinsic rewards were guided primarily by behavioristic principles—to the extent that rewards served as reinforcers, they were expected to enhance performance and increase motivation. On the basis of these assumptions, behavior modification procedures, including token economies and other heavy-handed classroom reward programs, were widely used. Thus, it was in stark contrast to this dominant paradigm when, in the early 1970s, researchers in three different laboratories arrived independently at similar hypotheses about the potentially detrimental effects of extrinsic rewards and constraints on subsequent intrinsic interest. Within a short period of time, all three groups had published studies demonstrating comparable adverse effects of superfluous extrinsic rewards (Deci, 1971, 1972; Kruglanski, Friedman, & Zeevi, 1971; Lepper, Greene, & Nisbett, 1973).

This convergence of findings across these different experiments seemed especially impressive and informative because each of these studies had used very different rewards, activities, contingencies, and subject populations. Deci (1971, 1972), for example, offered Carnegie Mellon undergraduates money for each spatial puzzle they solved correctly and observed later engagement with these puzzles during alleged breaks in the experiment. Kruglanski and associates (1971) offered Israeli high school students a visit to the psychology laboratories at Tel Aviv University in return for their completing measures assessing their creativity, task recall, reported enjoyment, and willingness to participate in future studies. Lepper, Greene, and Nisbett (1973) told preschool children that they could win a fancy "Good Player Award" for drawing pictures with a new type of magic markers and then observed these children's interest in this activity in their regular classrooms several weeks later.

At the same time, even these earliest studies included clear demonstrations of conditions under which extrinsic rewards failed to undermine, or even enhanced, intrinsic interest. Thus, Deci showed both that tangible rewards that were not contingent upon engagement in or success at the task did not decrease later intrinsic interest (1972) and that contingent verbal rewards could increase subsequent intrinsic interest (1971). Likewise, Lepper and colleagues (1973) showed that tangible rewards that were unexpected did not have the same detrimental effects on subsequent interest as did expected rewards. In addition, shortly afterwards, Calder and Staw (1975) published the first of a number of studies demonstrating that undermining effects occurred only with activities designed or selected to be of initial intrinsic interest to participants, whereas enhancement effects were more likely to occur with activities of little or no initial intrinsic interest.

In short, the results of the first half-dozen studies in this area, by themselves, presaged most of the important distinctions made by later narrative and meta-analytic literature reviews—that detrimental effects were less likely (and positive effects more likely) when rewards were clearly noncontingent, unexpected, and verbal, or when tasks were of little inherent interest to participants. Moreover, many of these early authors went to great lengths to make explicit the point that extrinsic rewards could, under appropriate

[1] Like Deci et al. (1999), we refer to the two papers by Cameron and her colleagues (Cameron & Pierce, 1994; Eisenberger & Cameron, 1996) as a single analysis. Both papers used very similar procedures and clearly arrived at very similar conclusions; unfortunately, the Eisenberger and Cameron report does not provide readers with sufficient methodological detail to identify the differences between the two.

conditions, have either positive or negative effects on later motivation and to argue that the relevant issue for further research was not whether rewards have negative, or positive, effects "in general," but rather when and why these different effects might occur (see, e.g., Calder & Staw, 1975; Deci, 1971; Lepper et al., 1973).

In historical context, therefore, the explicit goal of this research literature was precisely to provide a more nuanced perspective on behavioristic principles and assumptions. Thus, it should not be surprising that so many studies in this area involved complex, theoretically driven hypotheses and empirically critical moderator variables. In turn, we believe, these characteristics make this literature particularly poorly suited for some uses of meta-analytic techniques.

The sampling problem. Consider, for example, the most fundamental premise underlying meta-analysis—namely, that one can consider the collection of studies on a particular topic to constitute a reasonable sample from which one can generalize to some larger population of situations in the world. Such a proposal may work wonderfully in many situations, such as evaluations of new, clearly specified drug treatments or surgical procedures. Its applicability to the present literature, however, seems extremely dubious.

In fact, it makes almost no sense to consider the particular set of studies that have been performed and published in this area to be a random or representative sample of any meaningful larger population of reward procedures in the world. Because of the theoretical nature of this literature and the fact that even the very first studies on this problem had already established that there were conditions under which rewards could have both positive and negative effects, once the first half-dozen studies had been published, virtually every subsequent study to be published in reputable journals required the addition of some set of control procedures or cross-cutting moderator variables that would contribute to psychologists' understanding of the phenomenon. Thus, subsequent demonstrations of a significant detrimental effect of an extrinsic reward were almost necessarily accompanied by one or more conditions deliberately selected or designed so that no such effect, or even an opposite effect, would be obtained.

In addition, to produce the most theoretically telling comparisons, many important experiments employed exotic procedures or deceptive methods without counterparts in real life outside the laboratory. For instance, Pittman, Cooper, and Smith (1977; see also Fazio, 1981) sought to examine the importance of participants' attributions for their actions by directly manipulating those attributions by means of a false-feedback procedure. These participants were told that they were being hooked up to a physiological recording device and were then given false feedback about their pattern of arousal—either that it looked similar to others who had previously found the activity especially intrinsically motivating or that it looked similar to others who had previously found the extrinsic reward particularly attractive. Clearly, this procedure is unlikely to be replicated by any parent, teacher, or employer in real life. Moreover, unrealistic procedures of this sort, used to test particular theoretical alternatives while controlling for other factors, are widespread in this literature (e.g., Fazio, 1981; Kruglanski et al., 1975; Kruglanski, Alon, & Lewis, 1972; Lepper, Sagotsky, Dafoe, & Greene, 1982; M. Ross, Karniol, & Rothstein, 1976).[2]

Attempting to draw general or overall inferences from a sample composed of many of these sorts of studies is akin to recruiting participants for a decision-making experiment on the basis of their unusual but theoretically significant characteristics—some, perhaps, with a rare psychiatric disorder, others with abnormally high (or low) IQ scores, and still others with some unique speech impediment—then combining them, perhaps along with some "normal" individuals, into a single group and generalizing to how "people in general" make decisions. Although such an extrapolation would clearly be unwarranted, studies of these different types of individuals might nevertheless have provided critical information about how brain organization affects decision-making.

The singularities problem. Of course, one possible solution to such difficulties is to include as moderator variables in any meta-analysis those procedural variations that distinguished, for example, experimental from control conditions in the various studies under review. Unfortunately, this is often not possible in practice, whenever there are not enough cases of a particular variation to permit the use of meta-analysis. This, in turn, is most likely to be true of theoretically generated procedures that may simultaneously carry considerable information about underlying processes yet involve minimal mundane realism.

To take but one obvious example, consider an early study by Kruglanski, Alon, and Lewis (1972), in which the authors sought to demonstrate the theoretical importance of the participants' views of their own actions as instrumental to some extrinsic reward—while controlling for the participants' actual experiences and perceptions as they engaged in the activity itself—in producing detrimental effects on later intrinsic interest. To do this, Kruglanski and colleagues created a procedure in which participants were asked to undertake an interesting activity without any mention or expectation of a reward but were then told (falsely) after the completion of the activity that they were now receiving the reward that they had been "promised" at the outset. Not surprisingly, unlike the standard unexpected-reward conditions used in other studies, this ingenious and unusual unexpected-reward procedure did produce a decrease in later intrinsic interest in the activity, just as if the reward had indeed been expected.

Note, of course, that the same artificial procedure that makes this study of special theoretical interest—the use of a deliberate deception concerning the conditions under which participants had agreed to engage in the activity—also makes this condition totally unrepresentative of the normal uses of rewards in everyday life to which researchers would like to generalize the results of this literature. Note also that the unrepresentative and unrealistic procedure of this study also makes it one that is less likely to be replicated in subsequent studies, unless there were some specific controversy concerning its results, and hence, it is less likely to be represented in a sufficient number of studies to warrant quantitative treatment within a meta-analysis.

The problem of singularities can also apply to studies that do involve mundane realism. For example, M. Ross (1975, Exp. 2) conducted a study on the effects of reward salience in which preschool children were instructed to think about a reward as they engaged in an intrinsically interesting task (ideation condition),

[2] Although this set of studies may not necessarily be representative of the array of reward procedures used in the world outside the laboratory, many individual studies (as well as the findings of this literature on the whole) can have clear practical, as well as theoretical, significance.

were instructed to think about something else as they engaged in the task (distraction condition), or were not asked specifically to think about anything as they engaged in the task (nonideation condition). Subsequent levels of intrinsic motivation were then compared with those of a no-reward control group. Ross predicted and found that the undermining effect did not occur in either the distraction condition or the control group, and he used these results to suggest that extrinsic rewards are harmful only to the extent that they are attentionally salient to children. Although the ideation and distraction manipulations have both theoretical significance and mundane realism, such manipulations have rarely been employed in subsequent research, thus making their inclusion in meta-analytic comparisons problematic.

Hence, the significance of studies such as these is usually ignored by meta-analyses, in the sense that the unique, though theoretically telling, conditions and variations that they were designed to investigate are not even mentioned as elements that had been included in the designs of these experiments. Instead, such studies are occasionally excluded altogether or are, more typically, clumped together with other quite different studies—just as the Kruglanski et al. (1972) study is included as just another case of "unexpected reward" in all three recent meta-analyses of this literature, and the three theoretically distinct experimental conditions in the M. Ross (1975) study are collapsed into one group in the meta-analyses conducted by Cameron and Pierce (1994) and Deci et al. (1999). Clearly, such practices obscure the true meaning of these findings.

The interactions problem. Furthermore, in the case of some of the recent meta-analyses reported in this area, these problems are exponentially exacerbated by some authors' willingness to collapse results across diametrically opposed effects—even when (a) a sufficient number of studies are available to permit the testing of moderating variables, and (b) relevant statistical tests have established the heterogeneity of effect sizes within a given set. Cameron and Pierce (1994), for example, routinely collapsed both across explicitly designed experimental versus control conditions and across conditions shown to have produced significant opposing effects (Kohn, 1996; Lepper, 1998; Lepper, Keavney, & Drake, 1996; Ryan & Deci, 1996).

Some proponents of meta-analysis seem to justify such procedures by their "interest" in overall main-effect or bottom-line conclusions from a literature, as opposed to the more qualified (or nuanced) conclusions that might come from an interest in interaction effects. Although such reasoning may apply well to contexts in which sharp questions about agreed-upon treatments and outcomes are at issue (e.g., whether the use of a particular fertilizer increases the size of a given crop), its applicability to a literature in which investigators have known for over 20 years that different effects may occur under different conditions and in which over 80% of published studies were explicitly designed to include conditions in which different effects were predicted to occur seems highly inappropriate (see, e.g., Abelson, 1995; Kaplan, 1964). Indeed, even proponents of meta-analysis have noted that focusing solely on main effects can be grossly misleading when there are significant interactions present (Hunter & Schmidt, 1990; Matt & Cook, 1994; Press, 1990; Wachter & Straf, 1990).

In fact, if similar procedures were used in other comparable contexts, we suspect that they might well be considered to involve outright deception, if not fraud. For example, if a drug company were to report to the FDA that "overall" or "in general" there were no detrimental side effects to some new treatment—knowing full well that the prior evidence had actually shown that this treatment consistently produced completely opposite, and statistically significant, effects for different groups or under different conditions (e.g., increasing the frequency of strokes for men but reducing it for women, or vice versa)—we suspect that the government would simply not accept the argument that the company had been more "interested" in main effects than in interactions. Nor, we suspect, would our hypothetical drug company be without legal recourse and moral standing if the government were to publish an evaluation of the effectiveness of their new product based on a meta-analysis that focused only on main effects reported in studies that had been explicitly designed to elucidate interactions. That is, one can imagine a series of studies that included theoretically derived control conditions under which the effects of the drug had been predicted to disappear (e.g., a control group given an antagonistic drug or a blocking agent) in which it would be obviously inappropriate to sum across treatment and control conditions.

In fact, we believe that there is no way that any traditional reviewer would have been allowed to employ such procedures in summarizing studies in a narrative review. That the systematic use of such methods can be used to yield an anomalous "no difference" conclusion within a meta-analysis (Cameron & Pierce, 1994) tells us essentially nothing about the phenomenon or the actual literature under review. As the old saying goes, if you put one hand in the oven and the other in the freezer, it may not be terribly informative to report that your hands are "overall" at room temperature.

Summary. In short, we have argued that an unthinking use of meta-analytic techniques typically provides a less accurate and less complete picture of a research literature when that literature displays the following properties: (a) many theoretically significant but otherwise unusual conditions (the sampling problem), (b) many studies with theoretically critical procedural variations that are not represented in sufficient numbers to be analyzed separately (the singularities problem), and (c) many studies with moderator variables producing significant interactions that either cannot be, or have not been, considered within the meta-analytic framework (the interactions problem).

Because all three of these conditions are highly characteristic of the extant literature on the effects of extrinsic rewards on intrinsic motivation, we believe that any traditional meta-analysis of this literature must have limitations. At the same time, of the several recent meta-analyses of this literature that stimulated this comment, it is clear that those by Deci et al. (1999) and Tang and Hall (1995) are less subject to these criticisms than that of Cameron and Pierce (1994). In particular, both Deci et al. and Tang and Hall conducted their meta-analyses taking into account the theories that have guided the development of this literature over the past 30 years. These authors also recognized the importance of interactions by treating separately most experimental conditions designed to have opposing effects rather than collapsing across these conditions. Hence, it is not surprising that it is only the analysis by

Cameron and Pierce that produced the anomalous conclusion that negative effects of extrinsic rewards are merely a myth.[3]

Psychological or Functional Versus Purely Statistical Effect Sizes

In addition to the foregoing problems that may result from the application of meta-analytic techniques to complex theoretically derived literatures, other issues can arise when purely statistical effect-size measures are used to compare studies that vary greatly in their procedures and measures. In certain literatures in particular, measures of purely statistical effect size may not be good indicators of functional or psychological effect size, and in fact these two measures can be negatively correlated (Lepper, 1995). Again, greater attention to the social psychology of psychological research may help in understanding these additional problems with the unthinking use of meta-analytic techniques.

Consider two different meanings of effect size (Lepper, 1995; Lepper et al., 1996; Prentice & Miller, 1992). On the one hand, purely statistical measures of effect size depend solely on the means, standard deviations, and sample sizes in the relevant comparison conditions. For such measures, any procedures that minimize within-condition variance and/or maximize mean differences between conditions produce larger effect-size estimates.

By contrast, when evaluating the predictive power, or functional significance, of an experimental effect as psychologists, rather than as statisticians, researchers typically use a very different metric. In this more psychological sense of the term, the functional "effect size" and the consequent inferential power of a given comparison depend not only on the statistical significance of the difference between two conditions and the size of the samples but also on several additional factors: (a) the strength of the manipulation producing that difference, (b) the sensitivity and consequentiality of the dependent measures, and (c) the range and power of factors that were not explicitly controlled and thereby constitute the natural error variance against which the effect must "compete" for significance (see, e.g., Abelson, 1995; Lepper, 1995; Prentice & Miller, 1992; L. Ross & Nisbett, 1991). Thus, a deeper consideration of the specific contexts in which given effects are obtained may be needed for understanding the true impact of different variables.

Take a simple example: Imagine a set of studies all testing the same hypothesis that exposure to televised violence increases children's subsequent aggression but varying dramatically in their specific procedures and dependent measures. In some studies, the manipulation involves a brief exposure to a single film clip containing a modest level of violence; in others, the independent variable involves lengthy and/or repeated exposures to highly violent programming. Similarly, in some studies, the dependent measure involves a single self-report measure of the children's stated willingness to retaliate against hypothetical others who had hit them; in others, the dependent measure involves actual aggressive behaviors covertly observed in some naturalistic setting, like the school playground, several weeks after exposure to the manipulation.

Now suppose further that all of these studies used samples of an identical size and obtained identical statistically significant effects. Nonetheless, we argue, the functional significance, or psychological effect size, of these different studies could vary greatly. Thus, the less extensive, prolonged, and extreme the manipulation of televised violence—given equivalent samples and statistical effect sizes—the more powerful the underlying process would be inferred to be. Likewise, the more dissociated, distant, delayed, and consequential the measure of later aggression, and even the more uncontrolled "noise" present at the time of measurement—again, given equivalent samples and statistical effect sizes—the more powerful that particular demonstration would be taken to be.[4]

The important point, however, is not just that statistical effect size may sometimes be a poor indicator of functional effect size but that these two effect sizes are often negatively correlated. Both the use of more subtle and highly controlled independent variables and the use of more distant, dissociated, delayed, and consequential dependent measures, other things being equal, involve putting one's hypothesis to a more difficult and more stringent test and hence, reduce the likelihood of obtaining statistically powerful effects, even if the underlying hypothesis is valid. Ironically, the use of methodologically superior and more psychologically significant procedures must, ceteris paribus, lead to smaller average statistical effect sizes. Moreover, the use of statistical effect size estimates, in such cases, gives greatest weight to precisely those findings that would otherwise be found of least evidential value.[5]

In short, in procedurally diverse literatures like that concerning the effects of extrinsic rewards on intrinsic motivation, where measures of statistical and functional effect size are weakly or even negatively correlated, a simple adage may apply: One (effect) size does not fit all.

---

[3] In fact, with the exception of the age effects reported by both Deci et al. (1999) and Tang and Hall (1995), the conclusions reached in these two meta-analyses closely parallel those outlined in recent narrative reviews (Lepper & Henderlong, in press; Sansone & Harackiewicz, in press).

[4] Perhaps the most extreme, and most telling, illustration of this general point can be seen in pure demonstration studies like those of Asch (1951), Milgram (1974), and Rosenhan (1973). In each of these cases, the results from a single condition—35% conformity in the face of a unanimous majority in Asch's line-judging paradigm, 65% complete obedience to an authority figure's request for participants to shock an innocent fellow subject in Milgram's studies, or 0% detection of sane pseudopatients by real-world mental-hospital personnel in Rosenhan's observations—have remained among the most powerful and widely cited phenomena in psychology, despite the absence of any comparison group or statistical analysis. Indeed, the inferential power of these sorts of findings derives entirely from the functional or psychological effect sizes of these results, which, in turn, depend heavily on the precise procedures used. Note, thus, how little interest we would have if Milgram had shown that 100% of his subjects obeyed the experimenter's order that they use a black pen rather than a blue one or if Asch had shown that over 90% of his participants conformed to a group consensus that William McKinley was a more (or a less) effective president than James Garfield.

[5] Once again, a potential answer to this problem would be the inclusion of various study characteristics, like the type of manipulation or dependent measure, as an explicit factor in a meta-analysis (see Tang & Hall, 1995, for some examples). Once again, however, such a solution may be difficult in practice when, as in the present literature, the numbers of effect sizes of each type are small and many different study characteristics happen to be confounded with each other in the sample of studies that have been conducted.

The Hidden Costs of Quantitative Research Reviews

Our final set of concerns about the potential adverse effects of an overreliance on meta-analytic procedures is perhaps the most general—namely, that any automatic application of a technique that assigns precisely equal weights to studies of intrinsically different values may, in the long run, differentially "reward" research designs and procedures that give the most statistical bang for the buck, even when those designs and procedures may be less inherently valuable or informative. Certainly, in the extreme case, a rational researcher who was convinced that only a single overall statistical effect-size measure from his or her study would be included or considered in later research reviews might well decide that it would be a waste of time to include features such as subtle theoretical variables, multiple experimental or control conditions, real-world contexts, or distant and dissociated dependent measures.

Before going further, however, let us be perfectly clear: We do not believe that meta-analytic techniques are inherently inferior to traditional techniques; indeed, in literatures involving standard procedures and sharply posed questions, we believe that meta-analytic techniques should be generally superior to traditional reviewing methods. At the same time, in our view, the mindless or automatic application of such procedures to complex and theoretically derived literatures such as the present case can have negative effects on the level of discourse about the phenomena under study (Eagly & Wood, 1994). Let us examine these issues, then, in the specific context of the literature on extrinsic incentives and intrinsic motivation.

In the first place, it seems that meta-analysis can often encourage a preoccupation with the evaluation of simplistic overall or bottom-line conclusions—even in the face of literatures in which the existence of meaningful main effects of any sort has already been ruled out by decades of prior research. In most areas of research where findings prove to be complex, it has historically been seen as a sign of progress when the level of discourse shifts from a simplistic general question of whether some particular effect exists to more nuanced questions about when and why and how such effects occur (Abelson, 1995; Kaplan, 1964; Sansone & Harackiewicz, in press). We find it an unfortunate by-product of the present uses of meta-analysis that it seems to have encouraged some users to revert to drawing simplistic overall conclusions (Cameron & Pierce, 1994; Eisenberger & Cameron, 1996) even after 30 years of research demonstrating that extrinsic rewards can have both positive and negative effects on subsequent intrinsic motivation, depending upon the specific conditions of the study.

One specific, and particularly pernicious, contributing factor to this focus on simplistic questions is the tendency for many meta-analysts to average effects across competing cross-over interactions and across experimental and control procedures within experiments. Again, this seems a level of analysis that simply has not been considered acceptable in scholarly reviews until the issue was broached in the context of meta-analysis. For example, we believe that no reviewer using traditional methods would have been allowed—as Cameron and Pierce (1994) were—to take the array of studies showing that extrinsic rewards have opposite effects as a function of the interest level of the task (see Deci et al., 1999, Table 8) and to present them as if they had never even included a manipulation of task interest and had instead simply shown that there was no overall difference as a function of the presence of rewards.

Similarly, it seems that the use of meta-analysis can sometimes favor a reliance on across-study comparisons, in preference to inherently better controlled and more informative within-study comparisons. Thus, we find it remarkable in the current reviews that comparisons of average effect sizes for different types of rewards (expected, unexpected, contingent, noncontingent, etc.) are reported, relative to a no-reward condition, across all studies, without any mention of the host of more direct and probative within-experiment comparisons of, for instance, expected versus unexpected rewards (see, e.g., Lepper et al., 1973) or contingent versus noncontingent rewards (see, e.g., Ross et al., 1976) that have been reported in many of the primary sources cited in these recent meta-analyses.

Finally, as noted in the previous section, a sole reliance on meta-analysis in analyzing procedurally diverse literatures may lead researchers to give greatest weight to those findings that, in terms of the specific manipulations and measures they used, might otherwise have the least inferential power. For instance, studies that evaluated the benefits of some educational intervention in terms of students' performance on standardized achievement tests administered 3 months after the end of the experiment, compared with studies that examined students' scores on a test explicitly designed to cover only the specific material taught and administered the last day of the program, would be likely to produce smaller statistical effect sizes, even though we believe that their findings should rightfully be given much greater weight.

Thus, with respect to each of the above points, our concern is that more costly but inherently more probative methods will be replaced by less costly but less inherently valuable methods. A focus on simplistic conclusions and on overall main effects seems likely to hamper the future development of subtle hypotheses and the study of highly informative, but necessarily complex, effects. Likewise, an emphasis on across-study rather than within-study comparisons and on statistical as opposed to functional effect sizes may encourage procedures that make it easier to obtain statistically powerful findings, even if those studies prove, on average, less rigorous and less theoretically telling. Thus, in the marketplace of ideas, as in the marketplace of commerce, bad currency may drive out good.

General Conclusions

Are meta-analytic procedures inherently superior to, or automatically to be preferred to, more traditional reviewing methods? In the present comment, we have argued that they are not and that their superiority, or their inferiority, to traditional methods depends, in part, on both the nature of the literature being reviewed and the specific ways in which meta-analytic techniques are employed in those reviews.

Thus, we argue, meta-analyses may not be inherently more precise than traditional reviews when applied to literatures that are procedurally diverse, theoretically derived, and empirically complex. Rather, their apparent precision may sometimes prove merely an illusion produced when simplistic numerical estimates of effect size are used to compare the results of experiments that vary dramatically along dimensions associated with differences in func-

tional or psychological effect size that are independent of, or at odds with, statistical power.

Nor, we suggest, are meta-analyses inherently more objective than traditional reviews. In both cases, the conclusions reached depend as much on the specific characteristics of the research literature being reviewed and the particular choices made by those reviewing it as on the validity of the hypotheses being tested. Indeed, the dramatically different conclusions and recommendations presented in the recent meta-analyses by Cameron and Pierce (1994), Deci et al. (1999), and Tang and Hall (1995) provide a striking testament to the potential subjectivity of even quantitative literature reviews.

In short, researchers cannot rely on statistics alone to guarantee the truth or the accuracy of their conclusions.[6] Although it may be tempting to succumb to the facile assumption that meta-analysis provides clear-cut, objective answers to complex questions, as scientists we must resist this oversimplification. As one of the foremost experts on meta-analysis, Ingram Olkin, has noted, "doing a meta-analysis is easy, but doing one well is hard" (quoted in Mann, 1990). Sometimes, as in this particular literature, one might well question whether meta-analyses should be done at all.

[6] Indeed, as Abraham Kaplan has noted more generally, "the shortcomings . . . of the use of statistics in behavioral science are chiefly attributable to the tendency to forget that statistical techniques are tools of thought, and not substitutes for thought" (Kaplan, 1964, p. 257).

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership, and men.* Pittsburgh, PA: Carnegie Press.

Calder, B. J., & Staw, B. M. (1975). Self-perception of intrinsic and extrinsic motivation. *Journal of Personality and Social Psychology, 31,* 599–605.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research, 64,* 363–423.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society, 4*(Suppl.), 102–118.

Cooper, H., & Hedges, L. V. (1994). Potentials and limitations of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 521–529). New York: Russell Sage Foundation.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18,* 105–115.

Deci, E. L. (1972). The effects of contingent and non-contingent rewards and controls on intrinsic motivation. *Organizational Behavior and Human Performance, 8,* 217–229.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125,* 627–668.

Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 485–500). New York: Russell Sage Foundation.

Eisenberger, R., & Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American Psychologist, 51,* 1153–1166.

Fazio, R. H. (1981). On the self-perception explanation of the overjustification effect: The role of the salience of initial attitude. *Journal of Experimental Social Psychology, 17,* 417–426.

Hennessey, B. A., & Amabile, T. M. (1998). Reward, intrinsic motivation, and creativity. *American Psychologist, 53,* 674–675.

Hunt, M. (1997). *How science takes stock.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Kohn, A. (1996). By all available means: Cameron and Pierce's defense of extrinsic motivators. *Review of Educational Research, 66,* 1–4.

Kruglanski, A. W., Alon, S., & Lewis, T. (1972). Retrospective misattribution and task enjoyment. *Journal of Experimental Social Psychology, 8,* 493–501.

Kruglanski, A. W., Friedman, I., & Zeevi, G. (1971). The effects of extrinsic incentive on some qualitative aspects of task performance. *Journal of Personality, 39,* 606–617.

Kruglanski, A. W., Riter, A., Amitai, A., Margolin, B., Shabtai, L., & Zaksh, D. (1975). Can money enhance intrinsic motivation? A test of the content-consequences hypothesis. *Journal of Personality and Social Psychology, 31,* 744–750.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lepper, M. R. (1995). Theory by the numbers? Some concerns about meta-analysis as a theoretical tool. *Applied Cognitive Psychology, 9,* 411–422.

Lepper, M. R. (1998). A whole much less than the sum of its parts. *American Psychologist, 53,* 675–676.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology, 28,* 129–137.

Lepper, M. R., & Henderlong, J. (in press). Turning "play" into "work" and "work" into "play": 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic motivation: Controversies and new directions.* New York: Academic Press.

Lepper, M. R., Keavney, M., & Drake, M. (1996). Intrinsic motivation and extrinsic rewards: A commentary on Cameron and Pierce's meta-analysis. *Review of Educational Research, 66,* 5–32.

Lepper, M. R., Sagotsky, G., Dafoe, J. L., & Greene, D. (1982). Consequences of superfluous social constraints: Effects on young children's social inferences and subsequent intrinsic interest. *Journal of Personality and Social Psychology, 42,* 51–65.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative.* Cambridge, MA: Ballinger.

Mann, C. (1990, August 3). Meta-analysis in the breech. *Science, 249,* 476–478.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations.* Chicago: University of Chicago Press.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Miller, J. G. (1999). Cultural psychology: Implications for basic psychological theory. *Psychological Science, 10,* 85–91.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1, pp. 289–334). Cambridge, MA: Addison-Wesley.

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

Pittman, T. S., Cooper, E. E., & Smith, T. W. (1977). Attribution of

causality and the overjustification effect. *Personality and Social Psychology Bulletin, 3*, 280–283.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 6*, 228–233.

Press, S. J. (1990). Comments on the desegregation summary analysis. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 71–74). New York: Russell Sage Foundation.

Rosenhan, D. L. (1973, January 19). On being sane in insane places. *Science*, 179, 250–258.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. Philadelphia: Temple University Press.

Ross, M. (1975). Salience of reward and intrinsic motivation. *Journal of Personality and Social Psychology, 32*, 245–254.

Ross, M., Karniol, R., & Rothstein, M. (1976). Reward contingency and intrinsic motivation in children: A test of the delay of gratification hypothesis. *Journal of Personality and Social Psychology, 33*, 442–447.

Ryan, R. M., & Deci, E. L. (1996). When paradigms clash: Comments on Cameron and Pierce's claim that rewards do not undermine intrinsic motivation. *Review of Educational Research, 66*, 33–38.

Sansone, C., & Harackiewicz, J. M. (1998). "Reality" is complicated. *American Psychologist, 53*, 673–674.

Sansone, C., & Harackiewicz, J. M. (Eds.). (in press). *Intrinsic motivation: Controversies and new directions*. New York: Academic Press.

Tang, S.-H., & Hall, V. C. (1995). The overjustification effect: A meta-analysis. *Applied Cognitive Psychology, 9*, 365–404.

Tippett, L. H. C. (1931). *The methods of statistics*. London: Williams & Norgate.

Wachter, K. M., & Straf, M. L. (1990). Introduction. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. xiii-xxviii). New York: Russell Sage Foundation.

Wolf, F. M. (1986). *Meta-analysis*. Beverly Hills: Sage.

Received April 29, 1999
Revision received May 28, 1999
Accepted June 8, 1999 ■