

# Clinical attrition due to biased preclinical assessments of potential efficacy

Mark D. Lindner \*

271 Nob Hill Road, Cheshire, CT 06410, United States

## Abstract

Unless it is carefully controlled, bias often distorts the results of clinical trials, usually exaggerating the magnitude of true efficacy. For that reason, procedures to limit bias have been mandated by the FDA when assessing efficacy in clinical trials. The present review shows that the effects of bias in preclinical studies are at least as large as in clinical trials, and since bias is not usually controlled in preclinical proof of concept studies, compounds that actually have little or no therapeutic potential may often be advanced into clinical trials. This possibility is supported by the fact that lack of efficacy is the single biggest reason why compounds fail in the clinic. The shift to target-based discovery during the last 10–15 years may have further increased the effects of bias on preclinical assessments of potential efficacy, and contributed to the continuing decline in clinical success rates. Procedures are available to control for bias during preclinical assessments of potential efficacy, and their use could dramatically increase clinical success rates and substantially reduce the costs of drug discovery and development.

© 2007 Elsevier Inc. All rights reserved.

*Keywords:* Bias; Expectancy; Drug discovery; Productivity; Efficacy; Preclinical assessments

## Contents

1.	Introduction . . . . .	149
2.	The prevalence and magnitude of experimenter bias in clinical trials. . . . .	149
3.	The possibility that bias might affect preclinical proof of concept studies . . . . .	152
4.	Replication studies to address potential bias in preclinical proof of concept experiments . . . . .	154
5.	Why are scientists vulnerable to bias; why can't they be objective? . . . . .	157
6.	How bias affects experimental results. . . . .	158
6.1.	Effects of bias on attention and perception . . . . .	158
6.2.	Conditioning of bias . . . . .	158
6.3.	The demands of authority and the research setting. . . . .	159
6.4.	Bias in experimental design . . . . .	160
6.5.	Effects of bias on errors in recording and managing data . . . . .	161
6.6.	Bias in data analyses . . . . .	161
6.7.	Interpretation bias. . . . .	162
6.8.	Publication bias. . . . .	163
6.9.	Citation bias . . . . .	163
6.10.	Bias in selecting experts for feedback . . . . .	163
6.11.	Bias in the decision-making process . . . . .	164
7.	Bias is not limited to inexperienced and unethical scientists . . . . .	165
7.1.	Experimenter bias is not under conscious control . . . . .	165
7.2.	Even successful, experienced and productive scientists are vulnerable to bias . . . . .	165
8.	Target-based discovery may increase bias in preclinical assessments of potential efficacy . . . . .	166

\* Tel.: 203 804 8957.

E-mail address: [MDLindner@cox.net](mailto:MDLindner@cox.net).

9.	Controlling bias in assessments of potential efficacy could increase the productivity of drug discovery efforts . . . . .	167
10.	Limiting experimenter bias in preclinical assessments of potential efficacy . . . . .	168
10.1.	Creating and maintaining a culture that controls for potential bias . . . . .	169
10.2.	Procedures to limit bias . . . . .	169
10.3.	Objective decision-making . . . . .	170
11.	Summary and conclusions . . . . .	171
	References . . . . .	171

## 1. Introduction

Preclinical proof of concept studies are conducted to assess the therapeutic potential of novel compounds, and the results of these studies are used in the decision-making process to determine which compounds should be advanced into clinical trials. Despite evidence of potential efficacy in preclinical models, compounds often fail in the clinic due to lack of efficacy. In fact, the single biggest reason why drugs fail in clinical trials is lack of efficacy (Prentis et al., 1988; Kennedy, 1997; DiMasi, 2001; Schuster et al., 2005). This phenomenon is often attributed to the limitations in the predictive validity of the preclinical models, but it may also be due to the misleading effects of bias throughout the process of assessing potential efficacy in preclinical models.

In clinical trials, the hopes, beliefs, and expectations—together referred to as the biases of the patients and investigators can affect the results, usually to exaggerate the therapeutic effects of the treatment being evaluated. Partly for that reason, procedures intended to reduce or limit the effects of experimenter bias have been mandated by the FDA, the regulatory agency that has final authority over product approval. Since these procedures were mandated, results have accumulated that allow comparisons between studies that were very well controlled and studies that were not as rigorously controlled. These comparisons show that the phenomenon of bias is so common, and its effects are so profound, that unless it is carefully controlled, treatments that actually have little or no therapeutic potential often seem to produce fairly substantial beneficial effects.

In contrast to the thorough precautions taken to guard against experimenter bias in clinical trials, there is often little or no concern that experimenter bias might affect preclinical assessments of potential efficacy. This may not be surprising since clinical trials are characterized by extensive human interactions and subjective assessments which justify the rigorous control procedures mandated by regulatory agencies. It may be assumed that preclinical research is based on more quantitative objective measures that are fairly immune to subjective bias. In addition, in contrast to clinical trials, replications of preclinical proof of concept studies are fairly inexpensive and easy to conduct. For that reason it may be assumed that replication studies are routinely conducted and can be relied upon to efficiently validate the results of preclinical studies and rule out any concerns about experimenter bias.

The objectives of the present paper are (1) to review the literature on the prevalence and magnitude of experimenter bias

in clinical trials, (2) to assess the possibility that bias might affect preclinical proof of concept studies, (3) to determine if numerous reports of potential efficacy adequately address any concerns about potential bias, (4) to review why scientists and others might be vulnerable to bias, (5) to examine the mechanisms that mediate experimenter bias, (6) to distinguish bias from deceit and fraud and show that bias is not limited to weak or inexperienced scientists, (7) to discuss why the shift to target-based discovery might increase the effects of experimenter bias on preclinical assessments of potential efficacy, and (8) to discuss procedures that could help limit the effects of experimenter bias and the impact these procedures could have on the efficiency and productivity of drug discovery and development. For the purposes of this paper, significant or positive studies are those that reported effects that were statistically significant at  $P \geq 0.05$ .

## 2. The prevalence and magnitude of experimenter bias in clinical trials

As long ago as 1784, Benjamin Franklin literally blindfolded subjects to show that the therapeutic effects produced by mesmerism on a range of disorders were actually due to “imagination,” “illusions created by the mind,” they were not “real.” In 1799, a blind technique was used again to show that a device that supposedly conducted pathogenic fluid away from the body actually relieved pain only through the power of suggestion. In 1834, patients were treated with either bread pills (inert control) or pills with homeopathic solutions (extremely dilute solutions) to show that homeopathy was no more active than inert substances. In 1895, still more than 100 years ago, Wacław Sobieranski, a Polish pharmacologist, reported that he routinely used placebos to control for the effects of “autosuggestion” which “played a large role in healing,” and that “much of the healing properties of chemical medicine is in reality attributable to autosuggestion.” He argued that “in order to keep psychological influence to a minimum, all subjects needed to be unaware of the experimental substance” and he criticized his colleagues for neglecting the power of suggestion (see Kaptchuk, 1998, for an excellent historical review).

In the early 1900s several clinical trials were conducted with active treatments compared to placebo in which the physicians and/or the patients were blind to treatment allocation in order to prevent “bias” (Kaptchuk, 1998). For example, in the 1930s Harry Gold and his colleagues noted their concern that the clinical investigator, knowing whether a patient had received

the active treatment or placebo, might influence the patients' answers to questions about the effects of their treatment. In order to reduce the risk that bias might affect the results, neither the clinical investigators nor the patients were informed as to whether they were receiving the active treatment or the placebo until after the data had been collected. Gold initially referred to this as "the blind test" and later as "the double blind" procedure (Gold et al., 1937; Greiner et al., 1950). Using this double blind procedure, Gold and his colleagues showed that treatments previously believed to be effective in 80–90% of patients were actually no more effective than placebo, and they attributed the differences in the results between these studies to the effects of bias. These accounts demonstrate that the pervasive effects of experimenter bias in clinical trials and the need to control or limit potential bias by concealing treatment allocation were widely accepted as early as the 1950s (Hill, 1952; Gold, 1954; Modell & Houde, 1958).

In 1962 the Kefauver-Harris Amendments gave the FDA responsibility for verifying the therapeutic efficacy of novel treatments; and in 1970, when the FDA announced their criteria for determining clinical efficacy, it was no surprise that these criteria included numerous procedures to control for potential bias. These procedures limit the potential effects of bias on treatment allocation, data collection and measurement, and statistical analyses (Edwards, 1970). Since these rigorous controls were mandated to limit bias in clinical efficacy trials, literature has accumulated that shows clear differences in results based on the "quality" of the experimental design (i.e., the degree to which experimenter bias has been controlled; Green & Byar, 1984; Juni et al., 2001; see Table 1).

One of the simplest, weakest, lowest quality experimental designs, which is vulnerable to many types of bias, is treating patients and observing their responses without including a control group in the study. In these studies treatments are considered effective if they produce robust response rates well above what is expected by experts in the field. One of the problems with these uncontrolled studies is that they often underestimate the response rates of untreated patients or patients treated with a placebo. Based on Beecher's original review of placebo effects it was often assumed that placebo response rates were very stable and consistent at 35%. However, placebo response rates actually vary over a wide range from one study to another, from as low as 10% to as high as 90%. Treatment

effects can also vary over an equally wide range from one study to the next, and the size of the placebo effect is correlated with the size of the effect for active treatments: studies with larger treatment effects tend to have larger placebo effects as well, which makes it inappropriate to assume that a treatment is effective just because it produced a high response rate (Beecher, 1955; Kissin et al., 1968; Moerman, 1983; Moertel, 1984).

As mentioned above, Gold and his colleagues showed that treatments previously believed to be effective in 80–90% of patients based on the results of studies without control groups were actually no more effective than placebo (Hill, 1952; Gold, 1954). A review of psychiatric clinical studies also reported that 83% of studies without control groups concluded that the treatment examined was efficacious, while only 25% of controlled studies of the same treatments detected efficacious effects (Foulds, 1958). Other reviews of enthusiastically endorsed treatments reported an average improvement in 70–82% of patients, but these treatments were no longer considered effective and were abandoned based on the results of placebo-controlled trials (Roberts et al., 1993). And finally, a review of treatments for alcoholism reported efficacy in 94.5% of uncontrolled studies, while controlled studies of the same treatments reported efficacy only 6% of the time (Viamontes, 1972).

Slightly better than assessing treatment effects relative to the response rates expected by experts in the field is comparing patient responses to treatment with recovery rates previously recorded in non-treated patients, known as historical controls. Unfortunately, response rates of non-treated historical controls also vary over time, often improving over the years as life expectancies and quality of life improves. For this reason, and because non-treated patients do not enjoy the benefits of placebo effects, non-treated historical controls often over-estimate disease severity and progression relative to placebo-treated concurrent control groups (Diehl & Perry, 1986; Moertel, 1984). This explains why 79% of clinical trials produced beneficial results compared to historical controls, but the same treatments only produced beneficial effects in 20% of trials when controls were recruited concurrently and included in the study for direct comparisons (Sacks et al., 1982).

The use of a concurrent control group is more appropriate than using historical controls, but blinding procedures to conceal treatment allocation add yet another level of control over potential bias. As one example, in a clinical trial of treatments for

Table 1  
Degree of control over potential bias: effects on apparent efficacy

Control procedure	Outcome measure	More well controlled	Less well controlled	Study
With vs. without control group	% of studies reporting treatment efficacy	25	83	Foulds (1958)
		6	94	Viamontes (1972)
Concurrent vs. historical control group	% of studies reporting treatment efficacy	20	79	Sacks et al. (1982)
Blinded vs. not blinded	% of patients meeting criteria for significant improvement	0–19	27–63	Epstein (1996)
Randomized vs. non-randomized treatment allocation	% of Treatment-reduced mortality rates	20	53	Chalmers et al. (1977)
		8.8	58.1	Chalmers et al. (1983)
Active vs. inert placebo	% of studies reporting treatment efficacy	12	89	Carroll et al. (2001)
	% of studies reporting treatment efficacy	14	59	Shapiro and Shapiro (1997)

multiple sclerosis, clinical assessments made by a non-blinded neurologist produced statistically significant treatment effects, but assessments made by a neurologist blind to treatment condition failed to detect statistically significant effects of the treatments compared to placebo-treated controls (Noseworthy et al., 1994). A survey of non-blinded clinical trials of anti-CD4 monoclonal antibodies for rheumatoid arthritis also reported robust and statistically significant improvements, with 27–63% of patients meeting the criteria for significant improvement, while only 0–19% of patients in blinded clinical trials showed improvement. The treatment effects in these blinded trials were so much smaller that they were no longer statistically significant (Epstein, 1996).

Another control that further reduces potential bias, in addition to blinding the investigators and patients to treatment allocation during the study, is making sure that bias does not affect the initial allocation of patients into treatment groups. This is accomplished by using randomized treatment allocation, and by keeping the investigators blind to the treatment allocation. These procedures insure that patients that are more likely to improve are not preferentially allocated to the active treatment group. The effects of bias during treatment allocation are shown in surveys of clinical trials of anticoagulants for myocardial infarction: non-randomized trials produced an apparent 53% reduction in mortalities, compared to the placebo controls, while randomized trials of the same treatments detected only a 20% reduction in mortalities compared to placebo (Chalmers et al., 1977). In another survey, differences in case fatality rates between treatment and control groups were 8.8% in studies with blinded, randomized treatment allocation; 24.4% in studies with non-blinded randomized treatment allocation; and 58.1% in studies without randomized treatment allocation (Chalmers et al., 1983). Transcutaneous electrical nerve stimulation also produced significant analgesic effects compared to the placebo control in 89% (17/19) of non-randomized trials, but in only 12% (2/17) of randomized controlled trials (Carroll et al., 2001). Again, these results show that knowledge of treatment assignment leads to biased patient allocations which exaggerates the apparent efficacy of the treatments under study.

Even randomized, double-blind, placebo-controlled trials are vulnerable to experimenter bias. Another control for potential bias is the use of active placebos or current clinical standards, rather than inert placebos. Numerous studies have shown that active treatments often produce obvious adverse effects which suggest to the patient that they must be in the active treatment group; and this heightens the placebo effect for the patients actually receiving the active treatment beyond the placebo effect produced in patients taking an inert placebo (Fisher & Greenberg, 1993; Shapiro & Shapiro, 1997, p. 206).

The fact that side effects are sometimes taken as evidence of potential efficacy can be seen in a study of the anxiolytic meprobamate. There were no differences between placebo and drug unless the investigator suggested that specific side effects were indicative of drug efficacy. Under those conditions, the drug was clearly superior to the placebo (Fisher et al., 1964). In clinical trials with antidepressants, 59% of studies reported that the antidepressant was effective if the control was an inert or inactive placebo, but only 14% of the studies reported that the

antidepressant was effective if the control was an active placebo such as atropine, which produces dry mouth (Shapiro & Shapiro, 1997, p. 230). Finally, anti-depressants tested against inert placebos produced very large effects, but when compared to standard antidepressants both the novel treatment and the older standard treatment produced more modest effects, approximately one half to one quarter the size of the effects previously reported in the initial trials, which had used inert placebos (Greenberg et al., 1992). These results show that side effects suggest to the patients that they are receiving active treatment and thus exaggerate placebo effects.

In addition to requiring that efficacy be assessed in randomized, double-blind, placebo-controlled clinical trials, FDA-mandated control procedures also include the use of pre-defined and approved protocols with appropriate comparator treatments included in the experimental design, defined endpoints and planned statistical analyses, and predetermined criteria for decision-making. The overall effect in most studies that fail to control for experimenter bias is to exaggerate or overestimate the size of the treatment effects. For example, a review of 250 controlled trials concluded that the odds ratios were 20–40% larger in less well-controlled studies (Schulz et al., 1995). To put this into terms that might be more understandable, this difference in odds ratios would be produced if the true placebo response rate was 30%, the success rate in very well-controlled studies was only 33%, and the success rate with active treatments in less well-controlled studies was 50% (Fig. 1). In other words, in well-controlled clinical trials active treatments might produce, on average, a 10% increase in success rate above placebo; while poorly controlled clinical trials might produce, on average, a 67% increase in patient success rates above placebo.

The results of these studies make it clear that bias often produces robust effects on the results of clinical trials; and this

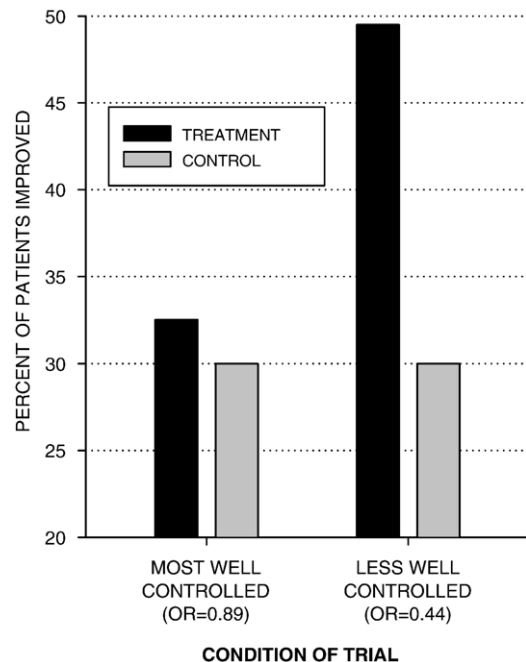


Fig. 1. Estimated differences in patient response rates in well-controlled versus less well-controlled clinical trials, based on odds ratios reported by Schulz et al. (1995).

phenomenon has been recognized as a significant problem that must be addressed when determining potential efficacy. Procedures to control for bias have been developed, and each level of control reduces bias further and further, up to the highest quality, most well-controlled experimental design. In these very high quality controlled studies the probability of detecting large, significant treatment effects is very low, but this is due to the fact that the effects of potential bias have been controlled, not because these studies are less sensitive to treatment effects. Regardless of whatever efficacy may be apparent in less well-controlled studies, the FDA usually bases their decision about potential efficacy entirely on the results of the highest quality experimental designs: large, randomized, double-blind, placebo-controlled trials, using pre-approved protocols, which most rigorously control for the potential effects of bias. This policy ensures that only those treatments that are truly efficacious get approved for use in patients.

### 3. The possibility that bias might affect preclinical proof of concept studies

It may be understandable that clinical trials are vulnerable to potential bias, since clinical studies are characterized by extensive human interactions and subjective assessments. For example, there is considerable variability and poor reliability between physicians with respect to their assessments of physical signs, interpretations of roentgenograms, electrocardiograms, electroencephalograms, and even physical measures of tumor size, in addition to variability in their diagnoses, treatment recommendations, and evaluations of the quality of care, all of which are therefore vulnerable to the effects of bias (Koran, 1975a,b; Moertel, 1984).

In contrast, it may be assumed that pre-clinical research is based on more quantitative, objective measures that are less vulnerable to experimenter bias. However, there is considerable evidence that experimenter bias is not limited to clinical trials, nor is it a problem that has emerged only recently. Bias has been recognized as a major concern in all areas of scientific research throughout the history of science. For example, it is well known that the scientific literature of the 16th and 17th centuries are full of experiments that were never conducted. Even some of the major scientific achievements by scientists such as Galileo and Pascal were based on experiments that could not have been conducted (Koyre, 1956, 1960). Galileo actually boasted that he did not need to verify his theories experimentally (Koyre, 1960).

In 1620, Francis Bacon emphasized the importance of cultivating the proper attitude of the mind, to be open and analytical, and to work to maintain objectivity. He felt that experimental results at that time were not being used to test but merely to illustrate the scientific doctrines being advanced, giving a misleading impression of strongly confirming them. He lamented that when an observation was inconsistent with a theory, the practice was to ignore it, cast doubt on it, exclude it as an exception, or explain it away by some “frivolous distinction,” but not to regard the theory itself as flawed. Bacon cautioned that investigators must work very hard to maintain objectivity and

avoid bias, especially when dealing with pet theories, “For man prefers to believe what he wants to be true” (Bacon, 1620, p. 59).

Concerns about bias continued into the early 1800s when Charles Babbage complained that scientists were often “trimming” and “cooking” their data, which meant that they were omitting or altering some of their original data in order to make their results fit more closely with their expectations (Babbage, 1830). Claude Bernard noted in the late 1800s that the greatest and most common stumbling block in research is accepting only those observations which support or confirm the experimenter’s hypothesis (Bernard, 1865, p. 23).

Even contemporary philosophers of science such as Thomas Kuhn have noted that data which are inconsistent with established theory is usually ignored or dismissed, often with little conscious awareness of its potential importance. He pointed out that only those with very open minds are able to perceive and recognize anomalous observations, and that significant discoveries are usually made by young investigators or those new to a field because most established investigators are so biased that they are unable to even perceive observations that are inconsistent with established theory (Kuhn, 1996).

Specifically with respect to preclinical models, classic studies published more than 40 years ago demonstrated that experimenter bias can affect the results of laboratory studies using non-human subjects. In an initial behavioral study in which experimenters tried to condition planaria to turn, it was noted that some experimenters were able to demonstrate conditioning but some experimenters were not (Fig. 2). In a follow-up study, planaria were given a 3-sec light cue followed by a 1-sec shock. Observers recorded whether the planaria contracted their bodies or turned their heads in response to the light, which were the dependent measures used to determine if the planaria were learning to associate the light with the shock. Increasing numbers of contractions and head turns in response to the light would be

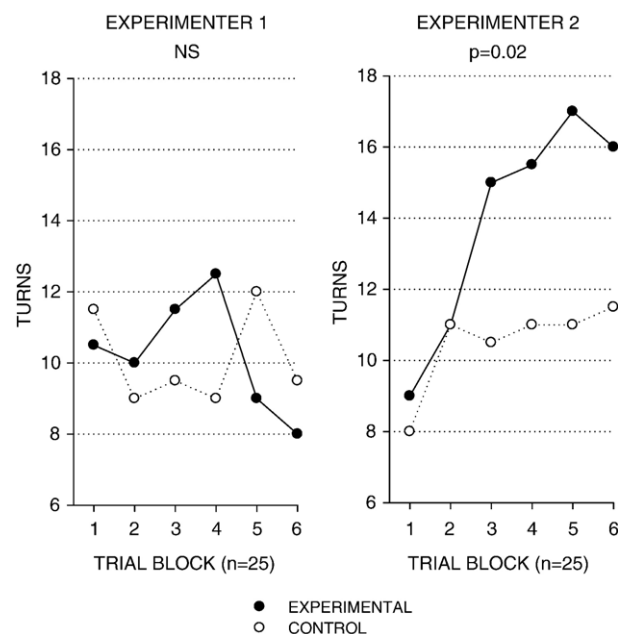


Fig. 2. Variability between experimenters in results of conditioning with planaria, from Rosenthal and Halas (1962) and Rosenthal (1966, pp. 7–11).

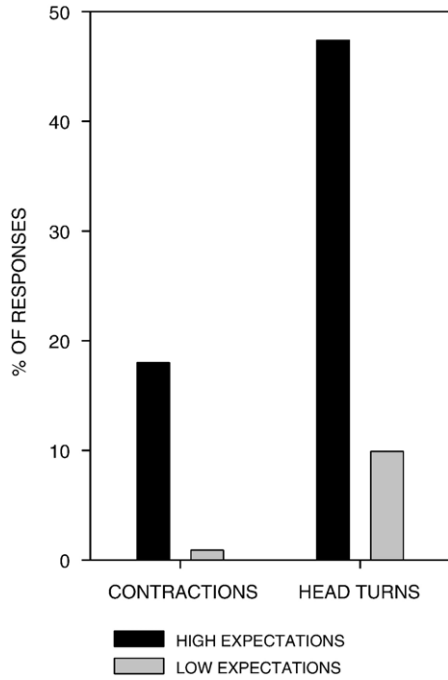


Fig. 3. Effects of experimenter expectations on results of conditioning with planaria, from Cordaro and Ison (1963).

evidence of classical conditioning. Observers that had been told to expect rapid conditioning recorded more responses to the light, 5 times as many head turns and 20 times as many contractions, compared to observers that had been told to expect little or no evidence of conditioning (Fig. 3).

In later studies, experimenters trained rats to run to the dark arm of a T-maze for a food reward. Rats were given 10 trials each day for 5 days. Experimenters that were told to expect their rats to acquire the task quickly did in fact record faster acquisition

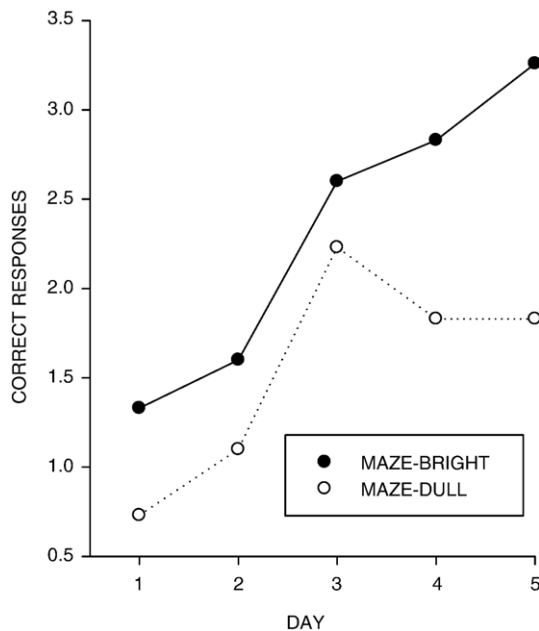


Fig. 4. Effects of experimenter expectations in rat T-maze acquisition, from Rosenthal and Fode (1963).

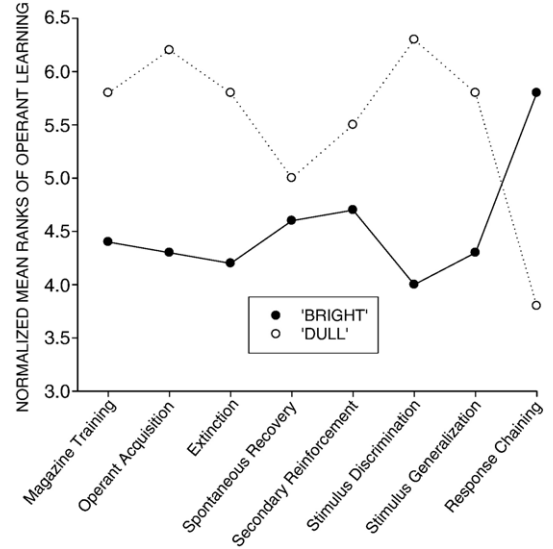


Fig. 5. Effects of experimenter expectations in a longitudinal study of operant learning in laboratory rats, from Rosenthal and Lawson (1963). Lower ranks indicate superior learning.

and better performance than experimenters that had been told to expect that their rats would be poor learners (Fig. 4). In another learning study, rats were trained to perform a series of tasks for a food reward. If the experimenter had been told that they were working with rats that had been bred for good performance in these tasks, they recorded better performance with their rats than if the experimenter had been told that their rats had been bred for poor performance in these tasks (Fig. 5). Burnham (1966, cf. Rosenthal, 1976, pp. 449–451) also reported a study in rats with

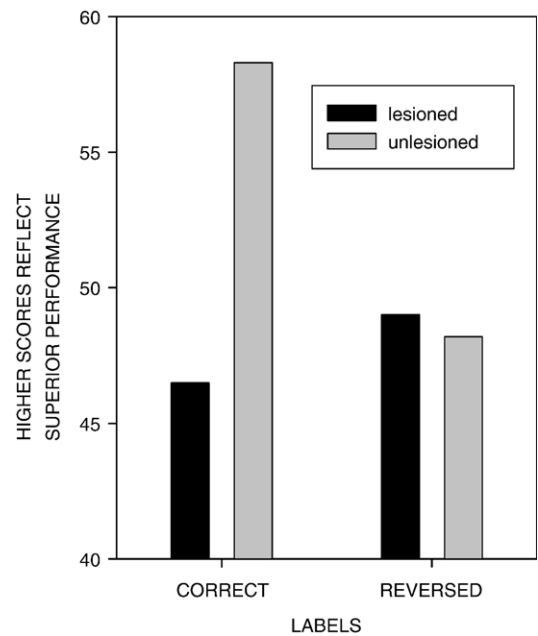


Fig. 6. Effects of experimenter expectations on T-maze discrimination in rats with brain lesions, from Burnham (1966, cf. Rosenthal, 1976, pp. 449–451). Higher scores reflect better performance. When rats were labeled correctly, the expected results were obtained. However, when the labels were reversed, no deficits were detected in the lesioned rats.

brain lesions versus non-lesioned controls. Half of each group was intentionally mislabeled, and these rats were then tested in a T-maze discrimination learning task. Impairments in this learning task were more dependent on the experimenter's expectations, based on the animal identification labels, than whether the rats actually had a brain lesion or not (Fig. 6).

The results of these and other studies showed that, consistent with the results of clinical trials, experimenter bias is both a general and robust phenomenon in experiments with animals (Rosenthal, 1963). An analyses of these experiments revealed that, surprisingly, the effects of experimenter bias were extremely large, and actually at least as robust and consistent in experiments with animals as in any of the research areas examined in humans, even including very subjective assessments such as perceptions of faces and ink blots (Fig. 7). All the animal studies included in this comparison were learning experiments, but there is no evidence that learning studies are more vulnerable to experimental bias than other kinds of preclinical models.

For example, experimenter bias can also affect research in the area of molecular biology and genetics. George Beadle, a major figure in the history of molecular biology who developed the 1 gene–1 enzyme hypothesis, stated, “As every experimenter in genetics knows, some classifications are difficult and may easily be unconsciously biased in favor of a preconceived hypothesis.” He recognized that some of his own observations were biased just in time to withdraw a manuscript that he had already submitted for publication, and he noted that many results reported in genetics are “too good” to be true (Beadle, 1967, p. 338). Even objective measures can be affected by bias. For example, standard laboratory procedures used to manually count blood cells were shown to require a degree of consistency that was not possible, given the equipment and procedures used, yet laboratory results typically conformed to these unrealistic requirements for consistency. This could only have occurred as a result of the bias of the laboratory technicians (Berkson et al., 1940).

While preclinical studies are at least as vulnerable to experimenter bias as clinical trials, preclinical proof of concept studies are rarely controlled for potential bias. For example, a survey of novel treatments for Alzheimer's disease found 11

compounds that were recently discontinued after moving into Phase III clinical trials. Fifty-five preclinical papers of potential efficacy had been published on these 11 compounds, and virtually 100% of these papers reported efficacy in terms of improving cognitive function, but none of them reported any procedures to control for potential bias (Table 2).

#### 4. Replication studies to address potential bias in preclinical proof of concept experiments

Since preclinical proof of concept studies are vulnerable to but do not protect against the effects of biases, they may exaggerate the potential efficacy of novel treatments. However, replications of preclinical proof of concept studies are not expensive and can be conducted fairly quickly, at least compared to most clinical trials, so it may be assumed that replication studies are routinely conducted and can be relied upon to efficiently validate the results of preclinical studies and address any concerns about potential experimenter bias. Among the first to establish the importance of replication was perhaps the German philosopher Immanuel Kant, who argued in *The Critique of Pure Reason* that before phenomena can be accepted as real and not just coincidences, events must occur repeatedly and consistently, according to well defined rules (Kant, 1787). Karl Popper emphasized and extended this point in *The Logic of Scientific Discovery*, where he wrote that, even phenomena that can be reproduced several times but can then no longer be reproduced, cannot be accepted as valid—they must continue to be regularly and consistently reproduced by anyone who replicates the procedures described in the original report (Popper, 1935, pp. 23–24).

According to the American sociologist Robert K. Merton, the scientific community widely accepts the requirement that experimental results must be reliably replicated in order to be accepted as valid: “... independently collected, systematic, and quantitative data supply the most demanding test of any scientific finding, ...even the most logical and reasonable theory remains in the stage of pure hypothesis until it is tested, and it is unhesitatingly rejected if the facts contradict it” (see Merton, 1973, pp. 163–164 and 270). This widely accepted need for

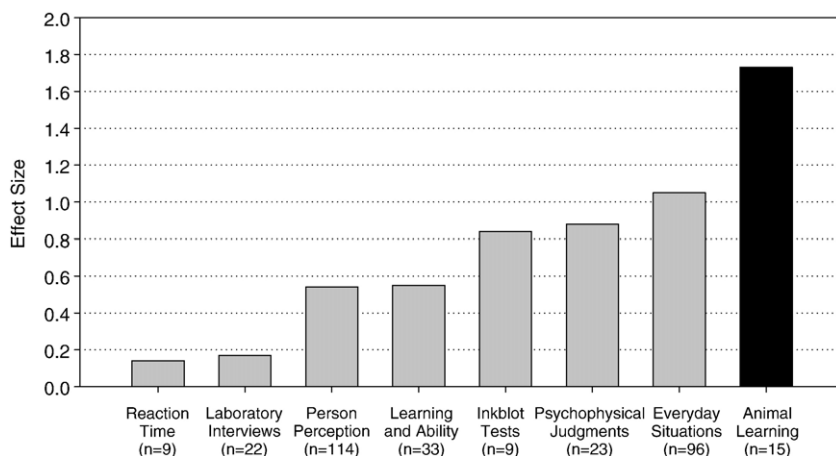


Fig. 7. Size, in standard deviations, of expectancy effects across eight research areas, from Rosenthal (1969), Rosenthal and Rubin (1978), and Robert Rosenthal, personal communication (4/28/2007).

Table 2  
Compounds for Alzheimer's disease and cognition disorders discontinued in phase III

Drug	Mechanism of action	Country tested	Reason discontinued?	No. preclinical efficacy pub's.	Control for bias
Adafenoxate (WON 150)	L-lactate dehydrogenase stimulants	Spain	Unknown	7	None
Ensaculin (Anseculin KA 672)	Undefined	Germany	Potential side effects	2	None
Eptastigmine (Heptylphosphostigmine, Heptylstigmin, L 693487, MF 201)	Acetylcholinesterase inhibitor	Italy, United Kingdom, USA	Aplastic anemia	7/8	None
Ipidacrine (Amiridin, Amiridine, NIK 247, Senita)	Acetylcholinesterase inhibitors; Potassium channel antagonists	Japan	Lack of efficacy	12	None
Lazabemide (RO 196327, Pakio, Tempium)	Antioxidants; Monoamine oxidase B inhibitors	Europe and Japan	Severe hepatotoxicity	0	N/A
Linopirdine (Aviva, DUP 996, Linopirine)	Acetylcholine release stimulants	Canada and USA	Lack of efficacy	8/10	None
Milameline (CI 979, Mirameline, PD 129409, RU 35926, Vivad)	Muscarinic receptor agonists	European Union and USA	Toxicity	3	None
ORG 2766 <sup>a</sup>	Adenylate cyclase stimulants	USA	Lack of efficacy	4	None
Suritazole (MDL 26479)	Benzodiazepine receptor inverse agonists	United Kingdom	Business decision	4	None
Xanomeline (LY 246708, NNC 110232, Memcor)	Muscarinic M1 and M4 receptor agonists	USA	Adverse effects	2	None
Zanapezil (TAK 147)	Acetylcholinesterase inhibitors	Japan	Lack of efficacy	3	None

English language preclinical publications assessing effects of compounds on cognitive function identified from search of PubMed.

<sup>a</sup> Only preclinical studies published before the results of the ORG 2766 clinical trials were published in 1985 were included in this survey.

consistent, reliable replication and confirmation is reflected throughout the literature by statements such as the following:

Replicability is the foundation of the value system of science (Zuckerman, 1977).

Science requires that a phenomenon be reliably produced in different laboratories for it to be accepted as genuine. Whoever claims to have discovered a phenomenon must describe in sufficient detail how it was produced so that other investigators, following similar steps, can reproduce it themselves. This requirement of replicability applies to all fields of science (Gilovich, 1991, pp. 168–169).

Replication studies are the cornerstone to any science and at the heart of the scientific method, the acceptance of any result as valid is contingent on demonstrating reliability. ... It is axiomatic that, for a scientific discipline [to be] cumulative, [it must] extensively utilize replication to test reliability, validity and generalizability (Campbell & Jackson, 1979).

Despite the widely accepted need to replicate results in order to verify their validity, replication studies are actually rarely conducted (Bornstein, 1990; Collins, 1992). Even in the highest ranking and most highly cited scientific journals, 55% of published studies are never even cited, much less replicated (Hamilton, 1990), and there are numerous examples of research results which were accepted as valid for some time until replications were finally conducted to reveal that the original findings could not be reproduced (Broad & Wade, 1982; Engler et al., 1987; Amir & Sharon, 1990; Bailey, 1991). In some cases investigators might not conduct replication studies because, based on the results of the statistical analyses, they are convinced that the original findings are reliable and valid. Findings that are extremely unlikely to have occurred by chance, based on statistical probabilities, are sometimes accepted as reliable and

valid without replication, but statisticians do not support this use of statistical analyses. Statistical analyses attempt to assess the future replicability of a particular phenomenon, but statisticians insist that the ultimate test of any significant finding is its repeatability (Fisher, 1971, p. 14; Keppel, 1982, p. 74).

The main reason why investigators do not conduct replication studies is that there is usually little incentive to do so. The scientific community rewards those who are first to report novel phenomena, and it is unlikely that any scientist will earn tenure or promotion by replicating other scientists' findings. There is certainly no reward for those who fail to reproduce statistically significant, novel phenomena. Reviewers and editors prefer to publish novel findings, and they feel that replications are a waste of time and journal space. In addition to the fact that there is little or no reward for those who conduct replication studies, the scientific community as a whole tends to treat anyone willing to conduct and report replications with disdain (Bornstein, 1990; Hendrick, 1990; Neuliep & Crandall, 1990, 1993), so it is not surprising that scientists usually choose not to publish or not to conduct replication studies in the first place (Mahoney, 1985, 1987).

When replication studies are conducted, experimenters can devote enormous amounts of time to them, even more time and effort than went into the original experiments (Collins, 1992), and if discordant results are produced, it can be very difficult to determine what might account for the differences. Because published methods must be concise, they only include information about factors that are known to be relevant at the time of the initial publication. Once discordant results are produced, investigators usually begin to consider if some other factor(s), not identified or included in the original, published methods, might account for the differences in the results (Rosenthal, 1966, p. 34; Collins, 1992, p. 55). Unless the original results are confirmed, investigators can continue to repeat the study, revising the procedures each time, sometimes

with extensive communication with the initial investigator after each failed replication, to try to insure that their procedures are identical to the original report, including all possible factors that were not included in the published methods (Hendrick, 1990; Collins, 1992, p. 55).

In addition to considering the possibility that some unidentified factor might account for the discordant findings, discussions about what could account for different results also often devolve into debates about whether the initially reported phenomena were not replicated because the replication studies were conducted by inexperienced or otherwise incompetent experimenters. Experimenters then try to demonstrate that they are competent, conducting one study after another as they address one potential source of error after another. Since there are an infinite number of potentially critical but unidentified factors, as well as potential sources of error and variability, the process of sequentially testing one factor after another can continue indefinitely (Collins, 1992). As long as the results continue to be discordant, it is impossible to determine whether the differences are due to some critical but unidentified factor, a flaw in the design or methods, artifact or confounding factors, random chance, simple error, experimenter bias, incompetence or fraud (Bornstein, 1990; Rosenthal, 1990; Collins, 1992).

This long-term, time-consuming, labor-intensive and often ultimately fruitless process is based on the assumption that the burden of proof for explaining why the original findings were not confirmed is on the investigator conducting the replication studies (Neuliep & Crandall, 1990). However, logicians and philosophers are in virtually unanimous agreement that the burden of proof regarding potential efficacy is on the scientist making the positive assertion (Gilovich, 1991, p. 181). Discordant findings simply raise questions about the validity of the original findings—and regardless of why the results are discordant, the original findings cannot be accepted as valid until they can be reliably reproduced (Zuckerman, 1977, p. 96). Not only is it unnecessary for investigators conducting replication studies to show precisely what might account for their discordant findings, it is also unnecessary to completely and precisely replicate every conceivable aspect of the initial study. In fact, not only are precise replications unnecessary, they are undesirable.

This is an important point to understand, in part because it is impossible to precisely replicate an experiment. It is always possible that the replication experiments are different in some way, perhaps yet unidentified, from the original experiment. Replication studies validate the original findings by showing that they can be reliably reproduced, but also that they can be generalized beyond the precise circumstances used in the original study. If confirmatory results can only be produced under conditions that are identical to the original experiment, it is always possible that those effects could be attributed to some artifact or unidentified confounding factors. More general replications show that the observed phenomena are not specific to a single set of precise circumstances, and for that reason more general replications are actually required to validate any experimental results (Campbell & Jackson, 1979; Amir & Sharon, 1990; Rosenthal, 1990; Collins, 1992).

Of course, it is important to demonstrate that the investigators conducting the replication study are competent, and that their materials and methods are sufficiently sensitive and reliable to detect significant effects. While this may be a problem in some fields of research, in preclinical proof of concept studies it is almost always possible to satisfy these criteria (Collins, 1992). For example, it is important to show that the measurement instrument is not limited by ceiling or floor effects, and that it is sensitive to changes within the range expected, based on the results of the original findings. Assays and other measurement procedures can be validated by demonstrating that they are sensitive to, and can reliably detect, dose-dependent effects of drugs, individual differences between subjects or differences based on age, cell number or other relevant treatment manipulations—all of which serve to validate the experimental results, regardless of whether the effects detected in the original experiments were confirmed or not.

It should also be pointed out that, while discordant results obviously raise questions about potential validity, consistently replicated and confirmed results cannot automatically be accepted as evidence of definite validity. Consistent, apparently reliable replications can be due to repeated error or bias on the part of the experimenters. Early, empirical investigations of this phenomenon were conducted by Karl Pearson who gave people simple measurement tasks, such as manually bisecting lines on a page. He found that the errors were not randomly distributed, people tended to over or underestimate the mid-point to the left or right, and these errors were correlated between different investigators (Pearson, 1902). Pearson attributed correlated errors between different investigators to undefined similarities in environmental and experimenter traits.

Further evidence that consistent, reliable replications can be due to experimenter bias has come from similar research on measurement error. For example, observers who know that the reliability of their measures will be calculated will produce more reliable measures—their measurements will be more closely related to the values being obtained by the other experimenters in their group (Zuckerman, 1977). Inter-observer reliability can increase over time, due to “drift” in the definition by the group of how to score their observations, an effect referred to as shifting “idiosyncratic group consensus” (Kent et al., 1974). The effects of bias are so common that theories of measurement error actually attribute error to several components, including person-specific and systematic biases (Cochran, 1968). For example, astronomers tend to record the transit times of bright stars as faster than faint stars. Thus, consistent, reliable results can be produced, not because the phenomenon is valid, but because the investigators are conducting their studies consistently with the same biased procedures.

This means that numerous confirmatory replications can be the result of consistent bias on the part of the investigators. For example, the portacaval shunt was developed to treat esophageal hemorrhaging and later expanded to treat a variety of ailments of the intestinal cavity. After 20 years of use, 51 clinical studies had been reported with this procedure. Among the 47 studies that were not very well controlled to prevent bias, 72% (34 of 47) reported that the treatment produced marked

improvement, another 20% (10 of 47) reported modest improvements. Only a small number of studies of this treatment were very well controlled (4 of 51), and the results of those studies clearly showed that this treatment did not produce therapeutic effects (Fig. 8). Thus, unless they have very carefully controlled for the potential effects of bias, even a large number of studies do not provide conclusive evidence of efficacy, even if they all report consistent, robust, therapeutic effects. As sample size increases, different sources of measurement error tend to cancel each other out, so samples tend to approximate the population more and more closely as sample size increases. However, large samples do not overcome the bias that exaggerates apparent efficacy (Gilovich, 1991).

Clearly, potential evidence of preclinical efficacy must be replicated, but it is also important that the results be replicated under very well-controlled conditions before they can be accepted as valid. To be clear, it is not necessary or even appropriate to conduct all experiments under rigidly controlled and blinded conditions. Especially during the initial, exploratory phase of research, before the effects of a novel compound have been identified or demonstrated, it is often helpful for investigators to be aware of the treatment conditions in the experiments so that they can more easily identify significant, possibly unexpected effects. A range of statistical procedures have also been developed for exploring data for any patterns or relationships, regardless of whether they were planned or expected. However, while these exploratory experimental and analytical procedures allow for the identification of unknown and unexpected effects during the initial discovery phase, they are also vulnerable to “identifying” relationships and effects that are spurious (Hartwig & Dearing, 1979; Hoaglin et al., 1983; Cliff, 1987, pp. 153–155). For that reason, once the optimal experimental procedures, endpoints and statistical analyses have been determined, potential efficacy must then be confirmed in very well-controlled replication studies. In fact, the decision-making process would be optimized by evaluating preclinical

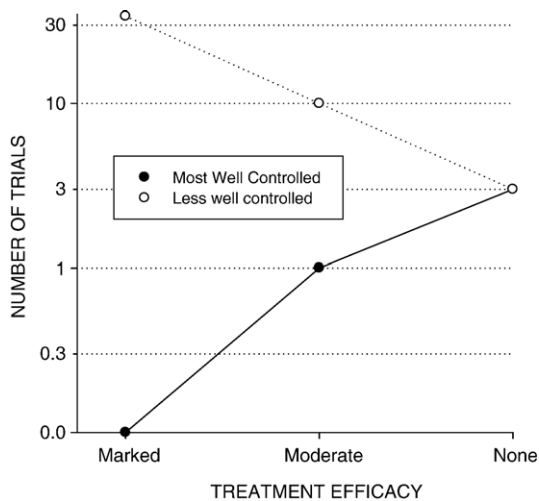


Fig. 8. Results of clinical trials on the portacaval shunt procedure, based on degree of control for potential bias (i.e., “most well controlled” versus “less well controlled”), from (Freedman et al., 1980, pp. 8–9; Grace et al., 1966; Gilovich, 1991).

studies in the same way that the FDA now evaluates clinical trials, with virtually all decision-making based on the results of only the most carefully controlled experiments.

### 5. Why are scientists vulnerable to bias; why can't they be objective?

Before discussing the specific types of bias that can affect experimental results, it is important to consider why scientists might be vulnerable to bias in the first place. Since the scientific method is dependent on objective observation, why can't we simply expect scientists to be objective?

Perhaps one of the most important reasons why it is difficult for scientists to maintain their objectivity is because of the integral role scientists play in deciding which areas of research to pursue, and in formulating the hypotheses and designing the experiments they conduct. Scientists pride themselves and are evaluated on their ability to make objective and accurate observations (Bernard, 1865), and the results of their studies and their ability to discover important, significant phenomena reflect their powers of observation, intelligence, expertise and creativity as scientists.

In addition to the desire to report significant, novel findings, scientists in drug discovery and development are also motivated to report potential efficacy, not only to establish their abilities as scientists, but also because they want to develop therapies to relieve human suffering. In addition, compounds are advanced into clinical trials based on preclinical evidence of potential efficacy. For scientists in drug discovery, their ability to maintain continued employment, to earn bonuses and promotions, and to advance up the corporate ladder into positions of greater prestige and power, is dependent on their ability to demonstrate the therapeutic potential of novel targets and compounds.

Scientists are motivated to demonstrate their ability to formulate and solve previously unsolved problems and to demonstrate significant, novel effects and efficacy with novel targets and compounds. They choose to do experiments, and they select the dependent and independent variables to be used because they expect a certain relationship to emerge between them. Max Weber wrote that scientists must be passionately devoted to their work, they must enthusiastically work long hours, with extreme focus and intensity, seeking inspiration in the same way that artists seek inspiration (Weber, 1946).

Scientists cannot be disinterested or skeptical—they must be committed to and involved with their ideas; otherwise they would not be willing to risk their time, effort, and occasionally their own money and careers (Ben-David, 1977). They naturally care about how the results turn out because their experimental results reflect on their ability to identify important, relevant questions, and to identify the critical factors involved to address those questions (Weber, 1946; Rosenthal, 1964a; Rosenthal, 1969; Ben-David, 1977). As Harrington so clearly stated, “caring deeply does not usually lead to affective neutrality” (Zuckerman, 1977).

It is almost impossible to expect scientists to be passionately committed to their research and at the same time to be completely objective and disinterested, but this raises questions

about why it is so often assumed that scientists are objective. Perhaps one explanation is that objectivity is expected and required only during the presentation and evaluation of the results, not during the actual process of conducting the research (Ben-David, 1977). For example, scientific papers written in the 1600s used a narrative style and reported exactly what scientists thought and what they found, including their sources of inspiration, guesses, failed experiments and frustrations; but journals no longer allow a narrative style, instead they require a very strict format which hides the actual research process and makes it appear as if scientists are completely unemotional and objective (Medawar, 1991; Silverman, 1991). This required format clearly helps to create and maintain the illusion of detached objectivity.

## 6. How bias affects experimental results

In this section, some of the different ways that bias affects experimental results will be reviewed before discussing specific procedures that can be used to limit bias.

### 6.1. *Effects of bias on attention and perception*

Perhaps the first and simplest functions that can be affected by bias are the experimenters' attentional and perceptual processes. It is well established that our conscious experience is not a direct reflection of the physical world (Marcel, 1983). Francis Bacon was already aware that sensory inputs are perceived differently for different individuals, based on their previous experiences and beliefs, "just as an uneven mirror" distorts images based on its particular, unique shape (Bacon, 1620), and his observations have been supported by the evidence up to the present (for discussion, see Hanson, 1958; Nisbett & Ross, 1980).

One example of how expectations affect perceptions is from an anecdotal report of a hunting accident in which 2 hunters, expecting to see a deer, saw a deer and shot it, not once, but twice. The "deer" turned out to be another hunter from their own party, but they continued to insist that they thought they saw a deer, despite the fact that under similar conditions (but when they were not hunting), they had no problem recognizing a man accurately at the same distance (Sommer, 1959). A good example of how prior experience affects our perceptions is the classic study by Turnbull, who reported that BaMbuti pygmies that grow up entirely in thick jungle forests, without access to open spaces and distant views, believed that objects were literally growing in size as they approached them, because they were unable to conceive of the effects of distance on apparent size (Turnbull, 1961).

Other examples of how our expectations affect our perceptions include a study in which subjects were given brief, tachistoscopic exposures to playing cards, some of which had been altered by reversing the colors (e.g., black 3 of hearts, red 2 of spades). Normal cards were identified quickly and with 100% accuracy after as little as 350 msec, but even after 1000 msec exposures, subjects often incorrectly identified the incongruous cards based on the color or shape of the suit,

without even realizing that the cards were incongruous (Bruner & Postman, 1949). In another study, observers' recognition of familiar 3-dimensional objects were unaffected when the objects' depth structure was scrambled, as long as their 2-dimensional projections were unchanged, and the observers were unaware of the depth anomalies introduced by scrambling (Bulthoff et al., 1998). In yet another example, practice golf balls are lighter than regular golf balls, and experienced golfers that are aware of this difference judged golf balls that they thought were practice balls to be lighter than regular golf balls, even though their actual weights were the same (Ellis & Lederman, 1998). Finally, subjects conditioned to a light paired with a very faint tone later reported that they heard the tone after the light, even when the tone had not been presented (Ellson, 1941). These examples support a top-down process in which expectations can actually override sensory inputs.

Research has also shown that individuals focus their attention on stimuli related to their concerns. For example, smokers who had abstained from smoking overnight attended more to words related to smoking and were distracted from focusing on other cues in an emotional Stroop task (Gross et al., 1993; Waters & Feyerabend, 2000), and students interpreted ambiguous pictures as having something to do with food and eating more often when they were tested before a meal than when they were tested after a meal (Sanford, 1936). As another example of this phenomenon, subjects asked to answer true or false questions were then provided with an answer sheet. Although some of the characters on the answer sheet were ambiguous (illegible), subjects often perceived the ambiguous character as the same "T" or "F" they had recorded on their test—just as they would have hoped (Stephens, 1936). Perceptual and attentional functions are more likely to be affected by hopes and expectations when the stimuli are more interesting and complex and when perceptions are based on recall, imagination, or inferences (Nisbett & Ross, 1980). It has even been reported that the human mind is so specially adapted to detect patterns that people often perceive patterns even when faced with completely random information (Gilovich, 1991).

### 6.2. *Conditioning of bias*

Bias can also be the result of conditioning. For example, subjects were shown lines of varying lengths or given objects of varying weights, and given a reward (money) associated with the longer lines and heavier weights, and a punishment (loss of money) associated with the shorter lines and lighter weights. When later asked to estimate the length or weight of objects, they tended to over-estimate the lengths of the lines and the weights of the objects (Proshansky & Murphy, 1942). In another study, subjects were shown line drawings and given similar monetary rewards and punishments. When these drawings were shown together in a way that was ambiguous, they tended to "see" the drawing that had been associated with a reward (Shafer & Murphy, 1943). In another study, if investigators were regularly given feedback in the form of approval or disapproval, with respect to whether the data being collected were in line with expectations or not, their ratings of videotaped behaviors were

affected in the expected direction (O’Leary et al., 1975). Subjects have also been covertly conditioned to start their sentences with the pronouns “I” and “we” by experimenters who responded with “good” when they did this, without specifying precisely what was good. As part of this experiment, experimenters also recorded better or worse conditioning, depending on whether they expected to see better or worse conditioning in their subjects (Rosenthal et al., 1966).

Experimenters may also be conditioned by experimental animals. In one case, a horse was thought to be able to read text and to solve complex math problems on command. Correct answers to questions and problems could be obtained from the horse by a number of different questioners, and a panel of expert veterinarians, jockeys, officers in the cavalry and animal psychologists, after careful study, determined that these accomplishments were not due to any intentional tricks or deceit on the part of the trainers, but were due to unintentional cues that the trainers were unwittingly giving the horse. Intense concentration on the correct answer enhanced the unconscious movements the questioners were making to cue the horse (Pfungst, 1911). As another example, experimenters that expected rats to learn faster tended to handle their rats more often and more gently than experimenters that expected their rats to learn more slowly. Rats that were handled more gently and more frequently were less stressed and more likely to perform better in learning tasks (Rosenthal, 1963; for review, see Rosenthal & Fode, 1963; Rosenthal & Lawson, 1963; Rosenthal, 1969).

Research assistants and their superiors may also be engaged in subtle, covert, reciprocal conditioning, the subordinates responding to cues from their superiors and also emitting cues about what they respond to, and the superiors emitting cues about the results they hope to see and responding to cues from the subordinates about how to attain those desired results. These reciprocal cuing and conditioning communications may not be under “conscious control,” and it is not clear how these signals are communicated, but this possibility is supported by the fact that these effects increase over time, with more interpersonal interactions and experience, and there is evidence that subtle cues are being conveyed that produce the expected and desired effects (Rosenthal, 1966; for review, see Rosenthal, 1969). In a study that reflects how reciprocal cuing and conditioning might occur, teachers told that certain students were very bright and had tremendous potential, smiled more at these students they thought had high IQs, they made more eye contact with them, leaned forward more and nodded their heads more. These behaviors made students more likely to enjoy school and work harder to improve, which further reinforced the teachers’ encouragements (Chaikin et al., 1974).

### 6.3. *The demands of authority and the research setting*

Experimenter bias can also be attributed to the force of authority and/or the demands of the research setting. For example, in a study of depressed patients given atropine (an active placebo which produces dry mouth), patients responded positively if the investigators told them that dry mouth was an indication of drug potency, but they responded poorly if they had

been told that dry mouth was an indication of extreme toxicity (Kast, 1961). In another study of the effects of zinc on people with taste disorders, when patients were blind to treatment allocation but the investigators were not, significant improvements were seen with zinc administration (Schechter et al., 1972). The expectations of authority figures also affect their subordinates, not just their patients. When a doctor suggested to mental health workers that a subject was neurotic or psychotic, these mental health workers were more likely to label the subject as psychotic, even though the subjects were actually normal (Sushinsky & Wener, 1975).

The force of authority has such a strong effect on most people that they will even do things that are unethical and immoral. In a classic psychology experiment, investigators were instructed by an authority figure to deliver electric shocks to a subject as a punishment whenever they gave an incorrect answer to a question. The shock intensity increased with each shock delivered, and although the investigators expressed their reluctance, 100% of them agreed to administer shocks up to the “intense shock” levels of 300 Volts, and 65% of investigators agreed to administer the shocks up to the highest levels of “severely dangerous,” 450 Volts. These investigators were normal people, and they were not threatened or coerced in any way, except with the pressure to obey authority, yet they continually agreed to administer more and more severe shocks, even beyond levels that they knew were extremely dangerous (Milgram, 1963). The authority figure in these studies was a research scientist, and these findings have been replicated in many studies over a period of decades (for review, see Blass, 1999). Perhaps one of the most interesting findings in this literature is that even experts (psychiatrists) are unable to predict the large percentage of investigators that will do things they think are unethical and even immoral, in order to please an authority figure. Milgram conducted a survey of 40 psychiatrists, and when the protocol was explained to them, they predicted that no more than 1 out of 1000 people would go to the highest shock levels, but in fact, more than 60% of subjects did so. A substantial portion of people do what they are told to do, without limitations of conscience, so long as they perceive that the command comes from a legitimate authority (Milgram, 1965).

In addition to the expectations and demands of authority figures, people also seem to respond to more subtle demands or contextual cues about what is expected of them. The phenomenon of demand characteristics was initially reported by Orne and his colleagues, in studies in which subjects were asked to simulate a hypnotic state. Subjects were able to simulate hypnosis so well that they could not be discriminated from subjects that were hypnotized. Subjects that were simulating hypnosis were able to endure pain and demonstrate feats of unusual strength even more than subjects that were actually hypnotized. These phenomena were demonstrated by simulating subjects even though they were not specifically instructed how to behave, they simply assumed or were able to surmise from the behavior, requests, instructions, and questions from the hypnotist, and other contextual cues surrounding the experiment, what was expected from them (Orne, 1959; Orne &

Scheibe, 1964; Orne & Evans, 1965). Subjects that were only simulating a hypnotic state were willing to administer shocks to themselves that were much more intense than hypnotized subjects; they were willing to pick up poisonous snakes, immerse their hands in concentrated acid, and even to throw acid into the face of another experimenter (Shor, 1964; Orne & Evans, 1965).

Orne concluded that the experimental situation takes place within the context of an explicit agreement of the subject to participate in a special form of social interaction. Within the context of our culture the roles of subject and experimenter are well understood and carry with them well defined mutual role expectations. People share with the experimenters their assumption that their participation with the research will contribute to science and perhaps to human welfare. Both subject and researcher share the belief that whatever the experimental task is, it is important, and no matter how much effort must be exerted or how much discomfort must be endured, it is justified by the ultimate purpose. Subjects also assume that their proper participation in the study is intended to validate the experimental hypothesis (Orne, 1962).

According to Orne, demand characteristics are part of any experiment and cannot be removed (Orne, 1962). The behavior that is expected of a subject is communicated only partly by the instructions, and also by subtle cues provided by the experimental context and procedure, and the experimental hypothesis under study. Subjects assume that adequate precautions for their safety and welfare will be taken, and the greater the investment of time, effort and discomfort on the part of the subject, the more he is compelled to care about its outcome and assume that it is important. The sum total of cues which communicate the purposes and intent of the experiment, particularly those aspects about which the subject is not explicitly informed, constitutes the demand characteristics (Orne, 1970).

Given the powerful effects that obedience to authority and contextual demand characteristics often exert, it is especially troubling to consider just how vulnerable many of the investigators that actually conduct most of the research are to these phenomena. Roth reported that hired hands in research have the same mentality as hired hands in any other production unit. Instead of following the ideal code of conduct of the “dedicated, honest and objective scientist,” they do not feel a sense of ownership of their work or work product, and they usually lack the extensive training on the importance of objectivity and accuracy, so they often cut corners and fabricate results that they think will “please their boss,” just as workers in any other production unit (Roth, 1966).

Taken together, the research by Milgram, Orne and Roth have shown that the research subordinates and associates that actually conduct most of the research are willing to satisfy the expectations of their superiors because of the contextual, social and behavioral control exerted over them by what they believe are competent, responsible investigators (Orne, 1959, 1962; Milgram, 1963, Orne & Scheibe, 1964; Milgram, 1965; Orne & Evans, 1965; Roth, 1966; Orne, 1970). As examples of this phenomenon, assistants working for Pavlov and Mendel apparently produced invalid experimental results in order to

demonstrate the effects their supervisors expected (Fisher, 1936; Zirkle, 1958; Razran, 1959).

#### 6.4. Bias in experimental design

There are any number of ways that an experimenter can bias the outcome of a study during the initial design. One way that bias can affect the design of the experiments is in the choice of the control group. For example, clinicians that select historical controls with higher levels of disease severity and progression are more likely to detect significant therapeutic effects with their novel treatments. Clinicians using historical controls reported that 79% of their trials produced beneficial results, but the same treatments only produced beneficial effects in 20% of trials when controls were randomly assigned as part of the study design (Sacks et al., 1982). As another example, non-profit organizations showed that new therapies were more effective than standard therapies only half the time (47%), while randomized trials supported by industry sponsors showed that the new therapies were more effective than the standard therapies much more often (74%). This difference is attributed in part to publication bias, but also to the selection of inappropriate comparator treatments in the industry sponsored trials (Djulgovic et al., 2000; Lexchin et al., 2003).

Perhaps the most common source of bias in the design of preclinical proof of concept studies is in the choice of tests and models. For example, in a survey of recent treatments for AD (Table 2), there was a clear predominance of preclinical experiments which demonstrated potential efficacy using the passive avoidance test, despite the fact that the passive avoidance task is widely recognized as a weak model with little clinical relevance and a high rate of false positives (Sarter et al., 1992a; Sarter et al., 1992b).

The use of small sample sizes might also be considered a source of bias in experimental design. Conducting 10 smaller studies instead of 1 large study will increase the probability that at least one of those studies will produce the desired result. If studies with positive results are then selectively reported and studies with negative results are simply discarded, conducting numerous smaller studies increases the probability of producing at least 1 positive result. Begg and Berlin (1989) showed treatment-related improvements in survival rates in clinical trials for cancer were largest (19%) in the smaller trials ( $n \leq 50$ ), and smallest (0%) in clinical trials with larger samples ( $n > 100$ ). A study on treatment of advanced cancer also reported response rates of 8–85%—the larger response rates were associated with the smaller studies, while response rates in larger studies ranged from 12% to 31% (Moertel, 1984; Begg & Berlin, 1989).

As one last example, when patients are selected based on some diagnostic assessment, their scores usually need to be above or below some criterion value, a value which is usually considered outside the range of normal values. Some people will score outside the normal range by chance alone, and when that group is retested, their retest scores are naturally more likely to be closer to the mean for the normal population. This is a phenomenon known as “regression towards the mean” (James, 1973). When using this procedure, if treatment effects are

assessed simply by comparing measures collected during treatment to baseline measures collected before treatment, regression to the mean will be interpreted as a significant treatment effect. For these experiments, control groups need to be included and effects of the treatments need to be based on changes that occur in the treatment groups relative to changes in the control group. There are of course many other ways that investigators can bias their results by manipulating the design of their experiments.

### 6.5. Effects of bias on errors in recording and managing data

Errors occur in recording observations, and these errors are generally not random, but are more often in the direction of the expectations of the observer. In one series of studies, errors occurred in observations only ~ 1% of the time, but more than 67% of these errors were in the direction of the observer's hypothesis (Rosenthal, 1969). In a study of extrasensory perception, 10 decks of cards were shuffled, an observer would focus on each card and the agent would try to read the mind of the observer (Kennedy & Uphoff, 1939). The agent would state what they thought the card was, and the observer would record what the card was and what the agent had called. By comparing the records of the observations recorded between the observer and another, independent observer, it could be shown that the observer made errors in favor of their expectations. If they believed in ESP, they made a small number of errors, but most of these mistakenly recorded that the subject had correctly guessed the card when they had not (Kennedy & Uphoff, 1939; Rosenthal, 1969). In a similar study with people who claimed that they could influence the roll of dice, photographic evidence revealed that these effects could be attributed to observer errors in how the results were recorded—believers in psychokinesis made errors favoring psychokinesis, but non-believers made errors that were in the opposite direction (Sheffield et al., 1952).

Experimenters can also bias their results through their selection and use of exclusion criteria, especially if they are excluding records after the blind has been broken, based on how well the data meet the experimenters' expectations (May et al., 1981; Gotzsche, 1989). In a clinical trial of anturane for treatment of myocardial infarction (Anturane re-infarction trial research group, 1980), subjects in both treatment groups were excluded from the analyses, but patients that died during the study were more likely to be excluded from the analyses if they were in the treatment group than if they were in the placebo group. In preclinical studies, some kind of procedure to check the quality of the data are required, but these quality control procedures also open the door to bias in the process of deciding which records should be excluded. As one example, it is common for investigators to make notes of any potential issues as the data are being collected, but to wait to see the results to decide which specific records should be excluded from the analyses. This practice dramatically increases the risk of bias; and since preclinical studies do not require that excluded data be reported, it is especially difficult to determine if bias might have affected the data exclusion process.

### 6.6. Bias in data analyses

In addition to affecting perceptions and experimental design, experimenter bias can also play a role in the way the data are analyzed. To begin with, simple computational errors can occur, and although they are rare, when they do occur they are usually in the direction of the observer's hypothesis. As one example, experimenters that committed errors in recording their observations also made computational errors, and just as with the recording errors, these computational errors were more often in the direction of the expectations of the experimenters and tended to be larger than errors in the opposite direction (Rosenthal et al., 1964; Rosenthal, 1966, p. 12; for review, see Rosenthal, 1969).

While computational errors occur, the most common ways that bias affects data analyses are by conducting multiple analyses without adjusting the level of significance, and deciding how to analyze the data after examining the results (i.e., post-hoc analyses). Repeated analyses can be conducted on raw scores, differences from baseline, and scores adjusted based on any number of potential covariates. Data can also be divided into subgroups based on a range of different criteria (Pocock et al., 1987; Assmann et al., 2000; Pocock et al., 2002). While it is often assumed that the results of statistical analyses will only be "significant" 5% of the time, this is only true if 1 planned analysis is conducted. In a study making 10 comparisons, there is a 40% chance of detecting at least 1 significant effect ( $P \leq 0.05$ ) just by chance alone (Tukey, 1977). If a study with 10 possible 1-tailed comparisons is conducted 3 times, there is an 85% probability that at least one of these experiments will produce a statistically significant effect in the desired direction.

Most preclinical proof of concept studies are sufficiently large and complex to include more than a few potential endpoints and several possible ways of adjusting or transforming each endpoint, and there are different types of pattern analyses that can be conducted, which often inflates the probability of detecting some evidence of efficacy beyond 40–50% for each experiment (Tukey, 1977). Experiments can then be repeated, which will virtually insure that at least 1 experiment will produce significant results. The data can also be analyzed as it is collected, and by continuing to collect and repeatedly analyze the data, statistical significance will eventually be assured (Armitage et al., 1969; Tukey, 1977; Diaconis, 1978). Finally, post-hoc analyses can also be conducted after examining the results of a study, and it is possible with hindsight to build a very favorable statistical analysis (Gilovich, 1991).

Even among the highest ranking journals studies often include analyses of multiple end points, repeated measurements over time, comparisons between subgroups, and comparisons between multiple treatments (Pocock et al., 1987). Even among clinical trials which tend to be much more carefully controlled than preclinical experiments, two-thirds of clinical trials present results of subgroup analyses without adjusting the  $P$  value, and without making it clear if the analyses were planned or not (Assmann et al., 2000). In 1 clinical trial almost 200 analyses were conducted, 8 produced "statistically significant" results ( $P \leq 0.05$ ), and although 10 false positives would be expected

out of 200 tests, all 8 statistically significant tests were interpreted as evidence of therapeutic effects (Anturane Reinfarction Trial Research Group, 1980). In a clinical trial of aspirin and streptokinase in patients with acute myocardial infarction, the authors demonstrated the dangers of post-hoc comparisons by showing that patients born under the Gemini or Libra astrological birth signs did somewhat worse on aspirin than on no aspirin, whereas for all other astrological signs, and overall, there was an impressive and highly significant benefit from aspirin (Friedman et al., 1998, p. 17).

Also in the Anturane study, the authors noted an apparent reduction in sudden deaths during the first 7 months of follow-up. In light of their observation, they defined sudden deaths during the first 7 months of follow-up as a new category of critical events and claimed that it was evidence of a statistically significant effect (Rose, 1982). These post-hoc subgroup analyses are inappropriate, and it has been shown that subgroup differences based on post-hoc analyses are often not replicated in subsequent trials (for review, see Yusuf et al., 1991). Simulation studies have also shown that post-hoc selection of covariates for adjustment, out of a larger set of potential covariates, will tend to lead to biased estimates of the treatment effect, especially with small studies (Beach & Meier, 1989).

Post hoc analyses are sometimes conducted after exploratory analytical procedures are used to identify patterns in the results. As mentioned previously, a range of statistical and analytical procedures have been developed for exploring data, post hoc, for any patterns or relationships, but these procedures are vulnerable to “identifying” relationships and effects that are spurious (Hartwig & Dearing, 1979; Hoaglin et al., 1983; Cliff, 1987, pp. 153–155). Therefore, once the optimal experimental procedures, endpoints and statistical analyses have been determined in initial exploratory studies, potential efficacy must then be confirmed with planned analyses in very well-controlled confirmation experiments (Pocock, 1997).

Sometimes it is difficult to distinguish between inappropriate analyses and inappropriate interpretation of the analyses. For example, it is inappropriate to conclude that 1 treatment is superior to another treatment based on the fact that 1 treatment is significantly different from the control but another treatment is not significantly different from the control. Treatments need to be compared to each other in order to determine if they are different from each other (Gotzsche, 1989), but this example leads us into the next section which continues with more discussion about how bias can affect the interpretation of different kinds of information.

### 6.7. Interpretation bias

At every step along the way of characterizing the results, there is room for interpretation. The same set of facts can be evaluated and interpreted in different ways, and once the results of experiments are communicated, their value, significance and impact are open to further interpretation and bias on the part of the consumers of the information—to over-rate or under-rate based on things like the prestige of the author, institution, and journal, and whether the results are consistent with their hopes,

expectations and beliefs (Nisbett & Ross, 1980; Owen, 1982; Vandembroucke, 1998).

As an example of bias in favor of findings consistent with expectations and beliefs, reviewers judged manuscripts to be of higher quality and they were more likely to recommend publication if the paper supported the reviewers' views and previous findings. The effects of bias were also larger among scientists with stronger prior beliefs (Mahoney, 1977; Koehler, 1993; Ernst & Resch, 1994). For example, subjects told that they scored high or low on tests of intelligence and social sensitivity later judged studies of the validity of those tests more or less favorably depending on how desirable their own scores had been (Wyer & Frey, 1983; Pyszczynski et al., 1985).

Not surprisingly, more subjective assessments, such as judgments about the degree of relevance, quality, and clarity of an experiment, are more vulnerable to bias than more specific, objective judgments (Koehler, 1993). For example, 1 investigator tended to interpret his results more favorably, concluding that 58% of all the compounds he tested in clinical trials were efficacious and safe enough to warrant regulatory approval, while another investigator tended to be much more conservative and reported that only 19% of the compounds he tested were sufficiently efficacious and safe to warrant regulatory approval (Greiner et al., 1950).

Obviously, different interpretations are due to differences in the way information is processed. Information that is consistent with expectations and desires is readily accepted, but more time is spent processing information that is inconsistent with expectations: arguments are developed to refute and discredit that information, and unexpected or undesirable information is subsequently judged to be weaker than information that is compatible with prior beliefs (Edwards & Smith, 1996; Ditto et al., 2003). There are many examples of this phenomenon. For one, journal reviewers rated papers higher if they were reporting results of an orthodox treatment than if they were reporting the same results due to an unconventional treatment (Resch et al., 2000).

Baar and Tannock (1989) constructed a hypothetical data set and reported the results 2 different ways, 1 low quality version with errors of reporting and omissions, similar to those seen in the literature, and a high quality version of the report produced with all the appropriate methods. The low quality report led to statements and conclusions that the treatment was effective and safe, while the higher quality version of the same report concluded that the treatment was ineffective and toxic. This exercise clearly illustrates that the same results can be interpreted and reported very differently, depending on the training and scientific rigor of the authors (Friedman et al., 1998, p. 336).

These biases can have profound effects on the progress of science and the development of novel treatments. For example, it took years to accept the link between *Helicobacter pylori* and peptic ulcer, despite clear supporting evidence, because the stomach was believed to be too acidic to support bacterial growth (Thagard, 1999). Polanyi (1963) also described his experience of publishing a paper which was rejected, and his career was almost ruined, because his theory and data were inconsistent with orthodox views of the time. It took 50 years

before his theory was finally accepted as valid. In fact, it is so difficult for people to accept new findings that are inconsistent with what they believe that Max Planck wrote, “A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it” (Planck, 1949), and this sentiment was supported by Thomas Kuhn (Kuhn, 1996).

#### 6.8. Publication bias

As discussed previously, preclinical scientists are more willing to report and publish studies that report statistically significant therapeutic effects. Studies that fail to detect statistically significant effects, including replication studies, are less likely to be reported and published (Dickersin et al., 1987; Dickersin & Min, 1993). Since some false positives are always produced, just by chance, the bias to publish significant, therapeutic effects means that false positives are over-represented and may dominate the published literature (Mahoney, 1977; Begg & Berlin, 1989; for review, see Sterling, 1959; Mahoney, 1985, 1987; Wilson et al., 1993).

A survey of clinical trials reported to the Finnish National Agency for Medicines showed that final, complete reports were generated twice as often for trials with significant, positive results than for trials with negative results: 38% of trials with positive results were reported, but only 19% of trials with negative or inconclusive results were reported (Bardy, 1998). In another survey of over 1300 controlled trials, publication bias was even more profound: 55% of published reports reported that the novel therapy under study was effective, but only 14% of unpublished trials suggested that the novel therapy was effective (Dickersin & Min, 1993; Dickersin et al., 1987).

Obviously, the bias to preferentially report and publish studies with positive results makes it difficult to rely on the published literature to properly assess the therapeutic potential of novel treatments. For example, novel treatments for cancer produce significant and fairly robust effects among the published trials, but the same treatments often produce effects that are much smaller and not statistically significant if all published and non-published studies are included in the analysis (Simes, 1986). A meta-analysis conducted on all published studies produced modest (16–19%) but statistically significant effects of treatment, but a meta-analysis including all registered studies (published and unpublished), produced smaller effect sizes that were not statistically significant (Simes, 1987).

Publication rates are highest among randomized, controlled clinical trials: in 1 survey, 98% of randomized, controlled trials with significant results were published, while only 86% of non-significant studies were published (Dickersin & Min, 1993); but, even with these high publication rates, publication bias can significantly affect assessments of the therapeutic potential of novel treatments. For example, in an initial publication of a non-randomized clinical trial, beta blockers produced significant reductions in mortality rates. Two randomized, controlled trials subsequently failed to detect any significant therapeutic effects of beta blockers, but those studies were not published (Snow,

1965; Furberg & Morgan, 1987). Someone relying on the published literature to assess the therapeutic potential of beta blockers would only see the results of the 1 study that reported positive effects and might not be aware of the 2 more well-controlled studies that failed to detect therapeutic effects.

The bias for publishing positive results among non-clinical studies is much higher than among randomized, controlled clinical trials (Easterbrook et al., 1991). One study showed that authors submitted 82% of their studies with positive outcomes but only 43% of their studies with negative outcomes. Editors then published 80% of the submitted reports with positive results and only 50% of the submitted reports with negative results. The combination of author and editor bias resulted in the publication of 67% of studies with positive results and only 20–33% of studies with neutral or negative results (Coursol & Wagner, 1986; Easterbrook et al., 1991).

The bias to publish studies with positive, therapeutic effects, but not studies that failed to detect significant effects, combined with other forms of bias that exaggerate potential therapeutic effects, results in a preponderance of publications that report positive effects: as many as 97% of non-clinical publications report positive effects (Sterling, 1959; Dickersin & Min, 1993). One consequence of this publication bias is that the results of studies included in the published literature cannot be accepted at face value (Dickersin et al., 1987; Easterbrook et al., 1991; Dickersin & Min, 1993), and the published literature cannot be used to reliably assess the therapeutic potential of novel compounds.

#### 6.9. Citation bias

Not only are studies with statistically significant results more likely to be published, they are also more likely to lead to a greater number of publications and presentations and to be published in journals with higher citation rates (Easterbrook et al., 1991). Studies reporting positive effects are then cited more than studies reporting negative effects (Gotzsche, 1987; Kjaergard & Gluud, 2002). Publications reporting insignificant effects are cited in favor of the author's hypotheses, as if they were statistically significant; unsupportive results are quoted as if they are supportive; and studies reporting negative results are often not cited at all. Clearly, authors cite papers that support their hypotheses and ignore or incorrectly cite papers that do not support their hypothesis, which results in biased views of the literature (Ravnskov, 1995).

#### 6.10. Bias in selecting experts for feedback

In addition to conducting studies and reading the literature, preclinical scientists also often identify experts they can rely on for guidance. Of course, bias can affect decisions about which experts to select. Their general beliefs and overall orientations, and even their positions on the specific targets of interest are usually well known from the same publications that serve to identify people as experts. Experts that support specific targets can be solicited, not to provide guidance in the decision-making process, but to provide support for targets and programs that have already been selected for development (Gilovich, 1991). In

other words, experts are actually often selected because they are advocates of some specific target, not because of their general knowledge and expertise in an area of research, and their input may serve to maintain an illusion that the decision-making process is based on the weight of the evidence.

Adding to the fact that experts are preferentially selected based on the targets they advocate, another problem with using experts is that they are often overconfident in their judgments, even when they are wrong, and they continue to maintain their confidence in their judgments regardless of the evidence (Fischhoff et al., 1977; Einhorn & Hogarth, 1978; Nisbett & Ross, 1980). For that reason, Sackett, who served as an expert, expressed his concern that experts actually do more harm than good. They accept or reject new information and hypotheses and add their prestige to hypotheses that are consistent with their own prior positions and publications, not on the basis of scientific merit. Their contribution supports established theories and makes them more difficult to replace with newer, more accurate observations and hypotheses. The net effect of expert contributions is thus often to impede the rate of scientific progress (Sackett, 1983).

#### 6.11. Bias in the decision-making process

Once preclinical data has been collected, analyzed, and reported, it must then be used in the decision-making process to determine if there is sufficient evidence to justify proceeding into clinical trials. Of course, the decision-making process is vulnerable to bias (Merton, 1948; Bornstein, 1990). For example, when a formal decision tree is not in place, complex problems are solved with the use of simple strategies or heuristics, and although these heuristics are generally quite useful, they sometimes lead to severe and systematic biases and errors (Tversky & Kahneman, 1974). Scientists, even experienced statisticians, use these same heuristics, despite their shortcomings, instead of more appropriate probabilistic reasoning (Einhorn & Hogarth, 1978).

One heuristic is to test hypotheses by searching for supportive information. This strategy is very efficient and often effective, but its use results in “confirmation bias” (Gilovich, 1991; Kerr & Tindale, 2004). Information that is consistent with prior hypotheses is accepted uncritically, at face value, while evidence that is inconsistent with prior hypotheses is either ignored or critically scrutinized until it can be discounted or distorted—consciously and unconsciously—until it seems consistent with prior beliefs (Merton, 1948; Bornstein, 1990; Gilovich, 1991).

Not surprisingly, 1 consequence of confirmation bias is that hypotheses are rarely rejected. No matter which side of an issue a person supports, their position actually tends to be strengthened over time. Even if they are exposed to information that is not always consistent with their position, with sequential evaluations initial positions are revisited and confirmed over and over again, which ultimately strengthens the initial bias. Even completely random empirical findings can be filtered and manipulated until they seem to strengthen prior hypotheses (Snyder, 1984; Jonas et al., 2001). Because people remember the results of their evaluations for each data set, they can be misled

by what in hindsight seems to be large, compellingly consistent samples of evidence, when in reality the remembered evidence is hopelessly tainted by biased interpretations (Einhorn & Hogarth, 1978; Nisbett & Ross, 1980). This phenomenon has been demonstrated convincingly by the fact that people with different perspectives believe more strongly in their initial hypotheses, even though they are polar opposites, after exposure to the same series of findings, some of which is consistent with 1 perspective and some of which is consistent with the opposite perspective (Lord et al., 1979).

Decision making can also be adversely affected by use of the “intuitive representativeness” heuristic. For example, when people are told that a population consists of 30% engineers and 70% lawyers, they correctly guess that an individual selected at random from this population will most likely be a lawyer (70% probability). However, when given a description of the personality of the individual (even though it actually provides no meaningful or predictive information about their occupation), predictions are based entirely on how closely the subjective description matches perceived stereotypes for each occupation. For example, if the description suggests the person is very meticulous, the prediction is that the person is probably an engineer, despite their knowledge that engineers make up only a minority (30%) of the population, and despite the fact that being meticulous doesn’t distinguish between engineers and lawyers (Kahneman & Tversky, 1973; Nisbett & Ross, 1980, pp. 140-150).

Decision-making can also be biased by incorrect beliefs about associations between variables (Chapman, 1967; Chapman & Chapman, 1969; Dawes et al., 1989). Co-variation is difficult to detect unless it is consistent with an existing theory, in which case very small correlations are “detected” even when they are not present (Nisbett & Ross, 1980). For example, with the Draw-A-Person test (DAP), patients draw a picture of a person and a psychiatrist examines the drawing to diagnose the patient, such that if the patient distorted or emphasized the eyes in their drawing, this suggests that they may be suspicious and perhaps paranoid (Chapman & Chapman, 1967; Nisbett & Ross, 1980, p. 94). Although empirical studies clearly show that the DAP is not valid, clinicians continue to use it because their biased interpretations of the drawings support their assumptions about the patients.

This brings us to another common problem in decision-making processes, which is that people are overconfident in their own judgments and continue to hold on to their beliefs—such as the belief that they are good at making decisions—even after they are faced with proof that their beliefs are incorrect (Ross et al., 1975; Fischhoff et al., 1977; Einhorn & Hogarth, 1978; Gilovich, 1991). In one study, clinicians were shown to be very confident in their diagnoses even when they were completely incorrect (Fischhoff, 1982; Faust et al., 1988).

Overconfidence in judgments may be due to reiteration effects and hindsight bias. Hindsight bias refers to the fact that when people know the outcome of some process, they tend to assume that they would have predicted that outcome, regardless of whether they actually would have predicted it or if the actual probability of the outcome is very low (Fischhoff, 1975; Fischhoff & Beyth, 1975). Hindsight bias is a problem especially

when results from preclinical assessments are evaluated without a predefined decision tree. Results from studies that fail to detect therapeutic efficacy can be discounted, in hindsight, because of some “flaw” in the experimental design.

The reiteration effect refers to the fact that hearing an assertion repeated over and over increases the degree of belief in that assertion (Hertwig et al., 1997). Reiteration effects are common to drug discovery programs because the rationale for selecting a target and the evidence supporting its potential efficacy are repeated over and over again over the years, as the program progresses through all phases of discovery and development, until tentative, suggestive evidence is eventually accepted as fact.

It is sometimes assumed that the bias of individuals can be overcome by making decisions in groups. If 1 member of the group can identify the correct solution to a problem, they can guide the rest of the group to accept that solution. However, correct solutions are not accepted unless they are supported by at least some members of the group, and usually the majority of the group will dominate the decision-making process out of proportion to the size of their majority, irrespective of the correctness of their decision, so that group decision making can actually amplify individual biases, rather than attenuate them (Kerr et al., 1996; MacCoun, 1998). One example of the fact that group decision-making does not always off-set the biases of individuals is the fact that groups that have made a decision are more likely to continue to support their decision and even escalate a chosen course of action despite evidence that it is failing (Staw, 1976; Whyte, 1993; Jones & Roelofsma, 2000).

## **7. Bias is not limited to inexperienced and unethical scientists**

The suggestion that any significant portion of the scientific literature might not be valid is perceived as an outrageous, personal attack on the credibility, integrity and honor of every member of the scientific community (MacCoun, 1998). Representatives of the scientific community respond to such suggestions defensively, by arguing that 99.9999% of scientific reports are accurate and truthful, and they deny that any significant changes are necessary (David, 1983; Koshland, 1987; Mahoney, 1987). Just as any other problem, bias cannot be addressed until it is recognized as a problem.

### *7.1. Experimenter bias is not under conscious control*

In order for bias to be accepted as a problem, it is important to clearly distinguish it from fraud. Fraud is defined as a conscious and willful act of deceit, which may include altering or completely fabricating data or results. In contrast, bias affects perceptions, interpretations and decision-making processes outside of conscious awareness. For example, with respect to sensory/perceptual processes, it is well known that what we think we see or observe is shaped by what we expect to see, and that these interpretations of our sensory inputs occur automatically, on an unconscious level, beyond our awareness or control. Since we are not aware that our expectations may have prejudiced our perceptions, we are left with a feeling of confidence and sometimes even strong conviction that

our perceptions are correct even when they are not (Johnson, 1953; Marcel, 1983; Fleming, 1992). For example, subjects interpret ambiguous stimuli as if they are consistent with their hopes and expectations, even when they are aware that their decisions are going to be checked and that any errors will be detected (Stephens, 1936).

With respect to errors in recording and processing data, it is unlikely that these are the result of conscious, intentional manipulations because errors are made in the direction of the observer’s expectations even when they know that independent records are being kept, and any errors they make will be identified (Kennedy & Uphoff, 1939; Rosenthal, 1969). Conditioning of bias is also attributed to unconscious processes since people are not aware that they are being conditioned or that they are emitting cues that serve to condition their subjects/subordinates. For example, the experimenters working with the horse, Clever Hans, were not aware that they were providing cues to the horse about how to respond correctly (Pfungst, 1911; Rosenthal, 1966; for review, see Rosenthal, 1969; Chaikin et al., 1974).

Even with respect to interpreting results and making decisions, processes which might at least seem to be within the realm of conscious awareness, “... it is difficult to avoid the subconscious tendency to reject ... data which weaken a hypothesis while uncritically accepting ... data which strengthen it” (Kety, 1959). Predecision processing restructures mental representations of the decision environment to favor one alternative before a decision is made (Brownstein, 2003). Decision makers develop a rationale and pull together information that supports and justifies that position, and they are not necessarily aware that they could just as easily put together an equally compelling argument to support the opposite position (Kunda, 1990).

Finally, the argument that the effects of bias cannot be attributed to conscious, willful fraud and deceit, is supported by the fact that the effects of bias are largest when expectations are shaped implicitly, using subtle powers of suggestion. When experimenters are offered bribes and other inducements and are explicitly asked to commit fraud and alter their results, they tend to resist this kind of overt pressure and show little or no effects of bias (Rosenthal, 1964b; Rosenthal et al., 1964). Thus, from the lowest level sensory/perceptual processes all the way up through the highest cognitive functions—interpreting results and making decisions about how to proceed—every step in the process is vulnerable to the effects of bias, outside of conscious awareness and control.

### *7.2. Even successful, experienced and productive scientists are vulnerable to bias*

It might be assumed that more experienced investigators would be less vulnerable to bias, but more experienced investigators are also higher in status, dominance, and usually behave more professionally—they appear more focused, oriented and competent and feel more confident about their ability to predict relationships between variables. More experienced investigators are therefore more likely to be biased themselves and to produce more bias in their subordinates and associates (Rosenthal et al., 1963; Friedman et al., 1965;

Rosenthal, 1966, pp. 362, 409–410; Rosenthal, 1969). As 1 example, in a study of child behavior modification, experimenter bias was more evident in the most senior investigator, despite the fact that this investigator was well aware of the phenomena of experimenter bias (Kent et al., 1974).

One classic example of how frequently the expectations or biases of the experimenter can unwittingly affect the experimental results, even among the most reputable scientists in the field, was reported by Rostand, who wrote that, “In 1903, Blondlot discovered ‘N-rays,’ which appeared to make reflected light more intense. This phenomenon was viewed by a great many observers, including many famous scientists of the day. Only a few were unable to detect the phenomenon.” While visiting Blondlot’s laboratory, R.W. Wood secretly removed a crucial part of the apparatus and looked on as Blondlot continued to “see” the effects, proving that these effects were really, “... a colossal compounded observer error.” After this incident was reported (Wood, 1904), the effects of “N-rays” could no longer be observed (Rostand, 1960; Collins, 1992, p. 45).

Many cases of bias and self deception have been discovered, even in the “hard” sciences such as physics, some of them corroborated in hundreds of publications even by the most prestigious scientists in their fields, including members of the national academies of sciences (Langmuir & Hall, 1989). The fact that Galileo and Pascal reported experiments that they had never conducted has been mentioned above. Galileo even boasted that he didn’t need to conduct experiments to verify his hypotheses—he felt he was just that good (Koyre, 1943).

Ivan Pavlov, a Nobel prize winner and one of the greatest figures in the history of psychology, believed that conditioned reflexes can be inherited, and he published a paper that supported this expectation (Pavlov, 1923). White mice that had been conditioned to run to their feeding place at the sound of a bell acquired the task in 300 trials, the second generation acquired the task in 100 trials, the third generation in 30 trials, the fourth generation in 10 trials, and the fifth generation in only 5 trials. Pavlov later reported that he was unable to replicate those results, and it has been suggested that they may have been generated by a research associate that was too eager to please (Zirkle, 1958; Razran, 1959).

Isaac Newton, one of the most influential scientists in history, is now known to have “fudged” many of his calculations in order to exaggerate their apparent precision (Westfall, 1973). Based in part on an extensive collection of correspondence between Newton and his editor, Newton maximized the precision of his calculations by adding new parameters and adjusting estimates of the maximum radius of the earth, the distance of the moon from the earth, and the density of air, until the solutions to his equations very closely approximated the expected values. His bias in these efforts is made apparent by the fact that his estimates of the values for individual parameters, such as the distance of the moon from the earth, were not based on attempts to most accurately estimate these individual parameters, but to select the value that produced the solution to the overall equation that most closely approximated the expected value.

Gregor Mendel, revered today as “the father of genetics,” crossed plants with different characteristics and observed the

offspring. He determined that physical characteristics were inherited, but his recorded results were too close to the results he predicted to have been collected experimentally (Fisher, 1936). These effects have been attributed to bias. Apparently, Mendel may have excluded results that varied too far from those he expected (Dunn, 1965; Wright, 1966; cf. Zuckerman, 1977).

Robert Millikan, another Nobel prize winner and perhaps one of the most renowned and influential scientists from the United States, developed an experimental apparatus and procedure that he used to calculate very precisely the charge of an electron, a finding which proved that electrons were finite quantities. While this discovery still stands as a remarkable scientific achievement, analyses of his original notebooks has shown that he routinely recalculated values and discarded data that did not fit with his expectations (Holton, 1978; Franklin, 1981).

The point of these examples is not to suggest that these great scientists were unethical or that they committed fraud, but simply to point out that even the greatest figures in the history of science were not necessarily objective. Even some of the most brilliant, prestigious and productive scientists had biases which affected their results. These cases simply drive home the point that experimenter bias is not dependent on a lack of experience, prestige, or professional integrity.

## 8. Target-based discovery may increase bias in preclinical assessments of potential efficacy

While bias has even affected the results of some of the greatest figures in the history of science, many of them conducted their research years ago, long before the development of the modern scientific and technological advances we now enjoy. It might be assumed that scientists may not have been aware of the problem of bias until more recently, and that modern research practices have improved to adequately control the effects of potential bias. However, bias has been recognized as a major problem from the early 1600s all the way up to modern times (Bacon, 1620; Babbage, 1830; Bernard, 1865; Kuhn, 1996), and the development of technologies and equipment that allow precise quantification, and the development of rigorous, standardized, experimental procedures may have created the illusion that research findings are completely objective (Reiser, 1993), but there is no evidence that more complex, technologically advanced research is any less vulnerable to the effects of bias. In fact, each level of complexity usually requires additional steps for quality control and other types of data interpretation and decision-making, and each of these steps is vulnerable to potential bias, which means that more complex, modern research practices may actually be more vulnerable to the effects of bias. Not only has research generally become more complex, but drug discovery efforts in particular have changed in other ways that increase the risk of bias.

In the past, drug discovery and development was heavily dependent on time-consuming and laborious *in vivo* studies, sometimes conducted with novel compounds on a trial and error basis until therapeutic potential was detected in a preclinical model. In the last 10–15 years, tremendous progress has been made in a number of areas relevant to drug discovery and

development. For example, gene-sequencing machines were developed, the human genome project was completed, and gene expression profiling techniques became available which simultaneously determine the gene expression pattern of the entire genome. At the same time, compound libraries have grown in size and chemical diversity, and automated assay procedures have been developed that allow hundreds of thousands of compounds to be quickly tested for binding and/or functional effects on the target. X-ray crystallography and computer automated 3-dimensional modeling technologies have also been developed to facilitate the design of compounds that precisely fit the binding site.

These scientific and technological advances have led to target-based drug discovery in which a molecular target related to a disease is selected at the beginning of the project, and hits can be quickly identified and optimized. Naturally, it was assumed that these changes would dramatically increase the speed and efficiency of drug discovery. Surprisingly, the cost of developing a new drug has increased to record levels and continues to grow at an annual rate of 7.4% above inflation (Dickson & Gagnon, 2004), and despite increased spending on research and development, the percentage of treatments that are being approved for clinical use has continued to decline. Although increased regulatory and safety requirements account for some of the increase in cost and higher drop-out rates, there is a general consensus that the productivity of discovery and development processes have been seriously declining (DiMasi, 2001; Tollman et al., 2001; DiMasi et al., 2003; FDA, 2004; Sams-Dodd, 2005).

Clearly, target-based discovery procedures have not yet been utilized in a way that capitalizes on their potential. One reason could be that the shift to target-based discovery has been accompanied by a concomitant increase in bias during preclinical assessments of potential efficacy. Proof of concept studies that used to be conducted fairly early, before strong attachments to individual targets had developed, are now conducted at the end of the lead optimization phase, 3–5 years into the program, at a point when considerable time and resources have already been invested. Proof of concept studies also used to be conducted under the authority of those with expertise with the preclinical models, and while there was naturally interest in identifying treatments with therapeutic potential, assessments of potential efficacy are now conducted under the direction of teams focused on the specific target of interest. Having successfully overcome numerous obstacles, these teams have usually developed a commitment to the target and a personal identification with, and a position of advocacy for, their program (Szymkowski, 2001; Sams-Dodd, 2005).

With target-based discovery programs, by the time potential efficacy is assessed, the teams directing the assessments tend to view the need to demonstrate preclinical efficacy as another obstacle to overcome in order to advance into clinical trials. At the same time, organizations now also use quotas to ensure productivity, requiring a defined number of programs to transition from exploratory, to early phase, to full phase and up into clinical trials. Since resources are tied up in target-based programs sometimes for years before proof of concept studies

are conducted, this means that there are a limited number of programs available for potential transitions, and high quotas produce a well-known “throw it over the wall” phenomenon where each group tosses less than optimal products “over the wall” to the next group in the development line in order to meet its quota (Cohen, 2003).

In addition, preclinical drug discovery personnel advance their careers largely based on the number of compounds they shepherd into clinical trials. Given the complexity and enormity of the task, simply getting a compound into clinical trials is an incredible achievement, and since overall success rates are so low and the delay between entry into clinical testing and final determination of actual efficacy is now so long, whether a particular compound is actually found to be safe and effective in the clinic has little effect on evaluations of preclinical drug discovery efforts. Altogether, with the shift to target-based discovery, the changes in the timing and the conditions under which preclinical proof of concept studies are now being conducted, dramatically increase the risk that experimenter bias might affect every step in the evaluation and decision-making process.

This possibility is supported by a considerable body of research which shows that the motivation to reach specific conclusions enhances the selective use of information and strategies that support the preferred outcome. People are more likely to arrive at conclusions that they want to arrive at (for review, see Kunda, 1990). After convincing themselves that their target should be advanced into clinical trials, drug discovery teams can then persuade their superiors to arrive at the same conclusion, using the same biased arguments and data sets. In part, this sequence of events is made possible because people are fairly insensitive to sampling bias, even if the biased nature of the information is obvious to them (for review, see Nisbett & Borgida, 1975; Ross et al., 1977; Kruglanski, 1983).

## **9. Controlling bias in assessments of potential efficacy could increase the productivity of drug discovery efforts**

The shift to target-based discovery may have increased bias in the preclinical assessment of potential efficacy, but preclinical research was conducted without any special safeguards to prevent experimenter bias even before the advent of target-based discovery. Therefore, compounds with little or no therapeutic potential may have already been advancing into the clinic, and this trend may have simply increased with the shift to target-based discovery. This possibility is supported by the fact that clinical success rates were already low, and the single biggest reason why drugs failed in clinical trials, historically, was lack of efficacy.

Clinical success rates have continued to decline over the last 10–15 years with the shift to target-based discovery, down to very low levels with the most recent novel targets (Fig. 9), and the percentage of clinical failures due to lack of efficacy has also continued to increase (Fig. 10). As clinical success rates continue to decline, the cost of development continues to increase, up to US\$800 million (Fig. 11), with some estimates as high as US\$1.2–1.6 Bn per drug (Windhover, Bain drug economics model, 2003, cf. Gilbert et al., 2003), and the return

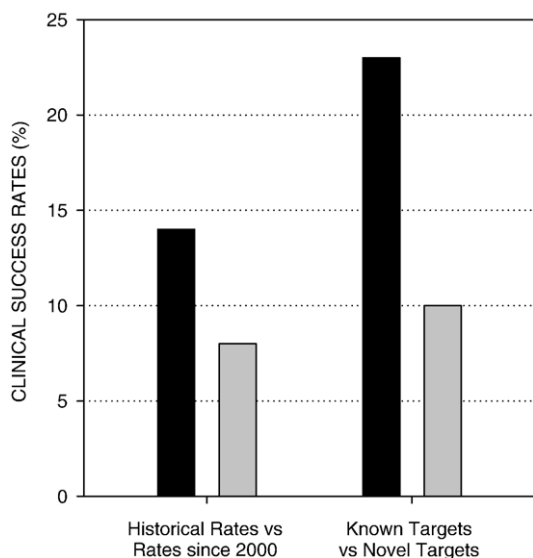


Fig. 9. Declining clinical success rates, especially with most recent targets (e.g., historical rates vs. rates since 2000; *CMR Internationale*, 2006; Windhover, Bain drug economics model, 2003, cf. *Gilbert et al.*, 2003) and novel targets (e.g., known targets vs. novel targets; *FDA*, 2004).

clinical success rates, and ultimately, reduce costs and increase productivity of preclinical drug discovery and development.

If the clinical success rate for novel targets could be increased by 10%, to the rate associated with known targets, it would be accompanied by a 25% decrease in the costs of discovery and development for each marketed drug (*DiMasi*, 2002). Since preclinical experimenter bias may have already accounted for some clinical failures even before the shift to target-based discovery, reducing experimenter bias might increase clinical success rates 10% above historical success rates, which would result in another 25% reduction in costs per drug (*Tollman et al.*, 2001). Clearly, even modest increases in clinical success rates could substantially reduce the costs and increase the productivity of drug discovery and development (*Fig. 12*).

More carefully controlling for bias during preclinical assessments of potential efficacy might improve clinical success rates in some therapeutic areas more than others. For example, lack of efficacy is not a problem with anti-infectives, only about 2% of anti-infectives fail in the clinic due to lack of efficacy. Other therapeutic areas have an average failure rate due to lack of efficacy as high as 46%, and those therapeutic areas, such as psychiatric and neurological disorders, are most likely to benefit from more carefully controlled preclinical assessments of potential efficacy (*Kennedy*, 1997; *DiMasi*, 2001).

#### 10. Limiting experimenter bias in preclinical assessments of potential efficacy

There is an assumption that simply making experimenters aware of the phenomenon is all that is needed to guard against bias, but this assumption is not supported by the evidence (*Rosenthal*, 1966, pp. 362–364). For example, experimenters that understood the problem of experimenter bias, knew that they were being monitored, and felt that their expectations had not affected their performance, nevertheless exhibited biases based on their expectations (*Troffer & Tart*, 1964).

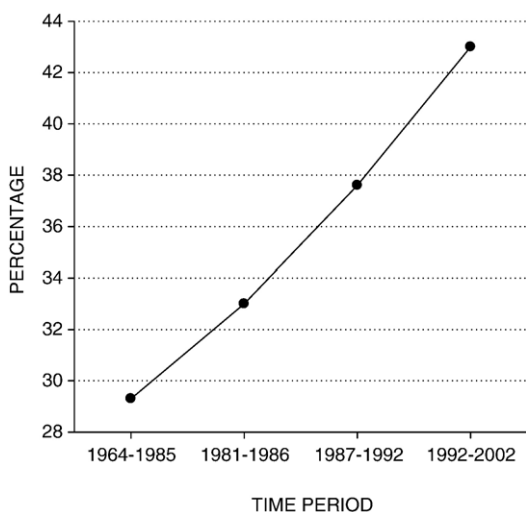


Fig. 10. Increasing percentage of clinical failure rates due to lack of efficacy, from (*Prentis et al.*, 1988; *Kennedy*, 1997; *DiMasi*, 2001; *Schuster et al.*, 2005).

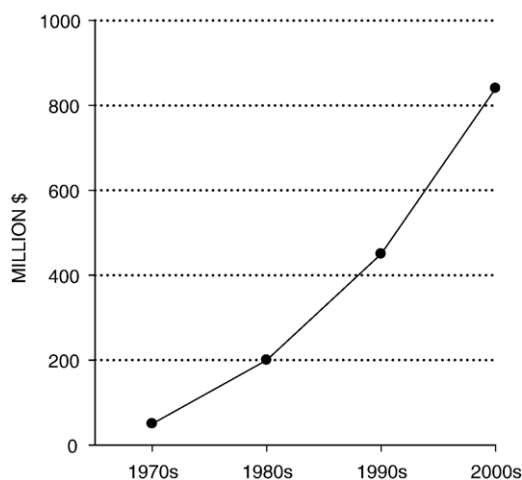


Fig. 11. Increasing cost of development per drug, from (*Niblack*, 1997; *DiMasi*, 2001; *Tollman et al.*, 2001; *DiMasi et al.*, 2003; *Dickson & Gagnon*, 2004; *FDA*, 2004; *CMR Internationale*, 2006).

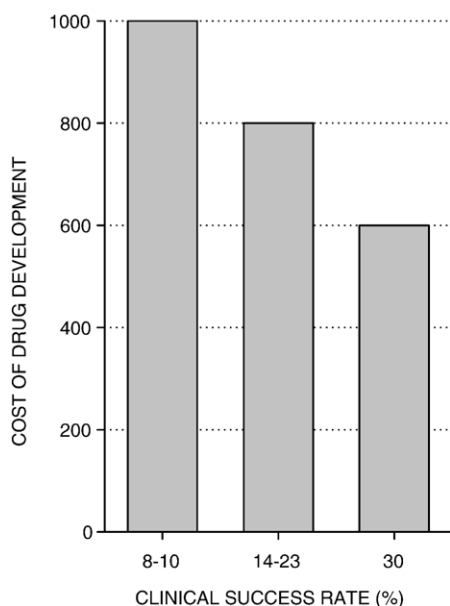


Fig. 12. Differences in clinical success rates and the costs of drug development in millions of dollars (Tollman et al., 2001; DiMasi, 2002).

### 10.1. Creating and maintaining a culture that controls for potential bias

However, it cannot be assumed that simply requiring the use of the relevant procedures will necessarily control for potential bias. It is possible to break blinds, and there is often considerable motivation to do so (Zifferblatt & Wilbur, 1978). More than half the people involved with conducting clinical trials are aware of at least 1 case in which procedures to conceal treatment allocation were overcome by those involved with conducting the trials. Patients have broken blinds by chewing and tasting their capsules (Karlowski et al., 1975; Miller et al., 1977), and investigators have opened or transilluminated envelopes and rifled through offices and desks to reveal treatment allocations (Carleton et al., 1960; Schulz, 1995).

Investigators may actively work to overcome protections against potential bias because they do not understand why bias is a problem and why it is important to control. Both clinical and preclinical investigators may believe that it is important to efficiently meet their objectives, and “efficiency” in the drug discovery process is sometimes defined as production of the “desired result” with a minimum expenditure of energy, which means getting a drug approved based on only the minimal amount of work necessary to get approval, rather than conducting a thorough assessment of whether the treatment is truly effective (Moertel, 1984). Investigators may also believe that their results are more a reflection of their abilities to detect treatment effects than a test of the therapeutic potential of the treatment under study. They may assume that if their results fail to detect efficacy, it will reflect poorly on the sensitivity of their measurements and their abilities and expertise as investigators. This concern is supported by the tendency to “shoot the messenger,” and to assume that undesirable research findings are attributed to the investigator’s personal disposition or incompe-

tence rather than a lack of therapeutic effect of the treatment under study (MacCoun, 1998).

In addition, investigators are also often aware of actual treatment allocations, despite the use of blinding procedures, even without making any effort to overcome them (Shapiro & Shapiro, 1997, pp. 191–201). For example, in some studies, simply based on their observations of the physical properties of the pills or other drug forms that were being administered, physicians were able to accurately differentiate between active treatment and placebo controls as much as 85–100% of the time, even though they were supposed to be “identical matching controls” (Blumenthal et al., 1974). Preclinical drug studies may also use coded solutions that are injected by the investigator, but coded solutions can often be discriminated based on their opacity and color. With this kind of blinding procedure, if the code is broken for 1 sample/animal, then the code is also automatically broken for all other samples/animals in that same treatment group. Inert placebos or vehicle controls can also serve to essentially break the blind if side effects of the active agents are not observed in the controls (Blumenthal et al., 1974; Zifferblatt & Wilbur, 1978). For example, preclinical researchers that are blinded can still determine which animals were treated with the active agent by observing side effects such as hyper- or hypoactivity, piloerection, or stereotypies.

Procedures that control for bias are often ineffective or can easily be overcome, but investigators usually will not admit that they are aware of the treatment allocations (Zifferblatt & Wilbur, 1978). Naturally, this makes it very difficult to control for the effects of bias, but this problem can be reduced by making sure that everyone involved is properly trained and supervised (Schulz, 1995). As Rosenthal (1966, pp. 35–36) has pointed out, when supervisors focus on the results investigators obtain, this quite naturally puts pressure on the investigators to produce a desired or expected result, and increases the risk that experimenter bias may affect the results. On the other hand, when investigators are made to feel that it is important to follow rigorous, controlled procedures and to obtain valid results, they obtain valid results regardless of what they expect (Adler, 1968; cf. Rosenthal, 1969). Clearly, it is critical to establish and maintain a culture that primarily values valid, accurate assessments, instead of emphasizing the desire to demonstrate efficacy.

### 10.2. Procedures to limit bias

As reviewed above, the FDA recognizes the impact that bias can have, and has mandated that efficacy be assessed almost entirely based on the results of clinical trials that carefully control for the effects of potential bias (Edwards, 1970). Clinical research has demonstrated that bias can be controlled (Friedman et al., 1998, pp. 10–11), and for the most part, the same controls should also be effective in preclinical proof of concept studies. These procedures will not be reviewed in detail here; but briefly, concealing treatment allocation, identifying exclusion criteria, conducting planned analyses only on the primary endpoints, and establishing a decision-tree for how different results will be interpreted before conducting the study, all help control for the effects of potential bias and also reduce

the risk of detecting spurious effects due to “data dredging” or “fishing” (Gilovich, 1991; May et al., 1981; Rose, 1982; Pocock, 1997; Friedman et al., 1998, p. 305; Pocock et al., 2002).

### 10.3. Objective decision-making

Bias can be controlled in preclinical proof of concept studies with the same kinds of procedures used to control the effects of bias in clinical trials, but these procedures control for bias only within the experiments in which they are used. Additional procedures must be used to control for bias in the larger decision-making process that determines whether compounds and programs should be advanced into clinical trials.

For example, in target-based drug discovery programs, potential efficacy is often assessed by members of the same team or group that are advocating for the advancement of their compound/target into the clinic. As advocates for their compounds and programs, it should be expected that these groups would selectively use and emphasize information that supports advancing their compound into the clinic (MacCoun, 1998). People motivated to arrive at a particular conclusion pull together information and produce an argument or presentation that justifies the conclusion they want to make, and they do it in a way that creates an “illusion of objectivity.” Every level in the hierarchical decision-making process is then given the same biased presentation, and decisions are made by people that are not aware that the presentation is biased (Pyszczynski & Greenberg, 1987), but even if they know the presentation is biased, the evidence suggests that it is virtually impossible to make adjustments in the decision-making process that adequately compensate for the effects of biased presentations (for review, see Nisbett & Borgida, 1975; Ross et al., 1977; Nisbett & Ross, 1980; Wilson & Brekke, 1994). Therefore, it is critical that the entire decision-making process be conducted in a way that protects against the effects of potential bias.

First, the potential efficacy of novel targets should be assessed by people that are primarily motivated to be accurate and to draw correct conclusions, not by people that are motivated to advance their compound into clinical trials. Emphasis on accuracy increases the duration and depth of processing and evaluation, and encourages use of the most appropriate and effective strategies and procedures (Kruglanski, 1983; for review, see Kunda, 1990). Groups also make better decisions when group members are more critical and analytical than when they are trying to reach a unanimous consensus (Postmes et al., 2001). In addition, if experts are selected to provide guidance, they should be selected based on their expertise in the field, not because they are advocates for the compound/program being evaluated, and they should be selected to represent a range of opinions in the field—some in favor and some opposed to the specific target of interest, in order to ensure a thorough evaluation.

Another important aspect of the decision-making process is to ensure that decisions are based on the right information. Assessments of potential efficacy in clinical trials are based almost entirely on the results of only those trials that have most

carefully controlled for the effects of bias (Green & Byar, 1984; Sackett, 1986; Juni et al., 2001). Preclinical assessments of potential efficacy might also be optimized by focusing on the results of studies that have very carefully controlled for the effects of potential bias. In addition, there is always at least a 5% chance that significant effects will be detected, even in the most well-controlled studies, and the results of these studies can be selectively pulled together to produce a presentation that suggests a particular compound/target/program should be advanced into clinical trials. These problems can be addressed by assigning group members to be responsible for different categories of information and splitting the decision task into 2 components—searching for and pulling together all the relevant information, and then integrating the information and making decisions (Kerr & Tindale, 2004).

Even if advocates for the compound/target under evaluation are not conducting the evaluation of potential efficacy, research has shown that decisions based on the experience of professional scientists, even among seasoned experts, are unreliable, because they are subject to the same biases and inferential failings as laymen (Nisbett & Ross, 1980; Sackett, 1986; Dawes et al., 1989). Decision-making skills can be taught, and formal training, especially in experimental design and statistical analyses, can help improve inferential abilities. For example, education in probabilistic disciplines such as psychology and medicine improve statistical and methodological reasoning and conditional logic, and provide training in how to deal with issues like false positives and negatives and the importance of considering potential confounding factors. Non-probabilistic professions such as chemistry and law do not address these issues (Nisbett & Ross, 1980; Nisbett et al., 1987; Lehman et al., 1988; Gilovich, 1991, pp. 190–192). Training in probabilistic sciences should be provided to the medicinal chemists and molecular biologists that often dominate the management of preclinical, target-based drug discovery efforts, since such training is not usually part of their formal education and training (Klayman & Brown, 1993).

Training in probabilistic sciences often results in the development and use of formal decision trees (Dawes et al., 1989). Simple statistical models for combining sets of empirical evidence consistently provide more accurate predictions than judgments by clinicians or other experts (Sackett, 1986; Grove et al., 2000; Grove & Lloyd, 2006). Virtually any type of data is amenable to the use of formal, mechanical decision trees, and such mechanical decision-making procedures have substantially outperformed the judgment of experts regardless of the task, amount of experience, or the types of data being examined. Adding expert judgment fails to improve over formal decision trees even when experts have access to extra information, because experts are unable to exercise proper restraint in overriding the formal decision-making process (Dawes et al., 1989; Grove et al., 2000; Grove & Lloyd, 2006).

Evidence-based medicine is 1 example of the use of formal decision trees (Thagard, 1999, p. 188). For example, students usually learn how to make diagnoses by studying the pattern of symptoms associated with each disease, but the accuracy of diagnoses can be improved by teaching students how to

differentiate between different diseases (Klayman & Brown, 1993). The same kinds of strategies could be applied to the complicated decision-making process in drug discovery and development. For example, proof of principle studies should be conducted after a consensus has been reached on how the results will be interpreted. This will help avoid problems with hindsight bias and explaining away data that is inconsistent with the desired outcome.

## 11. Summary and conclusions

The hopes, beliefs, and expectations, altogether known as the biases, of all the parties involved in the research process can affect the results. In studies assessing the therapeutic potential of novel treatments, bias usually exaggerates the therapeutic effects of the treatment being evaluated. Clinical research has shown that the effects of bias are so common and robust that unless they are carefully controlled they should be expected. The effects of bias in preclinical studies are at least as robust as in clinical trials, but preclinical proof of concept studies rarely control for potential bias, so many compounds that actually have little or no therapeutic potential may have advanced into clinical trials. This possibility is supported by the fact that lack of efficacy is the single biggest reason why compounds fail in the clinic.

Tremendous scientific and technological advances over the last 20 years have resulted in a shift to target-based discovery, but ironically, during the same time period, clinical failure rates due to lack of efficacy and the cost of drug development have increased, and return on investment has declined to levels that do not justify the risk associated with drug discovery—the pharmaceutical industry is approaching a state of crisis. Preclinical models may not provide perfect predictive validity, but they are at least as predictive as they were 30–40 years ago, so it is unlikely that limitations in the predictive validity of the models could account for the increasing rates of clinical failures, especially due to lack of efficacy. Instead, the advantages of target-based discovery may have been offset by increased bias in the assessment of potential efficacy, in part because potential efficacy is now assessed by the same groups that are advocating for the advancement of their compound/target into clinical trials. Procedures to control for bias have been developed in clinical trials and are available for use in preclinical proof of concept studies, but additional procedures are necessary to control for bias throughout the decision-making process used to determine which compounds should be advanced into clinical trials. The basic procedures that can be used to reduce bias are summarized here:

1. *Culture*: Instead of focusing on the pattern of results that would be expected or desired, a culture must be established that emphasizes the importance of adhering to the highest, most rigorous scientific standards. It must be accepted that experimenter bias is a natural part of the human condition, and all investigators must constantly be on guard against the possibility that bias might be affecting the results in previously unforeseen ways.

2. *Procedures*: Procedures known to reduce bias should be incorporated as a matter of routine, just as they are now in clinical trials. For example, protocols should be finalized and approved before starting the experiment, and the protocol should define the most relevant, primary endpoints, exclusion criteria, and analytical procedures, and all procedures used to conceal treatment allocation and limit bias should be explicitly documented and reported.
3. *Decision-making*: Decisions should be based on results of the most carefully controlled studies, all relevant data should be included in the decision-making process not just the data that fits with the desired result, and decisions should be made by individuals who are responsible for making the most accurate decisions, with training in probability and a commitment to continuously improving the decision-making process.

In conclusion, the present review suggests that cognitive and other psychological aspects of the process of assessing preclinical efficacy have been neglected, and while procedures are available to address the problems of potential bias, they will not be utilized until bias is recognized as a legitimate problem. For that reason, this review primarily focused on the evidence that demonstrates how common and robust the effects of bias are on the entire preclinical process of assessing potential efficacy. Controlling for bias could improve clinical success rates, and even modest improvements in clinical success rates could dramatically reduce costs and increase productivity and return on investment in drug discovery.

## References

- Adler, N. E. The influence of experimenter set and subject set on the experimenter expectancy set. 1968. Unpublished AB thesis, Wellesley College.
- Amir, Y., & Sharon, I. (1990). Replication research: a “must” for the scientific advancement of psychology. *J Soc Behav Pers* 5(4), 51–69.
- Anturane re-infarction trial research group. (1980). Sulfapyrazone in the prevention of sudden death after myocardial infarction. The Anturane Reinfarction Trial Research Group. *N Engl J Med* 302(5), 250–256.
- Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J R Stat Soc Ser A* 132, 235–244.
- Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 355(9209), 1064–1069.
- Baar, J., & Tannock, I. (1989). Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. *J Clin Oncol* 7(7), 969–978.
- Babbage, C. (1830). *Decline of Science in England*, Published in the United States by the IndyPublish.com, McLean, VA.
- Bacon, F. (1620). *Novum Organum: With Other Parts of The Great Instauration*. Chicago, IL: Open Court.
- Bailey, K. R. (1991). Detecting fabrication of data in a multicenter collaborative animal study. *Control Clin Trials* 12(6), 741–752.
- Bardy, A. H. (1998). Bias in reporting clinical trials. *Br J Clin Pharmacol* 46(2), 147–150.
- Beach, M. L., & Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Control Clin Trials* 10(4 Suppl), 161S–175S.
- Beadle, G. W. (1967). Mendelism, 1965. In R. A. Brink & E. D. Styles (Eds.), *Heritage from Mendel* Madison, Milwaukee: The University of Wisconsin Press.
- Becher, H. K. (1955). The powerful placebo. *JAMA* 159(17), 1602–1606.

- Begg, C. B., & Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *J Natl Cancer Inst* 81(2), 107–115.
- Ben-David, J. (1977). Organization, social control, and cognitive change in science. In J. Ben-David & T. N. Clark (Eds.), *Culture and Its Creators: Essays in Honor of Edward Shils* Chicago, IL: The University of Chicago Press.
- Berkson, J., Magath, T. B., & Hurn, M. (1940). The error of estimate of the blood cell count as made with the hemocytometer. *Am J Physiol* 128, 309–323.
- Bernard, C. (1865). *An Introduction to the Study of Experimental Medicine*. New York: Dover Publications Inc.
- Blass, T. (1999). The Milgram paradigm after 35 years: some things we now know about obedience to authority. *J Appl Soc Psychol* 29(5), 955–978.
- Blumenthal, D. S., Burke, R., & Shapiro, A. K. (1974). The validity of “identical matching placebos”. *Arch Gen Psychiatry* 31(2), 214–215.
- Bornstein, R. F. (1990). Publication politics, experimenter bias and the replication process in social science research. *J Soc Behav Pers* 5, 71–81.
- Broad, W., & Wade, N. (1982). *Betrayers of the Truth: Fraud and Deceit in the Halls of Science*. New York, NY: Simon and Schuster.
- Brownstein, A. L. (2003). Biased predecision processing. *Psychol Bull* 129(4), 545–568.
- Bruner, J. S., & Postman, L. (1949). On the perception of incongruity; a paradigm. *J Pers* 18(2), 206–223.
- Bulthoff, I., Bulthoff, H., & Sinha, P. (1998). Top-down influences on stereoscopic depth-perception. *Nat Neurosci* 1(3), 254–257.
- Burnham, J. R. (1966). Experimenter bias and lesion labeling. Purdue University. (Unpublished manuscript).
- Campbell, K. E., & Jackson, T. T. (1979). The role and need for replication research in social psychology. *Replication Soc Psychol* 1(1), 3–14.
- Carleton, R. A., Sanders, C. A., & Burack, W. R. (1960). Heparin administration after acute myocardial infarction. *N Engl J Med* 263, 1002–1005.
- Carroll, D., Moore, R. A., Mcquay, H. J., Fairman, F., Tramer, M., & Leijon, G. (2001). Transcutaneous electrical nerve stimulation (TENS) for chronic pain. *Cochrane Database Syst Rev* 3 (CD003222).
- Chaikin, A. L., Sigler, E., & Derlega, V. J. (1974). Nonverbal mediators of teacher expectancy effects. *J Pers Soc Psychol* 30, 144–149.
- Chalmers, T. C., Celano, P., Sacks, H. S., & Smith, H., Jr. (1983). Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 309(22), 1358–1361.
- Chalmers, T. C., Matta, R. J., Smith, H., Jr, & Kunzler, A. M. (1977). Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 297(20), 1091–1096.
- Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *J Abnorm Psychology* 72(3), 193–204.
- Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *J Abnorm Psychology* 74(3), 271–280.
- Chapman, L. J. (1967). Illusory correlation in observational report. *J Verbal Learn Verbal Behav* 6, 151–155.
- Cliff, N. (1987). *Analyzing Multivariate Data*. San Diego, CA: Harcourt Brace Jovanovich.
- CMR Internationale (2006). *2006 Global R&D Performance Metrics Programme: Industry Success Rates Report*. Novellus Court, 61 South Street, Epsom, Surrey, KT18 7PX, UK: Centre for Medicines Research International Ltd.
- Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics* 10(4), 637–666.
- Cohen, C. M. (2003). A path to improved pharmaceutical productivity. *Nat Rev Drug Discov* 2(9), 751–753.
- Collins, H. M. (1992). *Changing Order: Replication and Induction in Scientific Practice*. Chicago, IL: University of Chicago Press.
- Cordaro, L., & Ison, J. R. (1963). Psychology of the scientist: X. Observer bias in classical conditioning of the planarian. *Psychol Rep* 13, 787–789.
- Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Prof Psychol* 17, 136–137.
- David, P. (1983). The system defends itself. *Nature* 303(2), 369.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science* 243(4899), 1668–1674.
- Diaconis, P. (1978). Statistical problems in ESP research. *Science* 201(4351), 131–136.
- Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H., Jr. (1987). Publication bias and clinical trials. *Control Clin Trials* 8(4), 343–353.
- Dickersin, K., & Min, Y. I. (1993). Publication bias: the problem that won't go away. *Ann NY Acad Sci* 703, 135–146.
- Dickson, M., & Gagnon, J. P. (2004). The cost of new drug discovery and development. *Discov Med* 4(22), 172–179.
- Diehl, L. F., & Perry, D. J. (1986). A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? *J Clin Oncol* 4(7), 1114–1120.
- DiMasi, J. A. (2001). Risks in new drug development: approval success rates for investigational drugs. *Clin Pharmacol Ther* 69(5), 297–307.
- DiMasi, J. A. (2002). The value of improving the productivity of the drug development process: faster times and better decisions. *Pharmacoeconomics* 20(Suppl 3), 1–10.
- DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *J Health Econ* 22(2), 151–185.
- Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: the interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Pers Soc Psychol Bull* 29(9), 1120–1132.
- Djulgobovic, B., Lacevic, M., Cantor, A., Fields, K. K., Bennett, C. L., Adams, J. R., et al. (2000). The uncertainty principle and industry-sponsored research. *Lancet* 356(9230), 635–638.
- Dunn, L. (1965). Mendel, his work, and his place in history. *Proc Am Philos Soc* 109, 189–198.
- Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet* 337(8746), 867–872.
- Edwards, C. C. (1970). Regulations describing scientific content of adequate and well-controlled clinical investigations. *Fed Regist* 35(90), 7250–7253.
- Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *J Pers Soc Psychol* 71, 5–24.
- Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: persistence of the illusion of validity. *Psychol Rev* 85(5), 395–416.
- Ellis, R. R., & Lederman, S. J. (1998). The golf-ball illusion: evidence for top-down processing in weight perception. *Perception* 27(2), 193–201.
- Ellson, D. G. (1941). Hallucinations produced by sensory conditioning. *J Exp Psychol* 28, 1–20.
- Engler, R. L., Covell, J. W., Friedman, P. J., Kitcher, P. S., & Peters, R. M. (1987). Misrepresentation and responsibility in medical research. *N Engl J Med* 317(22), 1383–1389.
- Epstein, W. V. (1996). Expectation bias in rheumatoid arthritis clinical trials. The anti-CD4 monoclonal antibody experience. *Arthritis Rheum* 39(11), 1773–1780.
- Ernst, E., & Resch, K. L. (1994). Reviewer bias: a blinded experimental study. *J Lab Clin Med* 124(2), 178–182.
- Faust, D., Hart, K., & Guilmette, T. J. (1988). Pediatric malingering: the capacity of children to fake believable deficits on neuropsychological testing. *J Consult Clin Psychol* 56(4), 578–582.
- FDA. (2004). *Innovation or stagnation: challenge and opportunity on the critical path to new medical products*. U.S. Department of Health and Human Services, FDA White Paper.
- Fischhoff, B. (1975). Hindsight ≠ foresight: the effect of outcome knowledge on judgment under uncertainty. *J Exp Psychol Hum Percept Perform* 1(289), 299.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty* New York: Cambridge University Press.
- Fischhoff, B., & Beyth, R. (1975). “I knew it would happen”—remembered probabilities of once-future things. *Organ Behav Hum Perform* 13(1), 16.
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: the appropriateness of extreme confidence. *J Exp Psychol Hum Percept Perform* 3(4), 552–564.
- Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Ann Sci* 1, 116–137.
- Fisher, R. A. (1971). *The Design of Experiments*. New York: Hafner Publishing Co.

- Fisher, S., Cole, J. O., Rickels, K., & Uhlenhuth, E. H. (1964). Drug-set interaction: the effect of expectations on drug response in outpatients. In P. B. Bradley, F. Flugel, & P. Hoch (Eds.), *Neuropsychopharmacology* New York: Elsevier.
- Fisher, S., & Greenberg, R. P. (1993). How sound is the double-blind design for evaluating psychotropic drugs? *J Nerv Ment Dis* 181(6), 345–350.
- Fleminger, S. (1992). Seeing is believing: the role of 'preconscious' perceptual processing in delusional misidentification. *Br J Psychiatry* 160, 293–303.
- Foulds, G. A. (1958). Clinical research in psychiatry. *J Ment Sci* 104(435), 259–265.
- Franklin, A. D. (1981). Millikan's published and unpublished data on oil drops. *Hist Stud Phys Sci* 11, 185–201.
- Freedman, D., Pisani, R., & Purves, R. (1980). *Statistics*. New York: W.W. Norton & Company.
- Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1998). *Fundamentals of Clinical Trials*. New York: Springer.
- Friedman, N., Kurland, D., & Rosenthal, R. (1965). Experimenter behavior as an unintended determinant of experimental results. *J Proj Tech Pers Assess* 29(4), 478–490.
- Furberg, C. D., & Morgan, T. M. (1987). Lessons from overviews of cardiovascular trials. *Stat Med* 6(3), 295–306.
- Gilbert, J., Henske, P., & Singh, A. (2003). Rebuilding big pharma's business model. *In Vivo Business Med Report, Vol. 21(10)* (pp. 1–10).
- Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York: The Free Press.
- Gold, H. (1954). How to evaluate a new drug. *Am J Med* 17, 722–727.
- Gold, H., Kwit, N. T., & Otto, H. (1937). The xanthines (theobromine and aminophylline) in the treatment of cardiac pain. *JAMA* 108, 2173–2179.
- Gotzsche, P. C. (1987). Reference bias in reports of drug trials. *Br Med J (Clin Res Ed)* 295(6599), 654–656.
- Gotzsche, P. C. (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 10(1), 31–56.
- Grace, N. D., Muench, H., & Chalmers, T. C. (1966). The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology* 50(5), 684–691.
- Green, S. B., & Byar, D. P. (1984). Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med* 3(4), 361–373.
- Greenberg, R. P., Bornstein, R. F., Greenberg, M. D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under "blinder" conditions. *J Consult Clin Psychol* 60(5), 664–669.
- Greiner, T., Gold, H., Cattell, M., Travell, J., Bakst, H., Rinzler, S. H., et al. (1950). A method for the evaluation of the effects of drugs on cardiac pain in patients with angina of effort; a study of khellin (visammin). *Am J Med* 9(2), 143–155.
- Gross, T. M., Jarvik, M. E., & Rosenblatt, M. R. (1993). Nicotine abstinence produces content-specific Stroop interference. *Psychopharmacology (Berl)* 110(3), 333–336.
- Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *J Abnorm Psychology* 115(2), 192–194.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychol Assess* 12(1), 19–30.
- Hamilton, D. P. (1990). Publishing by—and for?—the numbers. *Science* 250(4986), 1331–1332.
- Hanson, N. R. (1958). Observation. In: *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science* Cambridge: Cambridge University Press.
- Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Beverly Hills, CA: Sage Publications.
- Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *J Pers Soc Psychol* 5(4), 41–49.
- Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychol Rev* 104(1), 194–202.
- Hill, A. B. (1952). The clinical trial. *N Engl J Med* 247, 113–119.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley and Sons, Inc.
- Holton, G. (1978). Subelectrons, presuppositions, and the Millikan-Ehrenhaft dispute. *Hist Stud Phys Sci* 9, 166–224.
- James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics* 29(1), 121–130.
- Johnson, M. L. (1953). Seeing's believing. *New Biol* 15, 60–80.
- Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *J Pers Soc Psychol* 80(4), 557–571.
- Jones, P. E., & Roelofsma, P. H. (2000). The potential for social contextual and group biases in team decision-making: biases, conditions and psychological mechanisms. *Ergonomics* 43(8), 1129–1152.
- Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 323(7303), 42–46.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychol Rev* 80(4), 237–251.
- Kant, I. (1787). *Critique of Pure Reason*. New York: Barnes and Noble.
- Kapchuk, T. J. (1998). Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bull Hist Med* 72(3), 389–433.
- Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Kapikian, A. Z., Lewis, T. L., & Lynch, J. M. (1975). Ascorbic acid for the common cold, a prophylactic and therapeutic trial. *JAMA* 231(10), 1038–1042.
- Kast, E. C. (1961). Alpha-ethyltryptamine acetate in the treatment of depression, a study of the methodology of drug evaluation. *J Neuropsychiatry* 2(Suppl 1), 114–118.
- Kennedy, J. L., & Uphoff, H. F. (1939). Experiments on the nature of extra-sensory perception. *J Parapsychol* 3, 226–245.
- Kennedy, T. (1997). Managing the drug discovery/development interface. *Drug Discov Today* 2, 436–444.
- Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *J Consult Clin Psychol* 42(6), 774–780.
- Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Kerr, N. L., MacCoun, R. J., & Kramer, G. (1996). Bias in judgment: comparing individuals and groups. *Psychol Rev* 103, 687–719.
- Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annu Rev Psychol* 55, 623–655.
- Kety, S. S. (1959). Biochemical theories of schizophrenia. I. *Science* 129(3362), 1528–1532.
- Kissin, B., Charonoff, S. M., & Rosenblatt, S. M. (1968). Drug and placebo responses in chronic alcoholics. *Psychiatr Res Rep Am Psychiatr Assoc* 24, 44–60.
- Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol* 55(4), 407–410.
- Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: an alternative approach to reducing error in diagnostic (and other) judgment. *Cognition* 49(1-2), 97–122.
- Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Org Behav Hum Decis Process* 56, 28–55.
- Koran, L. M. (1975). The reliability of clinical methods, data and judgments (first of two parts). *N Engl J Med* 293(13), 642–646.
- Koran, L. M. (1975). The reliability of clinical methods, data and judgments (second of two parts). *N Engl J Med* 293(14), 695–701.
- Koshland, D. E., Jr. (1987). Fraud in science. *Science* 235(4785), 141.
- Koyre, A. (1943). Galileo and the scientific revolution of the seventeenth century. *Philos Rev* 52, 333–348.
- Koyre, A. (1956). Pascal savant. *Blaise Pascal, l'homme et l'oeuvre, Vol. 1* (pp. 259–281).
- Koyre, A. (1960). Galileo's treatise "de motu gravium": the use and abuse of imaginary experiment. *Rev Hist Sci* 13, 197–245.
- Kruglanski, A. W. (1983). Bias and error in human judgement. *Eur J Soc Psychol* 13, 1–44.
- Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. Chicago, IL: The University of Chicago Press.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychol Bull* 108(3), 480–498.
- Langmuir, I., & Hall, R. N. (1989). Pathological science. *Phys Today* 42(10), 36–48.

- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: formal discipline and thinking about everyday-life events. *Am Psychol* 43(6), 431–442.
- Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 326(7400), 1167–1170.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol* 37(11), 2098–2109.
- MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annu Rev Psychol* 49, 259–287.
- Mahoney, M. J. (1977). Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. *Cogn Ther Res* 1, 161–175.
- Mahoney, M. J. (1985). Open exchange and epistemic progress. *Am Psychol* 40(1), 29–39.
- Mahoney, M. J. (1987). Scientific publication and knowledge politics. *J Soc Behav Pers* 2(2), 165–176.
- Marcel, A. J. (1983). Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes. *Cogn Psychol* 15(2), 238–300.
- May, G. S., DeMets, D. L., Friedman, L. M., Furberg, C., & Passamani, E. (1981). The randomized clinical trial: bias in analysis. *Circulation* 64(4), 669–673.
- Medawar, P. B. (1991). Is the scientific paper a fraud? In D. Pyke (Ed.), *The Threat and The Glory: Reflections on Science and Scientists*. Oxford: Oxford University Press.
- Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Rev* 8, 193–210.
- Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: The University of Chicago Press.
- Milgram, S. (1963). Behavioral study of obedience. *J Abnorm Soc Psychol* 67, 371–378.
- Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Hum Relat* 18, 57–76.
- Miller, J. Z., Nance, W. E., Norton, J. A., Wolen, R. L., Griffith, R. S., & Rose, R. J. (1977). Therapeutic effect of vitamin C. A co-twin control study. *JAMA* 237(3), 248–251.
- Modell, W., & Houde, R. W. (1958). Factors influencing clinical evaluation of drugs: with special reference to the double-blind technique. *JAMA* 167, 190–198.
- Moerman, D. E. (1983). General medical effectiveness and human biology: placebo effects in the treatment of ulcer disease. *Med Anthropol Q* 14(4), 13–16.
- Moertel, C. G. (1984). Improving the efficiency of clinical trials: a medical perspective. *Stat Med* 3(4), 455–468.
- Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *J Soc Behav Pers* 5(4), 85–90.
- Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *J Soc Behav Pers* 8(6), 21–29.
- Niblack, J. F. (1997). Why are drug development programs growing in size and cost. A view from industry? *Food Drug Law J* 52(2), 151–154.
- Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *J Pers Soc Psychol* 32, 932–943.
- Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science* 238(482), 625–631.
- Nisbett, R. E., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R. E., Yetisir, E., & Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 44(1), 16–20.
- O’Leary, K. D., Kent, R. N., & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *J Appl Behav Anal* 8, 43–51.
- Orne, M. T. (1959). The nature of hypnosis: artifact and essence. *J Abnorm Psychology* 58(3), 277–299.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am Psychol* 17, 776–783.
- Orne, M. T. (1970). Hypnosis, motivation, and the ecological validity of the psychological experiment. In W. J. Arnold & M. M. Page (Eds.), *Nebraska Symposium on Motivation*. Lincoln: University of Nebraska Press.
- Orne, M. T., & Evans, F. J. (1965). Social control in the psychological experiment: antisocial behavior and hypnosis. *J Pers Soc Psychol* 95, 189–200.
- Orne, M. T., & Scheibe, K. E. (1964). The contribution of nondeprivation factors in the production of sensory deprivation effects: the psychology of the “panic button”. *J Abnorm Psychology* 68, 3–12.
- Owen, R. (1982). Reader bias. *JAMA* 247(18), 2533–2534.
- Pavlov, I. V. (1923). New researches on conditioned reflexes. *Science* 58, 359–361.
- Pearson, K. (1902). On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philos Trans R Soc Lond A* 198, 235–299.
- Pfungst, O. (1911). *Clever Hans (The Horse of Mr. von Osten)*. New York: Henry Holt and Company.
- Planck, M. (1949). *Scientific autobiography and other papers*, Philosophical Library, translated from German by Frank Gaynor, New York.
- Pocock, S. J. (1997). Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 18(6), 530–545.
- Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 21(19), 2917–2930.
- Pocock, S. J., Hughes, M. D., & Lee, R. J. (1987). Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 317(7), 426–432.
- Polanyi, M. (1963). The potential theory of adsorption. *Science* 141, 1010–1013.
- Popper, K. (1935). *The Logic of Scientific Discovery, Translated to English and reprinted in 1959 by Routledge Classics, New York*.
- Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *J Pers Soc Psychol* 80(6), 918–930.
- Prentis, R. A., Lis, Y., & Walker, S. R. (1988). Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985). *Br J Clin Pharmacol* 25(3), 387–396.
- Proshansky, H., & Murphy, G. (1942). The effects of reward and punishment on perception. *J Psychol* 13, 295–305.
- Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: a biased hypothesis-testing model. *Adv Exp Soc Psychol* 20, 297–340.
- Pyszczynski, T., Greenberg, J., & Holt, K. (1985). Maintaining consistency between self-serving beliefs and available data: a bias in information evaluation. *Pers Soc Psychol Bull* 11, 179–190.
- Ravnskov, U. (1995). Quotation bias in reviews of the diet-heart idea. *J Clin Epidemiol* 48(5), 713–719.
- Razran, G. (1959). Pavlov the empiricist. *Science* 130, 916–917.
- Reiser, S. J. (1993). Overlooking ethics in the search for objectivity and misconduct in science. *Acad Med* 68(9 Suppl), S84–S87.
- Resch, K. I., Ernst, E., & Garrow, J. (2000). A randomized controlled study of reviewer bias against an unconventional therapy. *J R Soc Med* 93(4), 164–167.
- Roberts, A. H., Kewman, D. G., Mercier, L., & Hovell, M. (1993). The power of nonspecific effects in healing: implications for psychosocial and biological treatments. *Clin Psychol Rev* 13, 375–391.
- Rose, G. (1982). Bias. *Br J Clin Pharmacol* 13(2), 157–162.
- Rosenthal, R. (1963). On the social psychology of the psychological experiment: the experimenter’s hypothesis as unintended determinant of experimental results. *Am Sci* 51, 261–283.
- Rosenthal, R. (1964). Experimenter outcome-orientation and the results of the psychological experiment. *Psychol Bull* 61(6), 405–412.
- Rosenthal, R. (1964). The effect of the experimenter on the results of psychological research. *Prog Exp Pers Res* 72, 79–114.
- Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton Century Crofts. New York: Division of Meredith Publishing Co.
- Rosenthal, R. (1969). Interpersonal expectations: effects of the experimenter’s hypothesis. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in Behavioral Research* New York: Academic Press.
- Rosenthal, R. (1976). Interpersonal expectancy effects: a follow-up. *Experimenter Effects in Behavioral Research: Enlarged Edition*. New York: Irvington Publishers Inc.

- Rosenthal, R. (1990). Replication in behavioral research. *J Soc Behav Pers* 5, 1–30.
- Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behav Sci* 8, 183–189.
- Rosenthal, R., Friedman, C. J., Johnson, C. A., Fode, K., Schill, T., White, R. C., et al. (1964). Variables affecting experimenter bias in a group situation. *Genet Psychol Monogr* 70, 271–296.
- Rosenthal, R., & Halas, E. S. (1962). Experimenter effect in the study of invertebrate behavior. *Psychol Rep* 11, 251–256.
- Rosenthal, R., Kohn, P., Greenfield, P. M., & Carota, N. (1966). Data desirability, experimenter expectancy, and the results of psychological research. *J Pers Soc Psychol* 3(1), 20–27.
- Rosenthal, R., & Lawson, R. (1963). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *J Psychiatr Res* 2, 61–72.
- Rosenthal, R., Persinger, G. W., Kline, L. V., & Mulry, R. C. (1963). The role of the research assistant in the mediation of experimenter bias. *J Pers* 31, 313–335.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: the first 345 studies. *Behav Brain Sci* 3, 377–415.
- Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *J Pers Soc Psychol* 35, 485–494.
- Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm. *J Pers Soc Psychol* 32(5), 880–892.
- Rostand, J. (1960). *Error and Deception in Science*. New York: Basic Books.
- Roth, J. A. (1966). Hired hand research. *Am Sociol*, 190–196.
- Sackett, D. L. (1983). Second thoughts. Proposals for the health sciences: I. Compulsory retirement for experts. *J Chronic Dis* 36(7), 545–547.
- Sackett, D. L. (1986). Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 89(2 Suppl), 2S–3S.
- Sacks, H., Chalmers, T. C., & Smith, H., Jr. (1982). Randomized versus historical controls for clinical trials. *Am J Med* 72(2), 233–240.
- Sams-Dodd, F. (2005). Target-based drug discovery: is something wrong? *Drug Discov Today* 10(2), 139–147.
- Sanford, R. N. (1936). The effects of abstinence from food upon imaginal processes: a preliminary experiment. *J Psychol* 2, 129–136.
- Sarter, M., Hagan, J., & Dudchenko, P. (1992). Behavioral screening for cognition enhancers: from indiscriminate to valid testing: Part I. *Psychopharmacology (Berl)* 107(2–3), 144–159.
- Sarter, M., Hagan, J. J., & Dudchenko, P. (1992). Behavioral screening for cognition enhancers: from indiscriminate to valid testing: Part II. *Psychopharmacology (Berl)* 107, 461–473.
- Schechter, P. J., Freidewald, W. T., Bronzert, D. A., Raff, M. S., & Henkin, R. I. (1972). Idiopathic hypoguesia: a description of the syndrome and a single-blind study with zinc sulfate. *Int Rev Neurobiol (Suppl. D)*, 125–140.
- Schulz, K. F. (1995). Subverting randomization in controlled trials. *JAMA* 274(18), 1456–1458.
- Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273(5), 408–412.
- Schuster, D., Laggner, C., & Langer, T. (2005). Why drugs fail—a study on side effects in new chemical entities. *Curr Pharm Des* 11(27), 3545–3559.
- Shafer, R., & Murphy, G. (1943). The role of autism in a visual figure-ground relationship. *J Exp Psychol* 32, 335–343.
- Shapiro, A. K., & Shapiro, E. (1997). *The Powerful Placebo: From Ancient Priest to Modern Physician*. Baltimore: The Johns Hopkins University Press.
- Sheffield, F. D., Kaufman, R. S., & Rhine, J. B. (1952). A PK experiment at Yale starts a controversy. *J Am Soc Psych Res* 46, 111–117.
- Shor, R. E. (1964). A note on shock tolerances of real and simulating hypnotic subjects. *Int J Clin Exp Hypn* 12(4), 258–262.
- Silverman, W. A. (1991). Suspended judgment. Is the scientific paper a fraud? *Control Clin Trials* 12(2), 273–276.
- Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 4(10), 1529–1541.
- Simes, R. J. (1987). Confronting publication bias: a cohort design for meta-analysis. *Stat Med* 6(1), 11–29.
- Snow, P. J. (1965). Effect of propranolol in myocardial infarction. *Lancet* 286 (7412), 551–553.
- Snyder, M. (1984). When belief creates reality. *Adv Exp Soc Psychol* 18, 248–305.
- Sommer, R. (1959). The new look on the witness stand. *Can Psychol* 8(4), 94–99.
- Staw, B. M. (1976). Knee deep in the big muddy: a study of escalating commitment to a chosen course of action. *Organ Behav Hum Perform* 16, 27–44.
- Stephens, J. M. (1936). The perception of small differences as affected by self interest. *Am J Psychol* 48, 480–484.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from test of significance- or vice versa. *JASA* 54, 30–34.
- Sushinsky, L. W., & Wener, R. (1975). Distorting judgments of mental health. *J Nerv Ment Dis* 161(2), 82–89.
- Szymkowski, D. E. (2001). Too many targets, not enough target validation. *Drug Discov Today* 6(8), 397.
- Thagard, P. (1999). *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.
- Tollman, P., Philippe, G., Altshuler, J., Flanagan, A., & Steiner, M. (2001). *A Revolution in R&D: How Genomics and Genetics are Transforming the Biopharmaceutical Industry*. Boston, MA: The Boston Consulting Group.
- Troffer, S. A., & Tart, C. T. (1964). Experimenter bias in hypnotist performance. *Science* 145, 1330–1331.
- Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science* 198(4318), 679–684.
- Turnbull, C. M. (1961). Some observations regarding the experiences and behavior of the BaMbuti pygmies. *Am J Psychol* 74, 304–308.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science* 185, 1124–1131.
- Vandenbroucke, J. P. (1998). 175th anniversary lecture. Medical journals and the shaping of medical knowledge. *Lancet* 352(9145), 2001–2006.
- Viamontes, J. A. (1972). Review of drug effectiveness in the treatment of alcoholism. *Am J Psychiatry* 128(12), 1570–1571.
- Waters, A. J., & Feyerabend, C. (2000). Determinants and effects of attentional bias in smokers. *Psychol Addict Behav* 14(2), 111–120.
- Weber, M. (1946). In H. H. Gerth & C. W. Mills (Eds.), *Science as a Vocation. In: From Max Weber: Essays in Sociology*. New York: Oxford University Press.
- Westfall, R. S. (1973). Newton and the fudge factor. *Science* 179, 751–758.
- Whyte, G. (1993). Escalating commitment in individual and group decision making: a prospect theory approach. *Org Behav Hum Decis Process* 54, 430–455.
- Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychol Bull* 116(1), 117–142.
- Wilson, T. D., DePaulo, B. M., Mook, D. G., & Klaaren, K. J. (1993). Scientist's evaluations of research: the biasing effects of the importance of the topic. *Psychol Sci* 4, 322–325.
- Wood, R. W. (1904). The n-rays. *Nature* 70(1822), 530–531.
- Wright, S. (1966). Mendel's ratios. In C. Stern & E. Sherwood (Eds.), *The Origin of Genetics: A Mendel Source Book*. San Francisco, CA: W.H. Freeman.
- Wyer, R. S., & Frey, D. (1983). The effects of feedback about self and others on the recall and judgments of feedback-relevant information. *J Exp Soc Psychol* 19, 540–559.
- Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 266(1), 93–98.
- Zifferblatt, S. M., & Wilbur, C. S. (1978). A psychological perspective for double-blind trials. *Clin Pharmacol Ther* 23, 1–10.
- Zirkle, C. (1958). Pavlov's beliefs. *Science* 128(3337), 1476.
- Zuckerman, H. (1977). Deviant behavior and social control in science. In E. Sagarin (Ed.), *Deviance and Social Change*. London: SAGE Publications.