

Conflicts of Interest in Public Policy Research *

Robert J. MacCoun **

Abstract: In this essay, I discuss the difficulty of sustaining an inquisitorial system of policy research and analysis when it is embedded in a broader adversarial political setting. Conflicts of interest in public policy research exist on a continuum from blatant pecuniary bias to more subtle ideological bias. Because these biases are only partially susceptible to correction through individual effort and existing institutional practices (peer review, replication), I consider whether a more explicitly adversarial system might be preferable to the awkward hybrid that exists today. But there are important disanalogies between policy-relevant empirical debates and the kinds of conflicts we address with our adversarial legal system. If we are stuck with a muddled inquisitorial-adversarial hybrid, we need to encourage norms of “heterogeneous inquisitorialism,” in which investigators strive for within-study hypothesis competition and greater clarity about roles, facts, and values.

THE VARIETIES OF CONFLICT OF INTEREST

In public policy research, as in other domains of professional life, conflicts of interest (henceforth, COIs) are legion. Most policy researchers can readily provide many personal war stories from their professional experience. Generically, the most blatant cases tend to fall into four categories:

1. Investigators with a commercial or proprietary interest in the research outcome, or the use of funding from sources with a commercial or proprietary interest in the research outcome (e.g., Hilts, 2000; also see the 5 June 2002 special issue of *JAMA*).

* Forthcoming in D. A. Moore, D. M. Cain, G. Loewenstein, and M. Bazerman (eds.), *Conflicts of Interest: Problems and Solutions from Law, Medicine and Organizational Settings*. London: Cambridge University Press.

** Professor, Goldman School of Public Policy and Boalt Hall School of Law, University of California at Berkeley. Please send comments to <maccoun@socrates.berkeley.edu>

2. The use of funding from sources with a political agenda that would benefit from particular research outcomes (e.g., Revkin, 2003).
3. Paid expert testimony in an adversarial legal proceeding (see Faigman, 1999).
4. The use of proprietary data sources unavailable to other investigators (Metcalf, 1998).

To varying degrees, these four categories involve pecuniary motives and interests. But even if we could eliminate pecuniary motives in policy research, we would not eliminate conflicts of interest, because there is a fifth, more subtle category:

5. The influence, whether conscious or unconscious, of an investigator's allegiance to extra-political or ideological values and attitudes.

Even if pecuniary biases and ideological biases differ in their origins,¹ they can have similar effects on research conduct and evidence interpretation.

Pecuniary ideological biases are perhaps more troublesome in domains like medicine and engineering than in public policy, simply because there are few opportunities for researchers to reap financial reward for their research (patents, commercial applications). On the other hand, ideological biases may be less pervasive in domains like medicine and engineering, because disputes are more likely to center on means (the best techniques, author credit for innovations, etc.) than on ends (whether we should improve health, safety, or performance). In the public policy arena, the ends (income equality, reproductive rights, welfare entitlements, environmental preservation, a "drug-free society") are often as contested as the means that would achieve them.

Merriam-Webster's Collegiate Dictionary defines a conflict of interest as "a conflict between the private interests and the official responsibilities of a person in a position of trust." In the domain of public policy research, we can construe both "private interests" and "a position of trust" either narrowly or broadly. Narrowly defined, "private interest" would involve the potential attainment of money, prestige, or other resources for oneself or one's organization. A broader definition would include the researcher's personal values and political views. Narrowly defined, "a position of trust" would involve particular professional offices with explicit rules proscribing bias or the pursuit of personal gain. A broader definition might invoke Robert Merton's (1973) articulation of the norms of science that are widely shared in our culture:

¹ And they may not, since ideological attitudes probably influence the seeking and receiving of biased pecuniary support.

- Scientific accomplishments should be judged by impersonal criteria (“universalism”) rather than the personal attributes of the investigator.
- Scientific information should be publicly shared (“communalism”).
- Investigators should proceed objectively, putting aside personal biases and prejudices (“disinterestedness”).
- And the scientific community to hold new findings to strict levels of scrutiny, through peer review, replication, and the testing of rival hypotheses (“organized skepticism”).²

There are now a good many published case studies documenting conflict of interest in this broader sense in many research domains, including HIV/AIDS (Epstein, 1996), tobacco (Cummings, Sciandra, Gingrass, & Davis, 1991; Glantz, 1996), sexual orientation (LeVay, 1996), intelligence testing (e.g., Fraser, 1995), drug prevention (Gorman, 2003; Moskowitz, 1993), risk prevention (Fischhoff, 1990), marijuana statistics (MacCoun, 1997), and global warming (Gelbspan, 1997).

“Conflict of interest” can also be defined intrapersonally or interpersonally. In the traditional sense, the “conflict” is *intrapersonal* – a conflict between her role obligations and her behavior. But the term “conflict of interest” has also been used in a very different sense in the social psychology literature, one that defines the conflict *interpersonally*, between people or factions of people. For example, in the small group literature, McGrath’s (1984) group task circumplex defines “conflicts of interest” in terms of mixed-motive payoff structures. Of greater relevance to this essay, John Thibaut (a psychologist) and Laurens Walker (a lawyer) (1978) distinguish “cognitive conflicts,” where the parties have a joint interest in solving a problem, from “conflicts of interest,” where “a particular solution will maximize the outcome of one of the parties only at the expense of the other.” The intrapersonal definition is more conventional, but most real-world examples meet the interpersonal definition as well.

Identifying the Right Normative System

Thibaut and Walker (1978) identify the goal of cognitive conflicts as “truth,” and the goal of conflicts of interest as “justice.” These are lofty claims, but then, the authors had a lofty goal. Their work builds on a highly influential program of empirical research on citizen evaluations of

² Koehler (1993) presents evidence that scientists endorse such norms.

alternative conflict resolution procedures (Thibaut & Walker, 1975; also see Lind & Tyler, 1988; MacCoun, Lind & Tyler, 1992), but their theoretical goal is normative, not descriptive. They want to define the proper domains for inquisitorial vs. adversarial procedures of conflict resolution. In an adversarial process, as exemplified by the Anglo-American trial system, disputants retain “process control” by selectively presenting the facts most favorable to their position to a 3rd party decision maker. In an inquisitorial process, evidence is assembled by the 3rd party decision maker, or by a neutral investigator who reports to that decision maker. Some continental European legal systems are inquisitorial in this sense, but more relevant for present purposes, Merton’s norms of scientific practice are inherently inquisitorial.

Thibaut and Walker make two normative claims. First, “an autocratic system delegating both process and decision control to a disinterested third party is most likely to produce truth,” hence cognitive conflicts should be resolved through the inquisitorial method. Second, “a procedural system designed to achieve distributive justice...will function best if process control is assigned to the disputants,” as exemplified by “the Anglo-American adversary model.” Thibaut and Walker’s normative theory has been much less influential than their empirical program, and although their treatment is more sophisticated and nuanced than this brief sketch, I don’t find it entirely persuasive. Though it is useful to examine ideal types in the laboratory, few real-world problems seem to fit neatly into these cognitive-conflict and conflict-of-interest bins. Thibaut and Walker allow for the possibility of “mixed conflicts,” but this category – arguably the largest category – they call for a mix of inquisitorial and adversarial procedures.

Hence, I call attention to their theory because of the question they pose rather than the answers they offer. I believe that the central problem of conflict of interest in public policy research is the *blurring of adversarial and inquisitorial norms and roles*. Public policy research and the utilization of that research, falls far short of Merton’s inquisitorial ideals. Yet our allegiance to those norms, and our pretense to be operating under those norms, also keeps us from realizing some benefits of a more explicitly adversarial approach. We largely seem to muddle in the middle. In the abstract, a purely inquisitorial model might well be best, but we are unlikely to achieve one. An explicit, robust adversarial research process might be more attainable, and it might even have some advantages over a muddled mixed model, where some investigators play by one set of rules, some play by another, and some vacillate back and forth either strategically or unwittingly. But the adversarial model, whatever its merits in legal

settings (and those are decidedly mixed), has serious drawbacks outside the trial context. What are needed are clearer norms defining a realistically heterogeneous inquisitorialism.

What Isn't In This Essay

The topic of bias in politically relevant research is an old one, and it has been examined many times before from other angles. There is an enormous literature on the details of bias in research methodology, including biased research designs, biased statistical analyses, biased data presentation, and experimenter expectancy effects (e.g., Rosenthal, 1994). Note that the kinds of investigator biases examined here may express themselves through these methodological problems, but bad methodology may instead reflect ignorance or real-world data constraints rather than bias on the part of the investigator.

There is an extensive sociology literature on the effects of institutional factors, professional incentives, social networks, and demographic stratification on the scientific research process (see Cole, 1992; Zuckerman, 1988). And of course the troubled relationship between facts and values was a preoccupation of twentieth century philosophy of science (see Gholson & Barker, 1985; Laudan, 1990; Shadish, 1995). I also sidestep the postmodernist literatures on social constructivism, deconstructionism, hermeneutics, and the like, for reasons explained elsewhere (MacCoun, 1998, 2003; also see Gross & Levitt, 1994).

ATTRIBUTING BIAS TO OTHERS

It is very easy to attribute bias to researchers, and observers readily do so. But how are we to know whether the bias resides in the attributor, rather than (or in addition to) the investigator? The same forces that can produce bias in researchers can produce biases in consumers of that research. A case in point is the cottage industry in books denouncing “junk science.” These books are quick to criticize particular experts for sloppy and careless thinking – especially an overreliance on unsystematic and unrepresentative clinical case evidence and an underreliance on rigorous multivariate analysis and controlled experimentation. And what evidence do the authors offer for their indictment of junk science? Anecdotes about particular cases and particular experts, selected by an unspecified but surely non-random sampling process, with no correction for hindsight bias (the critiques make ready use of later science unavailable to the experts at the time in question), and no consideration of alternative motives for the expert

testimony (MacCoun, 1995). I am not questioning whether sloppy or biased expert testimony occurs – it surely does – but simply arguing that we are often willing to attribute bias based on the “junkiest” of evidence.

Obviously, some attributions of bias are self-serving; if an investigator presents findings you don't like, the quickest way to discredit her – much quicker and more reliable than conducting your own study – is to question her motives or her integrity. But there are also some more subtle cognitive phenomena that complicate the attribution process.

In a classic experiment by Jones and Harris (1967), students were enlisted to conduct an in-class debate on the topic of whether mid-1960s America ought to adopt a friendlier stance toward Fidel Castro. Half the participants were told that the debaters were assigned their positions by the debating coach, half were not told anything about how the debating roles were determined. After the debate, audience members estimated speakers' actual attitudes toward Castro. The audience overwhelmingly assumed the “pro-Castro” debater was indeed pro-Castro. This was true not only in the control condition, but among audience members who knew that the debating position was situationally determined (by the coach). Subsequent research using this “attitude attribution” paradigm has shown that this “shoot the messenger” tendency is quite robust (see Ross & Nisbett, 1991), exemplifying what Ross calls “the fundamental attribution error” – the tendency to give disproportionate weight to dispositional explanations (those internal to the actor, like traits and desires) for others' behavior, while discounting or overlooking situational influences.

The problem is a familiar one for policy analysts; audiences often assume that we favor (and, presumably, always favored) whichever political viewpoint our findings most readily benefit. In a recent statewide telephone survey of 1,050 California adults (MacCoun and Paletz, 2004), we found that for controversial social interventions -- gun control, capital punishment, medical marijuana, and school vouchers (but not for healthy eating habits) -- over half of all respondents were willing to speculate on a researcher's ideological beliefs knowing nothing other than whether a hypothetical study found that the intervention worked. For gun control and medical marijuana, positive findings led them to infer that the researcher was a liberal; for capital punishment, positive findings implied that the researcher was a conservative. Rather than viewing social science as an attempt to reveal facts about the world (the “discovery” model of

research), many citizens construe social science as a process of political exhortation, and social scientists are seen advocates who find what they want to find.

Naïve Realism

Ross and his colleagues (Pronin, Lin, & Ross, 2002; Robinson et al., 1995) have argued that humans are predisposed to assume that our views of the world are objective and veridical, and to neglect the ways in which our perceptions might be filtered by our biases and distorted by the evidence available to us. Because of this “naïve realist” stance, we tend to assume that those who disagree with us must be plagued by subjectivity, blinded by desire, or just plain confused. Thus, Pronin et al. (2002) have demonstrated that most people believe that they themselves are much less susceptible to judgmental biases than the average person in their peer group. And partisans on both sides of a dispute tend to see the exact same media coverage as favoring their opponents’ position (the “hostile media phenomenon”; Vallone, Ross, & Lepper, 1985). A corollary is that people are likely to genuinely believe that research that coincides with their own beliefs must be less biased and more objective than research that favors other positions (Lord, Ross, & Lepper, 1979; MacCoun & Paletz, 2004).

Adjusting for Perceived Bias

There is evidence that observers try to adjust their interpretation of new evidence based on their perceptions of bias in the source (Wegener et al., 2000). Thus, the results of a study by an ostensibly liberal research are assumed to be somewhat less favorable to a liberal position than actually reported. This kind of adjustment process sounds like good news, but the catch is that the adjustment is based not on actual bias, but on “individuals’ naïve theories of the biasing factor(s) at hand” (Wegener et al., 2000, p. 630). So bias correction improves validity when bias attributions are valid, but bias correction will introduce distortions when they are not. In research on communications between University officials and their seismic engineering consultants, we found anecdotal evidence for this distortion process (DeVries et al., 2001). Each side assumed that the other was biased, and adjusted their interpretations accordingly. Making matters worse, the more experienced actors on each side assumed the other side was making such adjustments, and *adjusted for the adjustments* so that their “adjusted positions” would match

their true positions. It would be desirable to “hit the reset button” so that all parties could see exactly where everyone else stood, but how to accomplish this wasn’t obvious to us.

One might hope that a source’s open disclosure of a COI would help observers to correctly adjust for the source’s bias. Distressingly, Cain, Loewenstein, and Moore (Chapter X in this volume) present new evidence suggesting that audiences fail to fully discount such biases. Moreover, they show that sources who disclosed a COI actually behaved in a more biased fashion. Cain and colleagues used a task that was explicitly inquisitorial rather than adversarial; it remains to be seen whether observers fare better in more blatantly adversarial settings.³

THE VARIETIES OF INVESTIGATOR BIAS

Documenting Bias

The biases I describe have been variously operationalized using one of the four methodological strategies discussed by Hastie and Rasinski (1988; Kerr et al., 1996; MacCoun, 1998, 2002): (a) documenting differences between observers (weak because it doesn’t show who is biased), (b) documenting a difference between a judgment and a normatively defined true value, (c) documenting the use of a normatively proscribed cue, or (d) documenting the failure to use a normatively prescribed cue. (The normative system in question can be Bayes theorem, classical statistics, decision theory, the legal rules of evidence, or Merton’s aforementioned norms of science).

It is easier to establish bias in the laboratory than in field studies. “Researcher allegiance” effects have been reported in meta-analyses of the research literatures on psychotherapy (Gaffan et al., 1995; Robinson et al., 1990; Shadish et al., 1993) and drug treatment (Prendergast et al., 2002). These analyses suggest that investigators with identifiable allegiances with a treatment program report significantly larger treatment effects than other, more disinterested investigators. For example, Gaffan, Tsaousis, and Kemp-Wheeler (1995) meta-analyzed outcome data from clinical trials of treatments for depression. They estimated that about half of the relative superiority for cognitive therapy reported in previous meta-analyses was attributable to researcher allegiance.

³ Also, Cain and colleagues obtained these effects in studies where the source provided quantitative information, leading to anchoring effects. Arguably, it might be easier to discount biased sources in domains where their information is qualitative or categorical (“vouchers worked,” “right-to-carry gun laws prevent crimes,” and so on).

On its face, this is a stunning claim, but the researchers used a very broad conceptualization of allegiance. Among the study features that, in their view, constituted strong allegiance were: “reference to previous published research showing the superiority of X to some other treatment,” “specific hypothesis or rationale as to why X should be superior to other treatments in this study,” “X was devised or first introduced by one of the authors,” and “X is the only treatment included in the study, and the authors regard it as superior to other available treatments.” Another problem is that indicators of allegiance are confounded with other factors, and may disappear when such factors are controlled (Shadish et al., 1993).

Still, more detailed case studies of program evaluation practices leave little doubt that such allegiance effects do occur. In a series of papers, Gorman (1998, 2002, 2003) has documented numerous highly misleading statistical practices deployed by prevention program designers evaluating their own programs. Reported outcomes effects are frequently based on only a carefully chosen subset of the study population and the dependent variables. Gorman’s own “intention-to-treat” re-analyses suggest that these programs are far less successful than reported. Still, when pressed, program advocates will claim such practices are meaningful in helping to provide “fair tests” of what the programs are capable of when “implemented with high fidelity.”

Perhaps the most unequivocal evidence for biased evidence evaluation comes from controlled laboratory demonstrations. Mahoney (1977) and Lord, Ross, and Lepper (1979) developed the basic experimental paradigm that is now used to study biased interpretation of research data. Participants – usually practicing scientists or graduate students with professional training in methodology – are asked to assess research studies. Unbeknownst to them, they have been randomly assigned to receive one of several experimental variants of a research manuscript, with the ostensible methods and results systematically varied. The participants are more persuaded by findings that support their own (previously assessed) political views, even when the methodology is identical. A given methodology is viewed favorably when it produces congenial results, and critically when it does not.⁴ These “biased assimilation” results (Lord et al., 1979) have been widely replicated, though there are boundary conditions (see MacCoun,

⁴ Lord et al. also argued that such situations actually produce *attitude polarization*, such that respondents became *more* extreme in the direction of their initial views. Happily, this truly perverse finding has not been replicated (see MacCoun, 1998).

1998 for a review). For example, biased assimilation effects are robust among judges with extreme attitudes, but more difficult to replicate among those with moderate views (Edwards & Smith, 1996; McHoskey, 1995; Miller et al., 1993).

Prototypical Forms of Bias

On its face, biased assimilation seems like a corruption of Merton's ideals of scientific reasoning. But a fair assessment – and a search for solutions – requires an inquiry into the causes of the bias. Elsewhere (MacCoun, 1998) I have sketched five prototypes of biased evidence processing (see Table 1). The prototypes vary with respect to intentionality, motivation, and normative justifiability. By *intentionality*, I refer to the combination of consciousness and controllability; a bias is intentional when the judge is aware of a bias, yet chooses to express it when she could do otherwise (see Fiske, 1989). *Motivation* is shorthand for the degree to which the bias has its origins in the judge's preferences, goals, or values; intentional bias is motivated, but not all motivated biases are intentional. Finally, *normative justification* distinguishes appropriate or defensible biases from inappropriate or indefensible biases, relative to some normative system (e.g., Bayesian decision theory, the rules of evidence in law).

Table 1. Five prototypical forms of investigator bias.

	<i>Intentional?</i>	<i>Motivated?</i>	<i>Justifiable?</i>
<i>Fraud</i>	Yes	Yes	No
<i>Advocacy</i>	Yes	Yes	Maybe
<i>Hot bias</i>	No	Yes	No
<i>Cold bias</i>	No	No	No
<i>Skepticism</i>	Maybe	Yes	Yes

Fraud. The first prototype is *fraud*—intentional, conscious efforts to fabricate, conceal, or distort evidence, for whatever reason—material gain, enhancing one's professional reputation, protecting one's theories, or influencing a political debate (see Fuchs & Westervelt, 1996; Murray, 2002; Woodward & Goodstein, 1996). At a macro level, they are often explicable from sociological, economic, or historical perspectives (Cole, 1992; Zuckerman, 1988). At a micro

level, they are sometimes explicable in terms of individual psychopathology. These cases are extremely serious, but again, I am interested here in less blatant, more subtle problems.

Absent direct evidence of intent, it can be difficult to distinguish fraud from ignorance or incompetence. For example, a study reporting a 1 to 2% average difference in various comparisons of marijuana prevalence between Dutch vs. U.S. youth (MacCoun & Reuter, 1997) was widely cited as finding a 32% difference (see MacCoun, 2001). This was done by selectively quoting estimates from different parts of a table providing fair year-by-age range comparisons. This could have been a simple misunderstanding of the study, but the estimates chosen were those most likely to make marijuana look more popular in the Netherlands (18 year olds in 1996) than the U.S. (12-17 year olds in 1992).

Advocacy. A second prototype is *advocacy*--the selective use and emphasis of evidence to promote a hypothesis, without outright concealment or fabrication. As I discuss below, advocacy is normatively defensible provided that it occurs within an explicitly advocacy-based organization, or an explicitly adversarial system of disputing. Trouble arises when there is no shared agreement that such adversarial normative system is in effect. When we speak of an investigator as being ideologically driven or biased, we imply that his or her attitudes or values have influenced his or her interpretations of the evidence. This is clearly a violation of the impartial inquisitorial model. But of course, the temporal and causal sequence is often reversed. Is it truly desirable, much less feasible, for an investigator's attitudes and beliefs to "kept in a lockbox," hermetically sealed from his or her research findings? I return to this quandary in a later discussion of inquisitorial vs. adversarial role conflicts.

Hot and cold biases. Contemporary psychology recognizes that most biased evidence processing can occur quite unintentionally through some combination of "hot" (i.e., motivated or affectively charged) and "cold" cognitive mechanisms. The prototypical *hot bias* is unintentional and perhaps unconscious, but it is directionally motivated—the judge wants a certain outcome to prevail. I suspect this is what most people have in mind when they speak of "biased" researchers.

But in professional psychology for the past several decades, the focus has been on "cold" unmotivated biases. The prototypical *cold bias* is unintentional, unconscious, and it occurs even when the judge is earnestly striving for accuracy. Numerous mechanisms have been identified in basic cognitive psychological research on memory storage and retrieval, inductive inference, and

deductive inference that can produce biased evidence processing even when the judge is motivated to be accurate and is indifferent to the outcome.⁵ Arkes (1991) and Wilson and Brekke (1994) have offered taxonomies for organizing these different sources of judgmental bias or error, and offer detailed reviews of the relevant research. These cold biases are an important source of bias in research (MacCoun, 1998) but because they are non-motivational they seem less relevant to conflicts of interest than the other prototypes.

Tetlock and Levi (1982) made a persuasive case for the difficulty of definitively establishing whether an observed bias is due to hot vs. cold cognition; the recent trend has been toward integrative “warm” theories (Cohen, Aronson, & Steele, 2000; Kunda, 1990; Kruglanski & Webster, 1996; Liberman & Chaiken, 1992; Pyszczynski and Greenberg, 1987). Most “hot” and “warm” accounts examine *directional* biases favoring a particular conclusion – what Kruglanski (1989) calls a “need for specific cognitive closure”. But another form of bias is the motivation to “find something” rather than finding nothing – what Kruglanski (1989) calls a “need for nonspecific cognitive closure.” The most obvious source of such a bias is a professional reward system that rewards studies that “reject the null hypothesis” (statistical jargon for “finding something”) – that find an effect of an intervention or a significant association rather than a lack of effect or a lack of association.⁶ But professional policy analysts often feel enormous pressure to “find something” to justify their efforts. And many a professional policy briefer has been on the receiving end of an angry policy maker’s tirade: “Don’t tell me you need more research! Don’t say ‘it depends’! Tell me right now, yes or no, what should I do?”⁷

Skepticism.⁸ Research on biased processing of scientific evidence has given somewhat less attention to the final prototype, which might be called *skeptical* processing. In skeptical

⁵ As an example, the availability heuristic (Kahneman, Slovic, & Tversky, 1982) is our tendency to give disproportionate weight to those items of evidence that come most readily to mind (because they are vivid, were encountered recently, or have received lots of media coverage).

⁶ The difficulty of publishing null results is well known, but it is defensible when studies are plagued by noisy measurement and/or inadequate sample sizes.

⁷ The briefer may rightfully respond “wait a second, that’s your job, this is my job,” but I have learned that this will make the briefing end very badly. The question seems to reflect a mix of alpha-dog posturing, genuine frustration, and some magical thinking about the possibility that the public will forgive a bad decision if the policy maker was poorly advised.

⁸ This section is a slight revision of a similar section in MacCoun (1998).

processing, the judge interprets the evidence in an unbiased manner, but her conclusions may differ from those of other judges because of her prior probability estimate, her asymmetric standard of proof, or both. This is arguably normative on decision theoretic grounds, but those grounds are controversial.

In a highly simplified decision theoretic analysis of scientific evidence evaluation, the judge assesses $p(H|D)$, the conditional probability of the hypothesis (H) given the data (D). Most of the research reviewed thus far has focused on this judgment process. Of course, in a simplified Bayesian model, $p(H|D)$ equals the diagnosticity of the evidence, $p(D|H)/p(D)$, weighted (multiplied) by the judge's *prior probability* (or "prior"), $p(H)$. (More complex models appear in Howson & Urbach, 1993; Schum & Martin, 1982).

For a Bayesian, the prior probability component is an open door to personal bias; so long as diagnosticity is estimated in a sound manner and integrated coherently with one's "priors," the updated judgment is normatively defensible (see Koehler, 1993). Of course, the normative status of this framework is a source of continuing controversy among philosophers and statisticians (see Mayo, 1996), especially the notion of subjective priors. Moreover, challenges to the theory's descriptive status (Arkes, 1991; Kahneman, Slovic, & Tversky, 1982; Pennington & Hastie, 1993) leave its normative applicability in doubt. And much of the evidence reviewed here implies that the diagnosticity component is itself a major locus of bias, irrespective of the judge's prior.

But decision theory also identifies a second, less controversial locus of potentially defensible "bias." Our probabilistic assessment of the hypothesis yields a continuous judgment on a 0-1 metric, yet circumstances often demand that we reach a *categorical* verdict: Will we accept or reject the hypothesis? This conversion process requires a *standard of proof*. Statistical decision theory, signal detection theory, and formal theories of jurisprudence share a notion that this standard should reflect a tradeoff among potential decision errors. A simple decision theoretic threshold for minimizing one's regret is $p^* = u(\text{FP})/[u(\text{FN}) + u(\text{FP})]$, where $u(\text{FP})$ equals one's aversion to false positive errors, and $u(\text{FN})$ denotes one's aversion to false negative errors (see DeKay, 1996; MacCoun, 1989). The standard of proof, p^* , cleaves the assessment continuum into rejection and acceptance regions. Thus the standard of proof reflects one's evaluation of potential errors, and this evaluation is extra-scientific, arguably even in the case of the conventional 0.05 alpha level.

When one error is deemed more serious than the other, the standard of proof becomes asymmetrical, and can easily produce greater scrutiny of arguments favoring one position over another. Thus, even for most non-Bayesians, there is a plausible normative basis for “bias” in assessments of scientific research (see Hammond, Harvey, & Hastie, 1992). Note however, that this form of bias is limited to qualitative, categorical decisions (“it’s true”; “he’s wrong”); it cannot justify discrepancies across judges (or across experimental manipulations of normatively irrelevant factors) in their quantitative interpretations of the diagnosticity of evidence.

Are These Biases Controllable?

Of the five prototypes, fraud and advocacy are arguably under the conscious control of the actor, but the other three often operate in an unconscious and automatic fashion (see Wegner & Bargh, 1998). Judgmental biases are remarkably resistant to eradication efforts; they tend to persist in the face of education (Arkes, 1991; Wilson & Brekke, 1994), incentives for accuracy (Camerer & Hogarth, 1999), and many forms of public accountability (Lerner & Tetlock, 1999). Three strategies that are at least somewhat successful at reducing bias are the so-called “consider the opposite strategy” (Lord, Lepper, & Preston, 1985), “devil’s advocacy” role-playing (Schwenk, 1990), and accountability to audiences of unknown or mixed viewpoints (Lerner & Tetlock, 1999). The mechanism may be the same in each case – getting the actor to actively consider alternative, competing points of view. This raises the question: Even without formal debiasing interventions, will the rough and tumble of collective adversarial debate correct these individual judgmental biases?

COMPARING THE ADVERSARIAL AND INQUISITORIAL MODELS

In her splendid book *The Argument Culture*, Deborah Tannen (1998, p. 3) describes “a pervasive warlike atmosphere that makes us approach public dialogue, and just about anything we need to accomplish, as if it were a fight. ...[This] argument culture urges us to approach the world—and the people in it—in an adversarial frame of mind.” This adversarial mindset is a self-fulfilling prophecy. Keltner and Robinson (1996) review evidence that the gap between partisans’ perceptions in a variety of attitudinal disputes are objectively much smaller than each side believes. Careful public surveys rarely show stark bimodal distributions of opinion; rather, there often is a continuum of viewpoints that gets bifurcated by the way journalists summarize

the results. But journalists, political parties, civil litigation, and our brain's own categorization processes dichotomize important policy problems, encouraging us to take sides.

In the Anglo-American adversarial legal system, advocates actively seek and selectively report the most favorable evidence for their clients. This approach is defended as a means of finding the facts; the traditional claim is that the "truth will out." Surprisingly few studies have directly compared the relative ability of adversarial and inquisitorial methods for accurately determining facts.⁹ It appears (see Lind & Tyler, 1988) that adversary proceedings may work well for legal discovery; mock attorneys playing an adversarial role seek out as much evidence or more evidence as neutral inquisitors. But this is offset by systematic distortions in fact presentation (Lind & Tyler, 1988). But when evidence strongly favors one party, evidence presented at trials is misleadingly symmetrical, exaggerating the facts in support of the other party. And in adversarial proceedings, witnesses slant their testimony in a direction that favors whichever party called them to testify.

Because of such problems, Thibaut and Walker (1978) argued that the inquisitorial method is to be preferred for "truth conflicts," purely cognitive disagreements in which the parties are disinterested (or have shared interests) and simply want to discover the correct answer. But they asserted that the adversarial approach is to be preferred for "conflicts of interest" in which the parties face a zero-sum (or constant sum) distribution of outcomes.

Even if legal disputes were to fit this dichotomy (and I don't believe they do), public policy disputes surely do not. There are purely technical policy analysis problems (queing, optimization, and the like), but anything that merits the label "dispute" involves a messy blend of truth conflicts and conflicts of interest, making it difficult to separate factual disputes from value disputes (see Hammond, 1996; Tetlock et al., 1996).

Making matters worse, features of the legal system that may promote good adversarial fact-finding are lacking in public policy research disputes (MacCoun, 1998; also see Burk, 1993). Five of these features are highlighted in Table 2.

⁹ There are lots of studies of accuracy *within* the adversarial context (eyewitnesses, jury comprehension, and so on).

Table 2. Features that Distinguish Legal and Scientific Fact-Finding

	<i>Legal Fact-Finding</i>	<i>Scientific Fact-Finding</i>
Explicit adversarial role?	Yes	No
At least 2 sides represented?	Yes	Not always
Explicit standard of proof?	Yes	Yes/No
Explicit 3rd party decisionmaker?	Yes	No
Positions bound the truth?	Usually	Rarely

Role clarity. In legal disputes, the adversarial roles of the participants are quite explicit; no one mistakes an American trial lawyer for a dispassionate inquisitor. Despite the popularity of lawyer jokes, surveys show that Americans (and Europeans) actually like the notion of a fierce adversary; as LBJ said in another context, “he may be a son of a bitch, but at least he’s *my* son of bitch.”

As expressed by Merton’s (1973) norms, citizens in our culture have very clear role expectations for scientists; if one claims the authority of that role, one is bound to abide by its norms or risk misleading the public. This surely doesn’t preclude advocacy activities on the part of scientists, but it does mean that we must be quite explicit about which hat we are wearing when we speak out, and whether we are asserting our facts (e.g., the death penalty has no marginal deterrent effect) or asserting our values (e.g., the death penalty degrades human life).

Graduate training in schools of public policy analysis is much more explicit about managing these conflicting roles. For example, Weimer and Vining’s (1992) textbook provides a neutral discussion of three different professional models: the *objective technician* who maintains a distance from clients, but lets the data “speak for itself,” avoiding recommendations; the *client’s advocate* who exploits ambiguity in the data to strike a balance between loyalty to the facts and loyalty to a client’s interests; and the *issue advocate* who explicitly draws on research opportunistically in order to promote broader values or policy objectives.

Who decides, and who holds the burden of proof? In legal disputes, there is explicit agreement about the standard of proof, burden of proof (who wins in a tie?), and ultimate decision maker (i.e., the judge or jury). Disputes over scientific findings typically lack an explicit standard of proof, and an explicit final decision maker. This contributes to the seeming intractability of many debates; when each observer is free to establish her own p^* , there are no grounds for consensus on who “won.” Expert panels assembled by the National Academy of Sciences and other organizations attempt to circumvent this problem, with mixed success. This isn’t entirely a bad thing. Research findings are rarely a direct determinant of policy decisions, and social scientists are sometimes strikingly naïve about the gaps between our research findings and the inputs needed for sound policy formation (see Weimer & Vining, 1992).

Trials are actually unusual in having a single burden of proof. Indeed, in policy disputes (and social argumentation more generally; Rips et al, 1999), there are multiple burdens in play. Participants and political factions have their own notions of where the burden lies – almost always with the other side. But political power (incumbency and/or public opinion) often creates an overarching burden that is greater than one side than for the other. For example, MacCoun and Reuter (2001) argue that the drug legalization involves three standards of proof – a philosophical burden on prohibitors to explain why liberty should be curtailed, a policy-analytic burden to prove that the benefits of legal change would outweigh the costs, and a political burden on legalizers to provide overwhelming evidence that drug use would not rise.

Policy analytic standards of proof have justified by principles of logic, statistics, and epistemology, but the result is not always politically neutral. A clear example is null hypothesis testing, where the .05 convention for statistical significance puts much greater emphasis on avoiding false positive findings (saying an intervention works when it doesn’t) than on avoiding false negative findings (failing to recognize a truly beneficial intervention). Increasingly rigorous econometric standards and the traditional emphasis on internal validity over external validity can have a similarly conservative effect. I recently publically defended a National Academy of Science critique of drug treatment research before an audience of angry treatment experts (see Horowitz, MacCoun, & Manski, 2001). Not one of them directly challenged our argument that treatment estimates were vulnerable to selection biases and regression to the mean; instead, they decried the patent unfairness of holding treatment to such a high standard when drug law enforcement is more generously funded without any evaluation.

Is anyone right? Finally, in many (though not all) legal disputes, the opposing positions “bound” the truth, either because one of the positions is in fact true, or because the truth lies somewhere between the two positions. But the history of science (e.g., Gholson & Barker, 1985; Thagard, 1992) reveals little basis for assuming that the truth is represented among those factual positions under dispute at any given moment (also see Klayman & Ha, 1987). This underscores the inherent ambiguity of using discrepancies among judges to locate and measure bias (Kerr et al., 1996)—all of us might be completely off target.

WILL COLLECTIVE JUDGMENT CORRECT INDIVIDUAL BIASES?

***Will “Truth Win” Via Collective Rationality?*¹⁰**

Institutional practices like peer review, expert panels (e.g., National Academy of Sciences, Institute of Medicine), and expert surveys (e.g., Kassin, Ellsworth, & Smith, 1989) are premised on a belief that collective judgment can overcome individual error, a principle familiar to small-group psychologists as the *Lorge-Solomon Model A* (Lorge & Solomon, 1955). (Model B having long since been forgotten.) In this model, if p is the probability that any given individual will find the “correct” answer, then the predicted probability that a collectivity of size r will find the answer is $P = 1 - (1 - p)^r$. Implicit in this equation is the assumption that if at least one member finds the answer, it will be accepted as the collectivity’s solution—the so-called *Truth Wins* assumption (e.g., Laughlin, 1996). This can only occur to the extent that group members share a normative framework that establishes the “correctness” of the solution. That framework might be acknowledged by most academicians (the predicate calculus, Bayes Theorem, organic chemistry), or it might not (e.g., astrology, numerology, the I Ching).

For almost half a century, social psychologists have tested the “truth wins” assumption for a variety of decision tasks (see Laughlin, 1996; Kerr et al., 1996). Even in purely intellectual tasks, “truth” rarely wins, in the strict sense that a solution will be adopted if a single member identifies or proposes it. At best, “truth supported wins”—at least some social support is needed for a solution to gain momentum, indicating that truth seeking is a social as well as intellectual process (see Laughlin, 1996). But even that only occurs in limited settings. When members lack a shared conceptual scheme for identifying and verifying solutions—what Laughlin calls

¹⁰ This discussion is adapted from MacCoun (1998).

“judgmental” as opposed to “intellective” tasks—the typical influence pattern is *majority amplification*, in which a majority faction’s influence is disproportionate to their size, irrespective of the truth value of their position (see Kerr et al., 1996).

In theory, collective decision making (or statistical aggregation of individual judgments) is well suited for reducing *random error* in individual judgments. (Indeed, this is a major rationale for meta-analysis, discussed below.) What about bias? A common assertion is that group decision making will correct individual biases, but whether in fact this actually occurs depends on many factors, including the strength of the individual bias, its prevalence across group members, heterogeneity due to countervailing biases, and the degree to which a normative framework for recognizing and correcting the bias is shared among group members (see Kerr et al., 1996). Elsewhere, my colleagues and I (Kerr et al., 1996; MacCoun, 2001) have demonstrated that under a wide variety of circumstances, collective decision making will *amplify* individual bias, rather than attenuating it. The collective will be *less* biased than its individuals when:

1. The correct answer is obvious to almost everyone, or
2. there is “strength in arguments,” such that a shared conceptual scheme allows participants to recognize a “correct” (relative to that scheme) result and endorse it. (This is the aforementioned Lorge-Solomon rule.)

The collective will tend to *amplify* individual bias when:

1. There is “strength in numbers,” such that large factions have influence disproportionate to their size, as will occur explicitly in a “majority rules” system and implicitly in any “majority amplification” process.
2. The case at hand is “close” rather than lopsided.

Our analysis focused on small groups reaching collective decisions. Collective research interpretation is of course quite different – aside from the occasional blue-ribbon panel or NAS/NRC/IOM committee, there is rarely any explicit group sitting in one place to determine what the evidence says. Rather, collective research interpretation is diffuse, spread over multiple audiences and decision makers over an indeterminate period of time.

Nevertheless, the mathematical framework used by Kerr et al. (1996) is sufficiently abstract that the conclusions summarized above are probably generally true. If there is a collective process that favors faction strength (where strength could refer to economic power

rather than faction size), and the case is close, the collective may well amplify bias. If there is a collective process that favors argument strength (such that an argument once voiced is highly persuasive), the collective will tend to be less biased than the individuals in it.

“Strength in arguments” sometimes prevails in adversarial systems; e.g., there are legal defenses or evidentiary problems that will lead most lawyers to reject a case. And of course “strength in numbers” effects are well-documented in the sociology of science (see Cole, 1992; MacCoun, 1998), where trendiness, networking, and social stratification can privilege some hypotheses and findings irrespective of their validity. But by its very nature, an inquisitorial system seems more likely than an adversarial system to favor “strength in arguments” over raw factional strength or resources.

DOES ACCUMULATING EVIDENCE DRIVE OUT BIAS?

In our analyses of collective bias (Kerr et al., 1996; MacCoun, 2001), groups were more likely to correct individual biases when strong evidence favored one position. This is observed, for example, in mock jury experiments using strongly slanted trial evidence; an occasional juror will endorse the less popular position, but he or she will quickly yield to overwhelming majority argumentation (and, sometimes, ridicule or disdain).

Kalven and Zeisel’s (1966) “liberation hypothesis” contends that jurors are most likely allow personal sentiments to influence their verdicts when the trial evidence is ambiguous. Similarly, physicist and science fiction author Gregory Benford (1980) proposes a “law of controversy”: “Passion is inversely proportional to the amount of real information available.” In support, MacCoun (1990; Kerr et al., 1996) cites several lines of individual- and group-level research demonstrating enhanced extra-evidentiary bias when evidence is equivocal.

Pyszczynski and Greenberg (1987) argue that while motivation influences hypothesis testing, most of us feel constrained by the desire to maintain an “illusion of objectivity.” Similarly, Kunda (1990, p. 482) argues that directional biases “are not unconstrained: People do not seem to be at liberty to conclude whatever they want to conclude merely because they want to. Rather... people motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer.” But we should be wary of overstating the case for an “objectivity constraint”. Even when evidence is strong and unidirectional, we may have difficulty recognizing it. “Cumulative

meta-analyses,” in which a running (weighted) average effect size is updated over time, show that research communities are often slow to realize that the accumulated evidence decisively favors a proposition (see Ioannidis & Lau, 2001; Mullen et al., 2001). And in real-world policy conflicts, the evidence is often too weak to constrain partisan judgment. Moreover, there is often evidence in support of each faction.

What Makes a Case “Close”?

Partisans operating in good faith can disagree on the facts for a variety of reasons (see Table 3). Normatively, any of the deadlocks in Table 3 might be expected to “give way” in the face of a sufficient accumulation of evidence favoring a given proposition. In reality, whether this will happen seems increasingly less plausible as one moves down the list.

Table 3. “Close cases” and their prospects for resolution

<i>What makes the case “close”?</i>	<i>Will more and better evidence resolve the dispute?</i>
Indifference on the part of the participants.	Probably, if they are motivated and able to process the evidence.
Lack of pre-existing evidence, such that individual choices “could go either way.”	Probably, if they are motivated and able to process the evidence.
Opposing factions with the same interpretation of facts, but different Bayesian priors.	Yes, if they are Bayesian updaters, but people tend to “anchor and adjust” instead.
Opposing factions with the same interpretation of facts, but different standard-of-proof thresholds.	One side will be persuaded long before the other, and the evidence may never be sufficient to cross very stringent thresholds.
Directly conflicting evidence regarding the same proposition.	The discrepancies will be viewed with suspicion on both sides, and until they are satisfactorily explained, each side will have grounds for holding firm.
Clear evidence for each of two or more propositions that evoke conflicting values.	The conflict will persist, but rather than reframing the debate as a value conflict, each side may cite its preferred “facts.”

In cases of indifference, or of genuinely equivocal evidence, people may quickly accept a proposition once the available evidence in its favor is overwhelming (or unopposed), provided they have the motivation and ability to comprehend it (Petty & Cacioppo, 1986).

To the extent that people are Bayesian updaters, the effects of differing “priors” should be attenuated, but several decades of research suggests that people often fail to update in a Bayesian fashion, though this is a matter of some controversy (see Gilovich, Griffin, & Kahneman, 2002). In many settings, inductive updating is better described by a weighting averaging rule. If people anchor on their priors and adjust insufficiently, they may well fail to converge on a consensus viewpoint.

Differing standards of proof will also discourage consensus on the evidence. For example, several studies indicate that the kind of biased assimilation effect documented by Lord et al. is largely mediated by more stringent processing of evidence supporting views contrary to one’s own. Ditto and Lopez (1992) found that students were significantly more likely to scrutinize a medical test when they tested positive for a potentially dangerous (fictitious) enzyme; they were also more than twice as likely to retest themselves. These reactions might appear to be normatively reasonable, but Ditto and Lopez also found that relative to students testing negative, students testing positive perceived the disease as less serious and more common. Similarly, Edwards and Smith (1996) find support for a “disconfirmation bias,” in which evidence inconsistent with the judge’s prior beliefs was scrutinized more extensively.

There are many literatures where evidence continues to accumulate on *both* sides of an empirical question. For example, evidence continues to accumulate on the question of whether marijuana is a “gateway” to hard drug use; almost every year some new studies present evidence that the association is causal, while other new studies suggest it is spurious (MacCoun & Reuter, 2001). Even putting aside the blatant partisanship and dubious logic that characterizes much of the debate, the truth fails to come into focus because strong inferential methods (like randomized experiments) are ethically precluded. In other cases, facts that appear contradictory from an adversarial perspective are not. For example, the Dutch decision to stop penalizing marijuana possession appears to have no effect on levels of marijuana prevalence, but their decision to allow coffeeshops to sell marijuana has probably increased its use (MacCoun & Reuter, 1997). These statements only seem contradictory when one bifurcates the debate into “Dutch policy is good” vs. “Dutch policy is bad.”

Finally, accumulating evidence sometimes shifts the terms of debate without bringing about resolution or consensus. This happens when ostensibly factual disagreements are a smokescreen for deeper differences in values. Specific deterrence was once the major argument of capital punishment supporters. As evidence accumulated questioning any marginal deterrent effect, the rationale shifted retribution. Most recently, a new rationale is cited – “closure” for victims’ families (see Zimring, 2001). There has been a similar evolution of stated rationales for the ban on gay and lesbian military service (MacCoun, 1996). First, gays were too effeminate to be soldiers, then they were a “security risk,” and in the 1990s, they were a threat to “unit cohesion.”

PROMOTING “HETEROGENEOUS INQUISITORIALISM”

Our system for introducing science and empiricism into policy discourse is clearly an awkward muddle of inquisitorial and adversarial methods. It would be quixotic to simply call for a return to a pure inquisitorialism that probably never has and never will exist. Indeed, some forms of bias are defensible. There are ample normative grounds for accepting differing opinions about imperfect and limited research on complex, multifaceted issues. There is nothing inherently wrong with differing standards of proof, and nothing shameful about taking an advocacy role – provided that we are self-conscious about our standards and our stance, and make them explicit. Fostering hypothesis competition and a heterogeneity of views and methods can simultaneously serve the search for the truth and the search for the good.

At the same time, I have argued that a shift toward a more explicit, robust adversarialism could make things worse instead of better. Instead, perhaps we need to better articulate the boundary between adversarialism and what might be called “heterogeneous inquisitorialism”—a partnership of rigorous methodological standards, a willingness to tolerate uncertainty, and the encouragement of a diversity of hypotheses and perspectives (MacCoun, 1998).

In principle, this should happen through the traditional scientific quality control procedures: Peer review and replication. Unfortunately, the evidence for the effectiveness of these institutional safeguards is pretty depressing (see MacCoun, 1998). Cicchetti (1991) and Cole (1992) report dismally low interreferee reliabilities in psychology journals (in the .19 to .54 range), medical journals (.31 to .37), and the NSF grant reviewing process (.25 in economics, .32 in physics). But with internet technologies, it is surely feasible to improve peer reviewing

practices. NSF and other funders are now using web-based peer-review portals. It is now logistically feasible to ask reviewers to review the theory section of a paper before receiving the method section, and the method section before they receive the results.

Exact replications are fairly rare (Bornstein, 1990), in part because editors and reviewers are biased against publishing replications (Neuliep & Crandall, 1990, 1993). An important development in this respect has been the dramatic growth of meta-analysis, the statistical aggregation of results across studies (e.g., Cooper & Hedges, 1994; Schmidt, 1992). Meta-analyses also have the benefit of being fairly explicit and transparent, and they've led to new standards for literature reviewing that seem likely to attenuate the citation biases that plague traditional reviews (e.g., Greenwald & Schuh, 1994).

Government agencies and private foundations need to actively promote and encourage greater use of randomized experiments, especially those conducted by independent investigators. A particularly encouraging development in this regard is the Campbell Collaboration, which is modeled on the influential Cochrane Collaboration for the dissemination of clinical trial evidence in medicine.¹¹

But we also need to take institutional steps to promote and encourage analytic methods that promote debiasing.¹² The key here is to discourage what might be called hypothesis promotion (testing single hypotheses and putting forth evidence on their behalf) and encourage investigators to engage in within-study hypothesis competition. Recall that devil's advocacy and the "consider the opposite" technique are among the few effective debiasing methods in the laboratory. A methodological stance that is similar in spirit was advocated by Platt (1964), who suggested that rapidly advancing research programs tend to employ a *strong inference* strategy, where the researcher designs tests of an array of plausible competitors rather than a single

¹¹ "The international Campbell Collaboration (C2) is a non-profit organization that aims to help people make well-informed decisions about the effects of interventions in the social, behavioral and educational arenas. C2's objectives are to prepare, maintain and disseminate systematic reviews of studies of interventions. We acquire and promote access to information about trials of interventions. C2 builds summaries and electronic brochures of reviews and reports of trials for policy makers, practitioners, researchers and the public." See <http://www.campbellcollaboration.org/>

¹² Glanz (2000) reports that particular physicists have developed a new procedure for reducing bias by blinding their analyses; a computer algorithm adds an unknown "offset" constant to their data, which is only removed when the analysis is complete. This might work well when competing theories make strong point predictions, but such is rarely the case in policy analysis. For us, the equivalent might be to add unknown "offsets" to each entry in our covariance matrices prior to multivariate analysis.

avored hypothesis. Greenwald, Pratkanis, Leippe, and Baumgardner (1986) applaud Platt's intent but suggest that his strategy is rooted in a naïve faith in falsificationism. Instead, they recommend a strategy of *condition seeking*, in which a researcher deliberately attempts to “discover which, of the many conditions that were confounded together in procedures that have obtained a finding, are indeed necessary or sufficient” (p. 223). For example, in Anderson and Anderson's (1996) “destructive testing” approach, the investigator tests alternative model specifications to identify what it takes to make an effect go away.

Discouraging bias doesn't necessarily invariably mean making our methods more stringent. In at least one sense, we might discourage mischief by being *less* stringent. One of the most expressions of investigator bias involves attempts to “cross the .05 threshold” of statistical significance via analytic “fishing expeditions,” dropping outliers, and so on. But as Cohen (1994) and others have pointed out, almost any comparison will be statistically significant given a large enough sample, yet levels of statistical power are scandalously low in the social sciences – a problem further compounded by the inherent noisiness of social science measurement. This is particularly a problem in field research, where there are logistical, economic, ethical, and political obstacles make it difficult to increase sample sizes. Thus while we obsess about false positives (Type I statistical errors), we arguably have a more serious problem with false negatives (Type II statistical errors; Cohen, 1994). But the growing influence of meta-analysis ought to diminish the value of null hypothesis testing. Meta-analysis helps solve the power problem, provided enough studies are found. It encourages lots of small studies (improving the heterogeneity and robustness of our research basis) rather than a handful of mega-studies. It doesn't solve the false-positive problem so much as make it moot – it provides robust estimation of the actual effect size across studies. Attention shifts to the magnitude of effects, their moderating conditions, and their *substantive* significance.

Finally, a more heterogeneous inquisitorialism would welcome investigators from the full spectrum of social categories and political beliefs (Redding, 2001; Tetlock, 1994), while at the same time encouraging greater clarity and candor about the values that motivate us, the facts that run counter to our views, and the role we are playing in any given setting (scientist vs. expert advocate). It may well be impossible, as philosophers and post-modernists insist, to separate facts from values, but we shouldn't accept this as a license for unfettered bias cloaked as “expertise.”

REFERENCES

- Arkes HR. 1991. Costs and benefits of judgment errors: Implications for debiasing. *Psy. Rev.* 110:486-498.
- Benford, G. (1980). *Timescape*.
- Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7-42.
- Cicchetti DV. 1991. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Beh. Brain Sci.* 14:119-186.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003
- Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26, 1151-1164.
- Cole S. 1992. *Making Science: Between Nature and Society*. Cambridge, MA: Harvard.
- Cummings KM, Russell S, Gingrass A, Davis R. 1991. What scientists funded by the tobacco industry believe about the hazards of cigarette smoking. *Am. J. Pub. Health* 81:894-896.
- DeVries, F., Mannen, K., Comerio, M., Ellwood, J., & MacCoun, R. (2001, 7 February). *Design faultlines: A study of seismic decisionmaking in universities*. Presented at the Annual meeting of the Earthquake Engineering Research Institute, Monterey, California.
- Ditto PH, Lopez DF. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *J. Pers. Soc. Psychol.* 63:568-584.
- Epstein, S. (1996). *Impure science: AIDS, activism, and the politics of knowledge*. Berkeley: Univ. Calif. Press.
- Faigman, D. L. (1999). *Legal alchemy: The use and misuse of science in the law*. Freeman.
- Fischhoff B. 1990. Psychology and public policy: Tool or toolmaker? *Am. Psychol.* 45:647-653.
- Fischhoff B, Beyth-Marom R. 1983. Hypothesis evaluation from a Bayesian perspective. *Psychol. Rev.* 90:239-260.
- Fiske ST. 1989. Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In *Unintended Thought*, ed. JS Uleman, JA Bargh, pp. 253-283. NY: Guilford Press.

- Fraser, S., (ed.) (1995). *The Bell Curve wars: Race, intelligence, and the future of America*. NY: Basic Books.
- Fuchs S, Westervelt SD. 1996. Fraud and trust in science. *Perspectives in Biology and Medicine* 39:248-270.
- Gaffan, E. A., Tsaousis, J., & Kemp-Wheeler, S. M. (1995). Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting & Clinical Psychology*. 1995 Dec Vol 63(6) 966-98.
- Gelbspan R. 1997. *The Heat Is On: The High Stakes Battle Over Earth's Threatened Climate*. New York: Addison Wesley.
- Gilovich, T., Griffin, D.W., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- Gholson B, Barker P. 1985. Kuhn, Lakatos, and Laudan: Applications in the history of physics and psychology. *Am. Psychol.* 40:755-769.
- Glantz SA. 1996. *The Cigarette Papers*. Berkeley: University of California Press.
- Glanz, J. (2000, 8 August). New tactic in physics: Hiding the answer. *New York Times*, D1, D4.
- Gorman, D. M. (1998). The irrelevance of evidence in the development of school-based drug prevention policy, 1986-1996. *Evaluation Review*, 22, 118-146.
- Gorman, D. M. (2002). The 'science' of drug and alcohol prevention: The case of the randomized trial of the Life Skills Training program. *International Journal of Drug Policy*, 13, 21-26.
- Gorman, D. M. (2003). Prevention programs and scientific nonsense. *Policy Review*, February & March, 65-75.
- Greenwald AG, Pratkanis AR, Leippe MR, Baumgardner MH. 1986. Under what conditions does theory obstruct research progress? *Psychol. Rev.* 93:216-229.
- Gross PR, Levitt N. 1994. *Higher Superstition: The Academic Left and Its Quarrels with Science*. Baltimore: Johns Hopkins.
- Hammond KR. 1996. *Human Judgement and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. NY: Oxford University Press.
- Hammond KR, Harvey LO, Hastie R. 1992. Making better use of scientific knowledge: Separating truth from justice. *Psychol. Sci.* 3:80-87.
- Hilts, P. J. (2000, September 1). Company tried to bar report that HIV vaccine failed. *The New York Times*.

- Horowitz, J. L., MacCoun, R. J., & Manski, C. F. (2002). Response to comments regarding the National Research Council report. *Addiction*, 97, 663-665.
- Howson C, Urbach P. 1993. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court. 2nd ed.
- Ioannidis, J. P., & Lau, J. (2001). Evolution of treatment effects over time: Empirical insight from recursive cumulative metaanalyses. *Proceedings of the National Academy of Sciences*, 98, 831-836.
- Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1-24.
- Kahneman D, Slovic P, Tversky A. eds. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.
- Kalven H, Zeisel H. 1966. *The American Jury*. Boston: Little, Brown.
- Keltner D, Robinson RJ. 1996. Extremism, power, and the imagined basis of social conflict. *Current Directions in Psychol. Sci.* 5:101-105.
- Kerr, N., MacCoun, R. J., & Kramer, G. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review*, 103, 687-719.
- Klayman J, Ha YW. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* 94:211-228.
- Koehler JJ. 1993. The influence of prior beliefs on scientific judgments of evidence quality. *Org. Behav. & Human Dec. Proc.* 56:28-55.
- Kruglanski AW. 1989. *Lay Epistemics and Human Knowledge: Cognitive and Motivational Bases*. NY: Plenum.
- Kruglanski AW, Webster DM. 1996. Motivated closing of the mind: "Seizing" and "Freezing". *Psychol. Rev.* 103:263-283.
- Kunda Z. 1990. The case for motivated reasoning. *Psychol. Bul.* 108:480-498. Kruglanski
- Laudan L. 1990. *Science and Relativism: Controversies in the Philosophy of Science*. Chicago: Univ. Chicago Press.
- Laughlin PR. 1996. Group decision making and collective induction. In *Understanding Group Behavior: Volume 1: Consensual Action by Small Groups*, ed. E Witte, JH Davis, pp. 61-80. Mahwah, NJ: Erlbaum.
- Lerner, J.S. & Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125, 255-275.

- LeVay, S. (1996). *Queer science: The use and abuse of research into homosexuality*. Cambridge, MA: MIT Press.
- Liberman A, Chaiken S. 1992. Defensive processing of personally relevant health messages. *Pers. Soc. Psychol. Bul.* 18:669-679.
- Lord CG, Lepper MR, Preston E. 1985. Considering the opposite: A corrective strategy for social judgment. *J. Pers. Soc. Psychol.* 47:1231-1243.
- Lord CG, Ross L, Lepper MR. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37:2098-2109.
- Lorge I., Solomon H. 1955. Two models of group behavior in the solution of Eureka-type problems. *Psychometrika* 20:139-148.
- MacCoun, R. J. (1989). Experimental research on jury decision making. *Science*, 244, 1046-1050.
- MacCoun, R. J. (1995). Review of K. R. Foster, D. E. Bernstein, & P. W. Huber (Eds.), *Phantom risk: Scientific inference and the law* (MIT Press, 1993). *Journal of Policy Analysis and Management*, 14, 168-171.
- MacCoun, R. J. (1996). Sexual orientation and military cohesion: A critical review of the evidence. In G. M. Herek, J. B. Jobe, & R. Carney (Eds.), *Out in force: Sexual orientation and the military*. Chicago: University of Chicago Press.
- MacCoun, R. (1998). Biases in the interpretation and use of research results, *Annual Review of Psychology*, 49, 259-287.
- MacCoun, R. J. (2001). American distortion of Dutch drug statistics. *Society*, 38, 23-26.
- MacCoun, R. (2002). Comparing micro and macro rationality. In M. V. Rajeev Gowda and Jeffrey Fox (Eds.), *Judgments, decisions, and public policy* (pp. 116-137). Cambridge University Press.
- MacCoun, R. (2003). Review of Shadish, Cook, & Campbell's *Experimental and quasi-experimental designs for generalized causal inference* (book review). *Journal of Policy Analysis and Management*, 22, 330-332.
- MacCoun, R., & Paletz, S. (2004). California survey of perceived bias in public policy research findings (manuscript in preparation). University of California at Berkeley.
- MacCoun, R., & Reuter, P. (1997). Interpreting Dutch cannabis policy: Reasoning by analogy in the legalization debate. *Science*, 278, 47-52.
- MacCoun, R., & Reuter, P. (2001). *Drug war heresies: Learning from other vices, times, and places*. Cambridge University Press. [Details and reviews at socrates.berkeley.edu/~maccoun/DWH.html]

- MacCoun, R. J., Lind, E. A., & Tyler, T. R. (1992). Alternative dispute resolution in trial and appellate courts. In D. K. Kagehiro & W. S. Laufer (Eds.), *The handbook of psychology and law* (pp. 95-118). New York: Springer Verlag.
- Mayo DG. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
- McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice-Hall.
- McGuire WJ. 1983. A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. *Adv. Exp. Soc. Psychol.* 16:1-47.
- Merton RK. 1973. *The Sociology of Science*. Chicago: University of Chicago Press.
- Metcalf, C. E. (1998). Research ownership, communication of results, and threats to objectivity in client-driven research. *Journal of Policy Analysis and Management*, 17, 153-163.
- Moskowitz JM. 1993. Why reports of outcome evaluations are often biased or uninterpretable: Examples from evaluations of drug abuse prevention programs. *Eval. & Planning* 16:1-9.
- Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, 27, 1450-1462.
- Munro GD, Ditto PH. 1997. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Pers. Soc. Psychol. Bul.* 23:636-653.
- Murray, B. (2002, February). Research fraud needn't happen at all. *APA Monitor*, 27-28.
- Neuliep JW, Crandall R. 1990. Editorial bias against replication research. *J. Soc. Behav. Pers.* 5:85-90.
- Neuliep JW, Crandall R. 1993. Reviewer bias against replication research. *J. Soc. Behav. & Pers.* 8:21-29.
- Pennington N, Hastie R. 1993. The story model for juror decision making. In *Inside the Juror: The Psychology of Juror Decision Making*, ed. R Hastie, pp. 192-221. NY: Cambridge University Press.
- Platt JR. 1964. Strong inference. *Science* 146:347-353.
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28, 369-381.
- Pyszczynski T, Greenberg J. 1987. Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Adv. in Exp. Soc. Psych.* 20:297-340.

- Redding, R. E. (2001). Sociopolitical diversity in psychology. *American Psychologist*, 56, 205-215.
- Revsin, A. C. (2003, August 5). Politics reasserts itself in the debate over climate change and its hazards. *The New York Times*, D2.
- Rips, L. J., Brem, S. K., & Bailenson, J. (1999). Reasoning dialogues. *Current Directions in Psychological Science*, 8, 172-177.
- Robinson, L. A., Berman, J. S. & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin*, 108, 30-49.
- Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, 68, 404-417.
- Rosenthal R. 1994. Science and ethics in conducting, analyzing, and reporting psychological research. *Psychol. Sci.* 5:127-134.
- Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. New York: McGraw-Hill.
- Schwenk CR. 1990. Effects of devil's advocacy and dialectical inquiry on decision making: A meta-analysis. *Org. Behav. & Hum. Dec. Proc.* 47: 161-176.
- Shadish, William R.; Montgomery, Linda M.; Wilson, Paul; Wilson, Mary R.; et al Effects of family and marital psychotherapies: A meta-analysis. *Journal of Consulting & Clinical Psychology*. 1993 Dec Vol 61(6) 992-1002
- Schum D.A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference. *Law & Soc. Rev.* 17:105-151.
- Shadish WR. 1995. Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation & Program Planning* 18:63-75.
- Tetlock PE. 1994. Political psychology or politicized psychology: Is the road to scientific hell paved with good moral intentions? *Poli. Psychol.* 15:509-529.
- Tetlock PE, Levi A. 1982. Attribution bias: On the inconclusiveness of the cognition-motivation debate. *J. Exp. Soc. Psychol.* 18:68-88.
- Tetlock, P. E., Peterson, R. & Lerner, J. (1996). Revising the value pluralism model: Incorporating social content and context postulates. In C. Seligman, J. Olson, & M. Zanna (Eds.), *Ontario symposium on social and personality psychology: Values*. Hillsdale, NJ: Erlbaum.
- Thagard P. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton Univ. Press.

- Thibaut J, Walker L. 1978. A theory of procedure. *Calif. Law Rev.* 26:1271-1289.
- Vallone RP, Ross L, Lepper MR. 1985. The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *J. Pers. Soc. Psychol.* 49:577-585.
- Wegener, D. T., Kerr, N. L., Fleming, M. A., & Petty, R. E. (2000). Flexible corrections of juror judgments: Implications for jury instructions. *Psychology, Public Policy, and Law*, 6, 629-654.
- Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4/e). Boston: McGraw-Hill.
- Weimer DL, Vining AR. 1992. *Policy Analysis: Concepts and Practice*. 2nd Ed. Englewood Cliffs, NJ: Prentice Hall.
- Wilson TD, Brekke N. 1994. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychol. Bull.* 116:117-142.
- Woodward J, Goodstein D. 1996. Conduct, misconduct and the structure of science. *Am. Scientist* 84: 479-490.
- Zimring, F. (2003). *The contradictions of American capital punishment*. Oxford University Press.
- Zuckerman H. 1988. The sociology of science. In *Handbook of Sociology*, ed. NJ Smelser, pp. 511-574. Beverly Hills, CA: Sage.