This assumption follows from, but is *not* dependent upon the following two conditions: (a) all *z*s above some *c* (e.g., 1.96) are reported in the literature, and (b) all reports in the literature contain *z*s above some *c* (e.g., 1.96).

Condition b appears to be reasonably accurate (e.g., 37 of the 42 studies we reported have *z*s above 1.96). Condition a is probably not very accurate because studies can be rejected for reasons other than small *z*s.

The alternative to the null hypothesis is that (because there is some effect of therapy) the distribution of *z*s—whatever it is—is centered to the right of 0 and hence the *z*s will be larger than predicted by the null hypothesis.

To test this null hypothesis, we constructed 200 random samples of *one* *z*-value greater than 1.96 from each of our 37 studies reporting at least one value that large. Thus, there was no within-study dependence between *z*s (and no reason to expect between-study dependence). The average of the average *z*s was 2.77, not 2.34 as predicted by the null hypothesis. Because the variance of the normal truncated above 1.96 is .14, the test *z* comparing 2.77 to 2.34 is 7.56 (.43 divided by $(.14/37)^{1/2}$). *p* is virtually 0. Similar results are found with cut points of 1.65, 2.33, and 2.58.

Unfortunately, the results counterindicate going "backwards" from the hypothesized tail to infer the location of the hypothesized mean. The reason is that all the variances of the *z*s actually computed are four to six times larger than those based on normal curve theory. All we can do is reject—soundly—the null hypothesis, without introducing the "small enough" ambiguity of the Rosenthal method. The discrepancy between theoretical and observed variances mitigates against any normal curve "correction" of effect size.

Sampling independent *z*s above 1.96 thus led to a mean very significantly larger than 2.34, the expected value if we were sampling *z*s above 1.96 from a unit normal distribution. Our conclusion is that we are *not* randomly sampling from a truncated normal. Specifically, the *z*s are larger. The significant effects of psychotherapy cannot be accounted for by selective reporting unless there is an additional bias that the larger the *z* beyond the standard significance level the more likely it is to be reported. If such a bias existed, we would expect the results above 2.58 to be nonexistent, or at least weaker than those above 1.96. But they are not (test $z = 7.42$, *p* virtually 0).

Our basic assumption is quite broad. In fact, when *c* approaches $-\infty$, it is the standard assumption underlying normal distribution significance tests. All we have done is to apply the same logic to a (truncated) part of that distribution.

REFERENCES

Kurosawa, K. (1984). Meta-analysis and selective publication bias [Comment]. *American Psychologist, 38,* 73–74.

Landman, J. T., & Dawes, R. M. (1984). Reply to Orwin and Cordray [Comment]. *American Psychologist, 38,* 72–73.

Rosenthal, R. (1979). The "file drawer" problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

The Citation Bias: Fad and Fashion in the Judgment and Decision Literature

Jay J. J. Christensen-Szalanski
University of Arizona

Lee Roy Beach
University of Washington

The usual procedure for research on the psychology of judgment and decision making consists of presenting subjects with a task for which some model from logic, probability theory, or the like prescribes a "correct" answer. The existence of a discrepancy between a subject's answer and the prescriptions of the model, and the size of that discrepancy, defines the adequacy of the judgmental or decision performance.

As one might expect, there usually is a discrepancy between what the subject does and what the model prescribes. The question is how large that discrepancy has to be before one concludes that the subject's performance is inadequate. In many cases, such conclusions depend a good deal on how the experimenter elects to interpret the data. As a result, the conclusions contained in the literature are quite mixed. There are numerous examples of poor performance (Fischhoff, 1975; Nisbett & Borgida, 1975), but there are also cases of good performance (Christensen-Szalanski, Diehr, Bushyhead, & Wood, 1982; Muchinsky & Fitch, 1975; Phelps & Shanteau, 1978), as well as cases in which it is difficult to really know (Bar-Hillel, 1979; Thorngate, 1980; see also Loftus & Beach, 1982; Christensen-Szalanski, in press).

The problem is that if one were to examine summaries of this research, either comprehensive summaries (e.g., Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980) or the introductions to journal articles, one would receive the distinct impression that the literature is not mixed at all, that the findings clearly show human judgment and decision making to be hopelessly inadequate. The unfortunate result of this is that persons who rely on such summaries and who do not closely examine the research and results, tend to accept the negative pronouncements as fact and to generalize them without hesitation (Berkeley & Humphreys, 1982).

It is our hypothesis that the widely held belief in the hopelessness of human judgment and decision performance results in part from the fact that only evidence to that effect gets much attention. That is, evidence for poor performance is cited more frequently than is evidence for good performance.

In the spirit of egocentric biases, the availability bias, the hindsight bias, the anchoring bias, and all the rest, we shall call selectivity in the citing of evidence the *citation bias.*[1] In what follows, we provide evidence for the existence of such a bias and examine the possible effects that the bias might have on the views of readers of the literature.

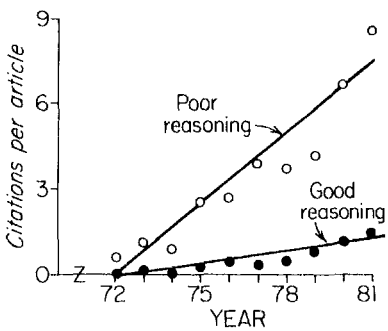
We start by comparing the citation frequencies of a representative set of published articles that reported either good or poor performance. To obtain these sets of articles, each of us separately read all of the more than 3,500 abstracts obtained from a search of *Psychological Abstracts* using the keywords: decision making, judgment, probability judgment, and problem solving. The search was limited to the 10-year period extending from 1972 through 1981. To be selected, the ar-

[1] Sackett (1979) identifies variations of this bias, including the *all's well literature bias,* the *one-sided reference bias,* and the *hot-stuff bias.*

ticles had to be empirical, they had to have been published in an English language journal, the subjects had to have been adult humans, studies of group problem solving were omitted, and the abstracts had to state that the subjects' behavior had been compared with the prescriptions from an explicit model from logic, probability theory, or the like. Each article was classified as having obtained either good or poor performance based on the reported findings. There were 37 good-performance articles and 47 poor-performance articles.

Figure 1

Relationship Between Average Number of Citations Per Article in Print and Chronological Year, Separated by Type of Article



The *Social Science Citation Index* was used to obtain the frequency with which each article was cited each year. The "popularity" or "citability" of the 24 different journals in which the articles of our sample were published was ascertained using the Journal Citation Impact Measures (*Social Science Citation Index: Journal Reports*, 1977). This measure is correlated with the journal's circulation and is relatively stable over time (Buffardi & Nichols, 1981).

During the 10 years in question, poor-performance articles were cited significantly more often than good-performance articles. The mean for the former was 27.8 times, whereas the mean for the latter was 4.7, $t(82) = 2.39$, $p < .02$, a ratio of almost six to one. There was no correlation between the type of article (good vs. poor performance) and the "popularity" of the journals or between the type of article and the year of publication. Thus the observed significant differences in the popularity of the articles cannot be accounted for by these factors.

We next examined whether the citation frequency of poor- and good-performance articles changed over the years. Figure 1 shows the average citation frequency per article in print for each of the 10 years surveyed. Although both types of articles were cited with increasing frequency in recent years (the slopes of both regression lines are significantly different from zero, $p < .01$), the citation frequency increased at a significantly greater rate for poor-performance articles than for good-performance articles ($p < .001$), demonstrating that the preferential popularity of poor-performance articles has increased markedly in recent years.

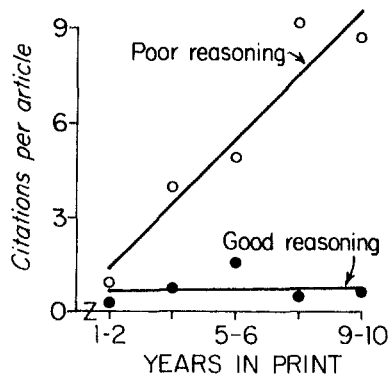
Figure 2 shows the average number of yearly citations per article in print as a function of how long the article has been in print by two-year intervals. The figure shows that the literature on good performance largely is ignored; the citation frequency remains small and is relatively unaffected by the time the articles have been in print. In contrast, the citation frequency of the poor-performance articles increased greatly with the passage of time.

Because newspaper and scientific journal coverage has been shown to be related to biases in readers' judgments (Christensen-Szalanski, Beck, Christensen-Szalanski, & Koepsell, 1983; Combs & Slovic, 1979; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978), we decided to examine whether the citation bias in favor of studies reporting poor reasoning performance was reflected in the opinions of those who read the literature. We expected the readers of this literature would think that humans are not very able judges or decision makers. Moreover, because it is a more familiar theme, the readers should be better able to recall documented examples (studies) of poor performance than they can of good performance.

Questionnaires were mailed to 149 United States members of the Judgment and Decision Making Society, a semiformal professional group, who were listed in the group's 1981 directory. Eighty people returned completed questionnaires. We asked the respondents to assess the overall quality of human judgment and decision-making abilities on a scale from 0 to 100 (0 = never optimal, 100 = always optimal), and to list four examples of documented poor judgment or decision-making performance and four examples of good performance.

Figure 2

Relationship Between Average Number of Citations Per Article and the Number of Years That the Article Has Been in Print, Separated by Type of Article



The respondents were divided into three groups representing three levels of research experience. This was based on the number of each person's self-reported publications in the field. Members of the low-experience group ($n = 39$) had fewer than 5 publications. Members of the medium-experience group ($n = 25$) had from 5 to 15 publications. Members of the high-experience group ($n = 16$) had more than 15 publications. The median ratings of the quality of human judgment and decision-making abilities were 45 for the low-experience group, 60 for the medium-experience group, and 65 for the high-experience group ($p < .05$, Jonckheere test of ordered alternatives).

As expected, an examination of the number of recalled examples of poor and good performance showed that respondents recalled significantly more examples of poor performance than of good performance (*Mdn*s 4 vs. 1, respectively, $p < .001$). However, the variety of poor-performance examples was extremely limited. Table 1 lists all responses that were given by at least 10% of the respondents. Eight of these responses accounted for 88% of the poor-performance examples given,

whereas all 10 of the responses accounted for only 31% of the good-performance examples given.

Note that all of the examples of poor performance were laboratory studies, usually with college students as subjects. In contrast, only 42% of the examples of good performance were laboratory studies; 58% were done in applied settings and/or used experts as subjects (e.g., livestock judges, weather forecasters, accountants).

In conclusion, these results demonstrate that there is indeed a citation bias in the judgment and decision-making literature. They corroborate the claims that studies which observe optimal behavior tend to be ignored in the literature (Berkeley & Humphreys, 1982; Cohen, 1981; Lopes, 1981).

Although our study design does not allow us to demonstrate that the citation bias influenced psychologists' opinions, the results show that a marked similarity exists between the citation bias and how the readers of that literature view human judgment and decision-making abilities. Less experienced researchers in the area appear to have a lower opinion of human reasoning ability than do highly experienced researchers. Perhaps the less experienced researchers are younger and obtained most of their experience only recently when concern about poor reasoning abilities has become so fashionable, or perhaps the more experienced researchers base their views less on what they read and more on their own experience with the vagaries of experimentation and data interpretation. Moreover, respondents recalled half again as many examples of poor performance as of good performance despite the fact that the variety of poor-performance examples was extremely limited.

Hammond (1982) recently called for psychologists to be more balanced in their approach to understanding human judgment and decision-making abilities. It would appear from the results of this study that Hammond's call might best be heeded. Although the study of reasoning errors can advance our understanding of reasoning processes (Kahneman & Tversky, 1982), so too can the study of good judgment (Christensen-Szalanski, 1978; Lopes, 1981; Lopes & Ekberg, 1980; Rachlin, Battalio, Kagel, & Green, 1981).

We do not believe that an editorial conspiracy has caused the citation bias. The nonsignificant correlation between type of study and citability of journal implies that good- and poor-performance articles are being published in the same or comparable journals, and there is not much difference in the proportions of each (44% and 56%, respectively). The poor-performance articles are just receiving most of the attention from other writers.

There probably are many reasons why this citation bias exists. For example, authors select citations to serve their personal goals (May, 1967) or to advocate their favored hypothesis (Armstrong, 1979), and readers overrate the results of prominent investigators (Owen, 1982). Sadly, scientific research is not immune to fads and fashions that result in a few hot topics getting the most attention (Armstrong, 1982; Boor, 1982; Dunnette, 1966; Thorne, 1977).

In addition to these problems, which are common to all academic disciplines, we think that there may be something else that is specific to research that involves reasoning. When people behave the way they are "supposed to," it often does not seem particularly remarkable. But when people behave in what appears to be an irrational manner, it is not only remarkable, but we can give it a name. Names like *representativeness, availability bias,* and *overconfidence bias* make the unfelicitous performance seem like a concrete phenomenon and create the illusion that we have actually explained something merely by naming it (Anderson, 1974). Psychology is full of names for various performance problems, and of late these seem to get labeled the something-or-other bias. Since the label *bias* is such an eye-catching term, perhaps to draw more attention to people's good performance, we should label that good performance the *"hats off" bias.*

Table 1
Frequency (%) of Response Among Participants (N = 80)

Response	Listed as example of poor reasoning	Listed as example of good reasoning
Availability	45	5
Representativeness	44	10
Overconfidence	33	0
Anchoring	28	2
(Mis)use of base rate	28	6
Conservatism	26	0
Weather forecasters	0	24
Hindsight	20	0
Livestock judges	0	15
Misuse of sample size	14	0
Total laboratory examples	100	42
Total applied/expert examples	0	58

REFERENCES

Anderson, N. H. (1974). Information integration theory: A brief survey. In D. H. Krantz et al. (Eds.), *Contemporary Developments in Mathematical Psychology* (Vol. 2, pp. 236–305). San Francisco: Freeman.

Armstrong, J. S. (1979). Advocacy and objectivity in science. *Management Science, 25,* 423–428.

Armstrong, J. S. (1982). Research on scientific journals: Implications for editors and authors. *Journal of Forecasting, 1,* 83–104.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance, 24,* 245–257.

Berkeley, D., & Humphreys, P. (1982). Structuring decision problems and the "bias heuristic." *Acta Psychologica, 50,* 201–252.

Boor, M. (1982). The citation impact factor: Another dubious index of journal quality. *American Psychologist, 37,* 975–977. (Comment)

Buffardi, L. C., & Nichols, J. A. (1981). Citation impact, acceptance rate, and APA journals. *American Psychologist, 36,* 1453–1456. (Comment)

Christensen-Szalanski, J. J. J. (1978). Problem solving strategies: A selection mechanism, some implications, and some data. *Organizational Behavior and Human Performance, 22,* 307–323.

Christensen-Szalanski, J. J. J. (in press). Discount functions and the measurement

of patients' values: womens' decisions during childbirth. *Medical Decision Making.*

Christensen-Szalanski, J. J. J., Beck, D., Christensen-Szalanski, C. M., & Koepsell, T. D. (1983). The effects of expertise and experience on risk judgments. *Journal of Applied Psychology, 68,* 278–284.

Christensen-Szalanski, J. J. J., Diehr, P. H., Bushyhead, J. B., & Wood, R. W. (1982). Two examples of good clinical judgment. *Medical Decision Making, 2,* 275–283.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences, 4,* 317–331.

Combs, B., & Slovic, P. (1979). Causes of death: Biased newspaper coverage and biased judgments. *Journalism Quarterly, 56,* 837–843.

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21,* 343–352.

Fischhoff, B. (1975). Hindsight is not equal to foresight. The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299.

Hammond, K. (1982, November). *To whom does the future belong? Is you is or is you ain't my baby?* Paper presented at the meeting of the Society for Judgment and Decision Making, Minneapolis, Minnesota.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11,* 123–141.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 551–578.

Loftus, E. F., & Beach, L. R. (1982). Human inference and judgment. Is the glass half empty or half full? *Stanford Law Review, 34,* 939–956.

Lopes, L. L. (1981). Performing competently. *The Behavioral and Brain Sciences, 4,* 343–344.

Lopes, L. L., & Ekberg, P. H. (1980). Test of an ordering hypothesis of risky decision making. *Acta Psychologica, 45,* 161–167.

May, K. O. (1967). Abuses of citation indexing. *Science, 156,* 890–891.

Muchinsky, P. M., & Fitch, M. K. (1975). Subjective expected utility and academic preferences. *Organizational Behavior and Human Performance, 14,* 217–226.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology, 32,* 932–943.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings in social judgment.* Englewood Cliffs, N.J.: Prentice-Hall.

Owen, R. Reader bias. (1982). *Journal of the American Medical Association, 247,* 2533–2534.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21,* 209–219.

Rachlin, H., Battalio, R., Kagel, J., & Green, L. (1981). Maximization theory in behavioral psychology. *The Behavioral and Brain Sciences, 4,* 371–388.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32,* 51–63.

Social Science Citation Index: Journal Reports. (1977). Philadelphia: Institute for Scientific Information.

Thorne, F. C. (1977). The citation index: Another case of spurious validity. *Journal of Clinical Psychology, 33,* 1157–1161.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science, 25,* 219–225.

This Comment is adapted from a paper presented at the Bayesian Research Conference, Los Angeles, February 10, 1983.

Correspondence concerning this Comment should be sent to Jay Christensen-Szalanski, Department of Family and Community Medicine, University of Arizona, Tucson, Arizona 85724.

The 20th International Congress of Applied Psychology

Sandra White
Advanced Research Resources Organization, Washington, D.C.

The 20th International Congress of Applied Psychology was held in Edinburgh, Scotland, on July 25–31, 1982. The Congress, organized every four years by the International Association of Applied Psychology (IAAP), was held in the George Square Complex of the University of Edinburgh. Approximately 1,500 colleagues from more than 60 countries attended. The Congress consisted of the scientific program, a coordinated social program, and organized professional visits to institutions of interest to psychologists.

The Rt. Hon. David Steel, rector of the University of Edinburgh, officially opened the congress during a session held in the McEwan Hall. At this session Edwin A. Fleishman (U.S.), president of IAAP, presented the presidential address. Addresses of welcome were given by the Rt. Hon. Tom Morgan, The Lord Provost of the City of Edinburgh; G. A. Randell (U.K.), chair of the Organizing Committee; W. T. Singleton (U.K.), chair of the Scientific Committee; and David Nelson (U.K.), chair of the Local Arrangements Committee. Also at this opening session, Gunnar Westerlund (Sweden), Roger Piret (Belgium), Claude Levy-Leboyer (France), and Fleishman (U.S.) were presented with scrolls for their contributions to the international development of psychology.

The program consisted of 20 invited keynote addresses, with simultaneous translation, by well-known colleagues from all over the world, 80 symposia and workshops, and 600 "interactive poster sessions," covering professional and cross-cultural issues, educational psychology, ergonomics, counseling and clinical psychology, industrial/organizational psychology, psychometrics, environmental psychology, and applied social areas.

Invited keynote speakers at the Congress (and their topics) included: Irwin Altman (U.S.), "Environmental Psychology: Promises and Prospects"; John Adair (Canada), "The Hawthorne Effect: A Reconsideration of the Methodological Artifact"; Chris Argyris (U.S.), "Problems in Producing Usable Knowledge for Implementing Liberating Alternatives"; Bernard M. Bass (U.S.), "Organizational Decision Processes: An Opportunity for Applied Psychology"; Donald Broadbent (U.K.), "Some Relations Between Clinical and Occupational Psychology"; Ching Chi-Cheng (People's Republic of China), "Applied Psychology in China"; Albert B. Cherns (U.K.), "Prerequisites for a Debureaucratized Society"; Hans Eysenck (U.K.), "The Conditioning Theory of Neurosis Revisited"; Fred Fiedler (U.S.), "Are Leaders an Intelligent Form of Life? A Long-Neglected Question of Leadership Theory"; Norman Frederiksen (U.S.), "Construct Validity and Construct Similarity"; Sol Garfield (U.S.), "The Effectiveness of Psychotherapy: The Perennial Controversy"; H. T. Himmelweit (U.K.), "Political Socialisation: A Socio-Psychological Study of Vote Choice"; Herbert Klausmeier (U.S.), "Improvement-Oriented Educational Research"; Paul Kline (U.K.), "Psychometrics: A Science With a Great Future Behind It?"; Jacques Leplat (France), "Error Analysis and Ac-