Published in final edited form as:
Law Hum Behav. 2009 June ; 33(3): 247–257. doi:10.1007/s10979-008-9133-0.

Can Jurors Recognize Missing Control Groups, Confounds, and Experimenter Bias in Psychological Science?

Bradley D. McAuliff
Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8255, USA

Margaret Bull Kovera
Department of Psychology, John Jay College of Criminal Justice, City University of New York, New York, NY 10019, USA mkovera@jjay.cuny.edu

Gabriel Nunez
Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8255, USA

Abstract

This study examined the ability of jury-eligible community members (N = 248) to detect internal validity threats in psychological science presented during a trial. Participants read a case summary in which an expert testified about a study that varied in internal validity (valid, missing control group, confound, and experimenter bias) and ecological validity (high, low). Ratings of expert evidence quality and expert credibility were higher for the valid versus missing control group versions only. Internal validity did not influence verdict or ratings of plaintiff credibility and no differences emerged as a function of ecological validity. Expert evidence quality, expert credibility, and plaintiff credibility were positively correlated with verdict. Implications for the scientific reasoning literature and for trials containing psychological science are discussed.

Keywords

Scientific reasoning; Internal validity; Expert testimony; Juror decision-making

Recent advances in DNA, blood type, and fingerprint testing have increased the likelihood that average citizens will confront complex scientific evidence when serving as jurors in civil and criminal cases. Nearly two-thirds (65%) of state court judges responding to a national survey indicated that they had some experience with DNA evidence in their courtrooms (Gatowski, Dobbin, Richardson, Ginsburg, Mertino, & Dahir, 2001). The role of psychological science in the legal system has burgeoned recently as well. Social or behavioral scientists constituted nearly one-quarter of all scientists in U.S. criminal appellate cases involving expert testimony from 1988 to 1998 (Groscup, Penrod, Studebaker, Huss, & O'Neil, 2002).

Research examining laypeople's scientific reasoning skills has enjoyed renewed interest among social scientists and legal scholars due to several recent U.S. Supreme Court rulings on the admissibility of expert evidence (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*,

Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8255, USA bradley.mcauliff@csun.edu.

1993; *General Electric Co. v. Joiner*, 1997; *Kumho Tire Co. v. Carmichael*, 1999). *Daubert* and its progeny have entrusted judges with a gatekeeping role in which they should base their admissibility decisions on the relevance and reliability of the expert evidence. Despite the Court's confidence in judges' ability to fulfill their gatekeeping role, many judges lack the scientific literacy required for a *Daubert* analysis (Gatowski et al., 2001) and have difficulty identifying methodologically-flawed expert testimony (Kovera & McAuliff, 2000a). Attorneys also struggle to effectively evaluate expert evidence (Kovera & McAuliff, 2000b). Their ability to make and successfully argue motions to exclude junk science, cross-examine an expert, or consult their own expert may be limited as a result (Kovera, Russano, & McAuliff, 2002).

Based on these limitations, it is likely that at least some invalid research will reach laypeople serving as jurors in court. Can they recognize variations in the validity of psychological science? We examined this research question by presenting jury-eligible community members expert testimony containing an internally valid versus invalid experiment in a simulated hostile work environment case.

SCIENTIFIC REASONING ABILITY

Previous research from the scientific reasoning literature has shown that laypeople have difficulty properly evaluating statistical and methodological information. Laypeople neglect base-rate information when judging the probability of certain outcomes (Kahneman & Tversky, 1973) and fail to recognize limitations associated with small sample size (Fong, Krantz, & Nisbett, 1986), sample bias (Hamill, Wilson, & Nisbett, 1980), and missing control groups (Mill, Gray, & Mandel, 1994). Similar deficits have been observed in legal contexts as well. Mock jurors underutilize expert probabilistic testimony compared to Bayesian norms (Faigman & Baglioni, 1988; Kaye & Koehler, 1991; Schklar & Diamond, 1999; Thompson & Schumann, 1987) and are reluctant to base verdicts on statistical evidence alone (Niedermeier, Kerr, & Messé, 1999; Wells, 1992). Mock jurors also have difficulty comprehending expert testimony on statistical matters. Only 14% of participants in one experiment correctly answered two questions designed to assess their understanding of a statistical expert's testimony and 43% provided incorrect answers to both questions (Faigman & Baglioni, 1988).

Other research reveals a more optimistic picture of laypeople's scientific reasoning skills in legal settings. In one study, a forensic serologist provided crucial blood- and enzyme-type evidence within the context of a videotaped simulation of a rape trial (Smith, Penrod, Otto, & Park, 1996). Mock jurors who learned that 20% of the population had an enzyme type that was shared by the assailant and the defendant were more likely to have judged the defendant as more guilty than were those who learned that 80% of the population had the same enzyme type. Mock jurors' guilty verdicts in another trial simulation decreased as the blood-type frequency shared between the defendant and the population increased (Goodman, 1992).

Two different trial simulations in which jurors were sensitive to the absence of control group information are particularly relevant to the present study. Citizens reporting for jury duty in an experiment by McAuliff and Kovera (2008) read a hostile work environment case in which the plaintiff's expert testified about a study she had conducted on men's reactions to sexualized television commercials. Only jurors who were high in the need for cognition (NC) (i.e., individuals who naturally engage in and enjoy effortful cognitive endeavors; Cacioppo & Petty, 1982) were more likely to find for the plaintiff and to rate the quality of the expert's study more favorably when it included a control group than when it did not. Similarly, jury-eligible citizens participating in a second study by Levett and Kovera (2007) read a child sexual abuse case containing defense expert testimony on witness suggestibility.

Only jurors who received opposing expert testimony highlighting the absence of an appropriate control group rated the defense expert's credibility less favorably, although this effect was only marginally significant.

Overall these mixed findings regarding laypeople's ability to reason about scientific issues in everyday and legal settings suggest that jurors may have difficulty differentiating valid research from junk science in trials containing expert testimony. Even if jurors are able to detect some internal validity threats such as missing control groups, they may fail to recognize others such as the presence of a confound or experimenter bias. If indeed this is the case, we are left to wonder what other characteristics of expert evidence influence jurors' evaluations and judgments? Information-processing models from the social-cognitive psychological literature on persuasion provide a much-needed theoretical framework to predict how jurors make decisions when confronting psychological science in court.

DUAL PROCESS MODELS OF PERSUASION

Two information-processing models are useful for understanding how jurors evaluate information presented at trial: the elaboration likelihood model (ELM; Petty & Cacioppo, 1986) and the heuristic-systematic model (HSM; Chaiken, 1980; Chaiken, Liberman, & Eagly, 1989). Both models propose that people generally want to hold correct attitudes and are willing to engage in varying levels of cognitive effort to achieve this goal. Central (ELM) or systematic (HSM) processing is highly effortful cognitive activity aimed at the careful analysis of a persuasive message's content. If the persuasive message contains valid, high-quality arguments, systematic processors are more likely to be persuaded than when the persuasive message contains invalid or weak arguments (Petty & Cacioppo, 1984; Petty, Cacioppo, & Goldman, 1981). Peripheral (ELM) or heuristic (HSM) processing is less cognitively taxing than systematic processing and entails the use of mental shortcuts or decision-rules to evaluate a persuasive message. Source-related cues including expertise, likeability, and physical attractiveness (Chaiken & Maheswaran, 1994) and message-related cues such as length or number of arguments (Petty & Cacioppo, 1984) can influence message evaluation for heuristic processors.

Systematic Versus Heuristic Processing of Psychological Science

According to the ELM and HSM, two important factors that influence whether people engage in systematic or heuristic processing are motivation and ability (Chaiken, 1980; Petty & Cacioppo, 1986). An individual must be both *motivated* and *able* to process a persuasive message systematically before such processing can occur. What factors affect one's motivation or ability to process systematically? Research has shown that personal relevance, personal responsibility, and NC all influence one's motivation to process a persuasive message systematically (Chaiken et al., 1989; Cacioppo, Petty, Feinstein, & Jarvis, 1996; Petty et al., 1981). Similarly, factors that determine one's ability to engage in systematic processing include information complexity, prior knowledge, distraction, and repetition (Petty & Cacioppo, 1986; Chaiken & Maheswaran, 1994; Ratneshwar & Chaiken, 1991).

If either motivation or ability is low, an individual is more likely to engage in heuristic processing when evaluating the persuasive message compared to someone who is motivated *and* able to process systematically. One message-related cue that jurors might rely on when processing heuristically is the representativeness of the research. When making complex decisions, people often rely on a representativeness heuristic to simplify the decision-making task (Tversky & Kahneman, 1974). Just as participants in that study saw certain characteristics (shy, withdrawn, structured, and helpful) as being more representative of certain occupations (librarian) than others, jurors may rely on a representativeness heuristic involving the study's ecological validity or the degree to which its methods, materials, and

setting approximate the real-life situation under examination (Brewer, 2000). One specific aspect of ecological validity that might affect jurors' evaluations of an experiment is the nature of its participant sample. Jurors may judge a study more favorably when it contains participants who are similar to members of the population to which the expert wishes to generalize his or her research findings than when it does not. This heuristic inference may occur irrespective of the study's internal validity.

At least two studies have investigated the effects of ecological validity on juror decision-making (Kovera, McAuliff, & Hebert, 1999; McAuliff & Kovera, 2008). Mock jurors used ecological validity as a heuristic cue to evaluate the trustworthiness and credibility of certain witnesses in the Kovera et al. study, but their verdicts and evaluations of expert evidence quality did not differ as a function of ecological validity in either study. Thus, it appears ecological validity may affect some trial-related judgments but not others.

OVERVIEW AND HYPOTHESES

The present study examined two primary questions related to jurors' ability to reason about psychological science: Can jurors differentiate an internally valid study from one containing a missing control group, confound, or experimenter bias? If they cannot, do jurors instead rely on heuristic cues associated with the ecological validity of the expert's research when judging its quality? We sought to answer these research questions by presenting jury-eligible community members a written trial summary of a hostile work environment case containing expert testimony. The expert described a study she had conducted on the effects of sexual primes on men's behavior. We varied the study's internal validity (valid, missing control group, confound, and experimenter bias) and ecological validity (high, low) to determine whether those variables moderated mock jurors' evaluations of expert evidence quality and other trial-related judgments.

We generated two hypotheses that flow directly from our research questions and the literature reviewed. Based on previous work by McAuliff and Kovera (2008) and Levett and Kovera (2007), we predicted that mock jurors would be sensitive to missing control group information but not the more sophisticated internal validity threats of a confound or experimenter bias. Support for this hypothesis would consist of a statistically significant main effect for the study's internal validity in which mock jurors' ratings of expert evidence quality are higher for the valid versus missing control group version of the expert's study but no different from the confound and experimenter bias versions. Second, given their limited ability to process systematically, we predicted that mock jurors in the confound and experimenter bias conditions would rely on the study's ecological validity as a heuristic cue when rating its quality. Support for this hypothesis would consist of a statistically significant interaction between the study's internal validity and ecological validity. Specifically, the mock jurors' ratings of expert evidence quality should be higher for the study high versus low in ecological validity when the study contains a confound or experimenter bias; however, no such differences should emerge for the valid or missing control group versions.

METHOD

Participants

Two hundred forty-eight community members residing in southern California participated in our study in exchange for $5.00. We recruited community members by distributing a flyer that described the research participation opportunity in our local community and by offering students extra-credit for referring extended family members to participate in the research. All participants met the California requirements for jury eligibility: a U.S. citizen who is at

least 18 years of age, able to understand English, and who has not been convicted of a felony (*California Code of Civil Procedure*, §203).

Participants averaged 38 years in age and most were female (52%), had not served on a jury before (79%), and had never been involved in legal proceedings (76%). Members of various ethnic groups participated including: Caucasians (45%), Hispanics, Central/South Americans, Mexicans (31%), African Americans (5%), Middle Easterners (9%), Asians (6%), Native American (1%), and Others (3%). Community members' average ratings of statistical knowledge ($M = 4.01$), motivation ($M = 4.55$), and cognitive effort expended while participating in the experiment ($M = 4.95$) were above the 7-point Likert-type scale midpoints with larger numbers indicating more positive responses.

Trial Stimulus

Participants read a 15-page summary of a simulated civil case in which the plaintiff alleged she was the victim of gender discrimination due to a hostile work environment. The fact pattern of the trial simulation was derived from an actual case (*Robinson v. Jacksonville Shipyards Inc.*, 1991). Certain facts were modified to prevent the possibility that participants might recognize the case from media coverage. The trial simulation consisted of opening statements and closing arguments from both attorneys, direct-and cross-examined testimony from five witnesses, and standard California judicial instructions.

The plaintiff was the sole female mechanic who worked with a male maintenance crew in a trucking company service garage. She alleged that sexual materials were displayed throughout workplace and that she was the target of unwelcome sexual advances. A female coworker corroborated the plaintiff's allegations. On cross-examination, the plaintiff admitted that she sometimes used crude language and told sexual jokes to her male coworkers. A social psychologist who testified on the plaintiff's behalf discussed conditions that increase the likelihood of sexual harassment (rarity of women in the workplace, a paucity of information available when evaluating workers for promotion, ambiguity of evaluation criteria, and a sexualized working environment). Two defense witnesses testified. A shift supervisor claimed that the plaintiff never complained about the sexual materials until she was reprimanded for her tardiness and absenteeism from work and that once the plaintiff complained about the posters, he removed all of them. He added that the plaintiff often used profanity and joked about sexual matters with her male coworkers. A mid-level administrator for the company testified that the plaintiff had shown him examples of materials she claimed were offensive but that those materials were similar to many ads appearing on television. The administrator stated that he offered to follow up on the issue but that the plaintiff did not provide the names of the alleged harassers.

Experimental Manipulations

The expert described a study she had conducted on the effects of sexually suggestive materials on men's behavior toward women. This study was based on an experiment by Rudman and Borgida (1995), who found that men who viewed sexualized commercials sat closer to a female confederate, evaluated her more negatively, and asked her a greater number of sexually inappropriate questions compared to men who viewed nonsexualized commercials. Moreover, the female confederate in that experiment, who was blind to experimental condition, rated the men who had viewed the sexualized commercials to be more sexually motivated than those who had viewed the nonsexualized commercials. Within the expert's description of her study, we manipulated its internal validity and ecological validity.

Design

This study used a 4 Internal Validity (Valid, Missing Control Group, Confound, and Experimenter Bias) × 2 Ecological Validity (High, Low) fully-crossed factorial design. We randomly assigned participants to one of eight experimental conditions in which they read a version of the expert's study that varied in internal validity and ecological validity. With the exception of these manipulations, all information presented in the different versions of the trial summary was identical.

Internal Validity—The first version of the study was identical to the original Rudman and Borgida (1995) study and contained no threats to internal validity (valid condition). Our assessment that this was a valid study is supported by the fact that it was published in a peer-reviewed journal and won SPSSI's Gordon Allport prize for best intergroup relations paper of the year. Thus, social scientists in the field judged the study to be methodologically sound and worthy of high commendation.

Unlike participants in the valid condition who viewed either sexualized or nonsexualized commercials, participants in a second version of the study viewed the sexualized commercials only (missing control group condition). Because this version of the expert's study did not include an appropriate control group, any conclusions regarding the effects of viewing sexualized commercials were invalid.

In the third version of the study, there were two female confederates (confound condition). One research assistant interacted exclusively with men who had viewed the sexualized commercials and the other interacted exclusively with men who had viewed the nonsexualized commercials. Because men's exposure to sexualized commercials was confounded with the identity of the female research assistants, differences in men's behavior could be attributable to the experimental treatment or to the unique characteristics of each female research assistant.

In the fourth version, the female research confederate knew in advance whether the men had viewed sexualized or nonsexualized commercials (experimenter bias condition). Such knowledge may have caused the female research assistant to treat the men in each condition differently and may have resulted in the observed differences between groups.

Ecological Validity—In the high ecological validity condition, participants were employees at a trucking company similar to the plaintiff's workplace. In the low ecological validity condition, participants in the expert's study were college undergraduates and therefore were less similar to the population to which the expert wished to generalize her findings.

Dependent Measures

Participants decided whether the plaintiff had demonstrated by a preponderance of evidence that the trucking company constituted a hostile working environment using a dichotomous scale (defendant is liable/defendant is not liable).

Participants rated the quality of the expert's study using a series of 7-point Likert-type scales where 1 = Strongly Disagree and 7 = Strongly Agree (see Table 1). We summed jurors' responses to those nine items to create a single index of study quality, with higher numbers representing more positive evaluations (hereinafter referred to as “expert evidence quality,” Cronbach's alpha = .87).

Participants evaluated the expert and plaintiff on eight 7-point bipolar adjective pairs: not believable/believable, certain/uncertain (R), convincing/not convincing (R), not credible/

credible, intelligent/unintelligent (R), good/bad (R), immoral/moral, and respectable/not respectable (R). Participants' ratings were averaged across these items to form a single index of credibility for each witness (hereinafter referred to as “expert credibility” and “plaintiff credibility, Cronbach's alpha = .86 and .85, respectively). Items followed by (R) were recoded so that higher numbers on the final scales represented more positive witness evaluations.

Four additional items served as manipulation checks for the internal validity and ecological validity variables. With respect to internal validity, jurors responded to three forced-choice questions asking what type of commercials the expert included in her study (sexualized only, both sexualized and nonsexualized), how many women posed as applicants for the research assistant position in the expert's study (one, two), and whether the research assistant in the expert's study knew that the male participants had watched the sexualized or nonsexualized ads (yes, no). Jurors also used a 7-point Likert-type scale to indicate how similar the participants in the expert's study were to the workers at the plaintiff's trucking company (1 = Not at All Similar, 7 = Very Similar).

Mock jurors concluded their participation in our research by providing demographic information about their gender, age, jury eligibility, racial/ethnic identity, history of jury service, and previous involvement in legal proceedings (civil or criminal, plaintiff or defendant). We used a series of Likert-type scales to measure participants' self-reported knowledge of statistics and research methodology, their motivation level while reading the trial summary and answering the questions, and how much cognitive effort they expended while reading the trial summary and answering the questions (1 = None or Not at All, 7 = A lot or Extremely).

Procedure

We collected data from participants in groups of 5–20 participants per session. At the beginning of each session, the experimenter read standardized instructions and distributed an informed consent sheet for participants to read and sign. After providing informed consent, participants received the trial stimulus and dependent measures. Participants did not deliberate or confer with one another at any point during the study. Once participants completed the dependent measures, they were debriefed and paid $5.00.

RESULTS

Manipulation Checks

Participants noticed the different levels of the internal validity manipulation that were included in our research.

Missing Control Group—Those who read the valid, confound, or experimenter bias versions were more likely to report that participants had viewed both sexualized and nonsexualized commercials (90%) than that participants had viewed the sexualized commercials only (10%), χ^2 (1, N = 189) = 120.64, $p < .001$, $\Phi = .80$. Those who read the missing control group version of the expert's study were more likely to report that participants had viewed sexualized commercials only (83%) than that participants had viewed both types of commercials (17%), χ^2 (1, N = 58) = 24.90, $p < .001$, $\Phi = .66$.

Confound—Participants who read the valid, missing control group, or experimenter bias version of the study were more likely to report that it had included only one research assistant (86%) than two (14%), χ^2 (1, N = 183) = 93.78, $p < .001$, $\Phi = .72$. Similarly, those who read the confounded version of the expert's study were more likely to report that it had

included two research assistants (74%) as opposed to one (26%), χ^2 (1, N = 62) = 14.52, p < .001, Φ = .48.

Experimenter Bias—Participants who read the valid, missing control group, or confound version of the expert's study were more likely to report that the female research assistant was blind to experimental condition (89%) rather than that she was not blind (11%), χ^2 (1, N = 186) = 114.60, p < .001, Φ = .78. Those who read the experimenter bias version of the study were more likely to report that the female research assistant knew what type of commercials the men had viewed (75%) than that she did not know (25%), χ^2 (1, N = 61) = 15.75, p < .001, Φ = .51.

Ecological Validity—Our experimental manipulation of the study's ecological validity was successful as well. Participants who read the version of the study that included a sample of trucking employees judged those participants to be more similar to the workers at the plaintiff's workplace than did jurors who read the study that used an undergraduate psychology student sample, F(1, 242) = 4.99, p = .03, partial η^2 = .02, Ms = 4.23 and 3.80, respectively.

Data Analytic Strategy and Primary Results

We began by subjecting jurors' dichotomous liability verdicts to a logistic regression. We regressed verdict on internal validity, ecological validity, and the interaction of these two variables. Neither the main effects nor interaction were statistically significant, χ^2 (3, N = 248) = 2.08, p = .56, Φ = .09.

Next we subjected the data collected from the expert evidence quality, expert credibility, and plaintiff credibility dependent measures to a multivariate analysis of variance (MANOVA) to explore potential differences in mock jurors' responses as a function of the manipulated variables. We used the Pillai's Trace criterion multivariate statistic to test the significance of all main effects and interactions. A 4 Internal Validity 9 × 2 Ecological Validity MANOVA revealed a statistically significant main effect for the study's internal validity, Mult. F(9, 717) = 1.99, p < .05, partial η^2 = .02. No other main effects or interactions were statistically significant.

We followed-up the significant multivariate main effect for the Internal Validity variable using univariate F-tests for each of the four dependent measures. Only the tests for the expert evidence quality and expert credibility dependent measures reached traditional levels of statistical significance (see Table 2 for results). Post-hoc comparisons revealed that mock jurors' ratings of evidence quality were higher for the valid study than for the missing control group study. Mock jurors' ratings of evidence quality in the confound and experimenter bias conditions did not differ from each other or any of the other means (see Table 2). Similarly, mock jurors' ratings of expert credibility were higher for the valid study compared to the missing control group study and their ratings of evidence quality in the confound and experimenter bias conditions did not differ from each other or any of the other means (see Table 2).

We reran the MANOVA and univariate F-tests using only participants who answered the internal validity manipulation checks correctly. The results and pattern of effects were consistent with those obtained when the entire participant sample was included.

Correlations Among the Dependent Measures

Mock jurors' verdicts were positively related to their ratings of expert evidence quality and plaintiff credibility (see Table 3). The more favorably participants viewed the expert's study, the more likely they were to find the defendant liable for a hostile work environment.

DISCUSSION

We designed the present study to examine two research questions involving jurors' ability to reason about psychological science effectively. Can jurors differentiate an internally valid versus invalid study? And if they cannot, do they instead rely on other heuristic cues associated the expert's research? We discuss the specific findings relevant to each of these hypotheses and conclude by considering the limitations of our work and its implications for the scientific reasoning literature and trials containing psychological science presented by experts.

Can Jurors Recognize Missing Control Groups, Confounds, and Experimenter Bias in Psychological Science?

Jurors in our study recognized one threat to internal validity (missing control group) but not others (confound, experimenter bias). Other trial simulation research has observed similar juror sensitivity to missing control group information, but only under special circumstances. High NC jurors were able to distinguish a valid study from one missing a control group (McAuliff & Kovera, 2008). There is limited evidence that jurors who heard opposing expert testimony that highlighted the importance of control groups might be able to differentiate between an expert's research that contained or did not contain a control group (Levett & Kovera, 2007).

Why did jurors detect the missing control group threat to internal validity in our study even when these “special circumstances” were not present? First with respect to McAuliff and Kovera (2008), we simply do not know whether our jurors shared similar or different levels of NC with those who participated in the earlier study because our jurors did not complete the NC scale. We do know that jurors in our study reported expending cognitive effort when reading the trial stimulus and answering the questions (M = 4.95 on a scale where 1 = None and 7 = A lot). It is possible that this high level of cognitive effort is similar to the high NC group in McAuliff and Kovera's study. Second with respect to the Levett and Kovera (2007) study, we used a different trial stimulus that contained different expert testimony than those researchers. Recall that Levett and Kovera presented jurors with a criminal child sexual abuse case containing expert testimony about the influence of suggestive questioning on children's reports. They varied the presence/absence of control group information by including whether a group was questioned using nonsuggestive techniques. It is possible that missing control group information of this nature presented in this context is more difficult to recognize than it was in our study.

Moving beyond the missing control group issue, our study is the first to examine jurors' sensitivity to the presence of a confound and experimenter bias in psychological science presented in court. Jurors were insensitive to these variations in internal validity across all four dependent measures. This result is consistent with the previous work by Kovera et al. (1999) who observed that variations in the construct validity of an expert's study did not affect mock jurors' verdicts or ratings of expert/plaintiff credibility and trustworthiness.

In considering potential explanations for the null effects observed in our study, it is critical to keep in mind our manipulation checks demonstrated that jurors were sensitive to the inclusion of a single versus multiple female research assistants (confound) and whether she knew in advance what condition the male participants had been exposed to (experimenter

bias). The statistically significant effect for the missing control group version of the expert's study helps us rule out the possibilities of insufficient power and insensitive dependent measures to detect differences involving the confound or experimenter bias versions should they have existed. Additional calculations confirm that our study had adequate power (.85) to detect a small- to medium-sized effect given the number of participants in our study and $\alpha = .05$. It also seems unlikely that our results can be explained by decreased levels of juror motivation in the confound and experimenter bias versions of the expert's study because a univariate ANOVA revealed that jurors' self-reported motivation did not differ across the four internal validity conditions [$F(1, 238) = .07$, $p = .80$, partial $\eta^2 = .00$].

What seems more plausible to us is that internal validity threats in the form of confounds and experimenter bias are inherently more difficult to detect than a missing control group because they require a more sophisticated knowledge of research methodology. One need not search far in today's popular media to find examples of television programs and advertisements touting the benefits of Drug *X* to improve hair loss, Toothpaste *Y* to promote decay, and Detergent *Z* to get clothes cleaner compared to some placebo, control, or other comparison group. Yet media coverage of confounds and experimenter bias is virtually nonexistent in popular culture. As a result, these internal validity threats may require more specialized training to comprehend and later identify.

Did Jurors Who Failed to Recognize Certain Internal Validity Threats Instead Rely on the Representativeness Heuristic to Render Judgments?

If individuals are unable or unmotivated to process a message systematically, the ELM and HSM propose that they will rely on heuristics involving source- or message-related cues to make decisions. We predicted that mock jurors in our study who failed to systematically process the psychological science presented by the expert would judge the study's quality to be higher when it contained participants who were similar to the population to which the expert intended to generalize her results (high ecological validity) than when participants were less similar (low ecological validity). This hypothesized effect would be consistent with the representativeness heuristic such that mock jurors would view the expert's study more favorably when it included trucking company employees versus college students because the former are more representative of the plaintiff's coworkers who allegedly harassed her. The data did not support our hypothesis as the expert evidence quality ratings (or any other dependent measure) did not differ as a function of the study's ecological validity.

Why did mock jurors not use the heuristic cue of ecological validity? We must begin by ruling out the possibility that these null effects are due to statistical artifacts. First, recall our analyses indicated that jurors attended to the manipulated variables when reading the trial stimulus. Jurors in the high ecological validity condition reported that participants in the expert's study were more similar to the employees at the plaintiff's workplace than participants in the low ecological validity condition, although this effect was small in size (partial $\eta^2 = .02$). Second, our study had power equal to .85 to detect a small-to medium-sized effect given the number of participants in our study and $\alpha = .05$. Third, as further evidence supporting the power of our tests, differences relatively small in size (partial η^2s ranging from .02 to .05; Cohen, 1988) reached traditional levels of statistical significance in several of the ANOVAs used to examine our data. Finally, there was no evidence that the null effects were the result of a restricted response range. Jurors' responses varied greatly within and among the various dependent measures and there was no evidence of a floor or ceiling effect as most means fell more toward the scale mid-point than either extreme. Based on these indices, we believe the null effects associated with ecological validity reflect a true lack of differences and not statistical artifacts.

We believe more suitable explanations exist for why jurors did not rely on the ecological validity of the expert's study. Our heuristic cue manipulation required jurors to (1) attend to specific methodological features of the expert's study, namely the participant sample and (2) use that information as a proxy for study quality. This is a more sophisticated, research-oriented task than what is typically required in persuasion studies where participants rely on heuristics such as “Experts can be trusted” and “Argument length implies strength” to make decisions. Although jurors in our study did attend to the participant sample information (as evidenced by our successful manipulation check), recognizing how these features relate to study quality may have exceeded their lay knowledge or ability. It is also possible that jurors were knowledgeable and able in this regard but simply made mistakes, chose to disregard the information altogether (after all, low ecological validity does not necessarily equate to low internal validity), or relied more heavily on other features of the expert's study to evaluate its quality.

Placing our ecological validity results in the broader context of previous research, we should note that mock jurors' insensitivity to high versus low ecological validity in our study is consistent with the previous research by McAuliff and Kovera (2008) who found that actual jurors' decisions did not vary as a function of ecological validity; however, our findings are also inconsistent with earlier research by Kovera et al. (1999) who observed that mock jurors used ecological validity as a heuristic cue to evaluate expert credibility and plaintiff credibility/trustworthiness. All three experiments used similar fact patterns and case materials; however, the Kovera et al. study was unique in that it included jury-eligible undergraduate psychology students and a videotaped trial stimulus (as opposed to jury-eligible community members and a written trial stimulus). Perhaps these or other differences enhanced jurors' sensitivity to ecological validity in Kovera et al.'s study compared to the other two. Additional research is needed to reconcile these disparate results and to better understand if and when jurors rely on ecological validity as a heuristic cue.

Correlations Among the Dependent Measures

Correlations among the dependent measures revealed that mock jurors' verdicts were positively related to their ratings of expert evidence quality and plaintiff credibility. On the one hand, these results are encouraging when we consider that jurors were able to identify at least one threat to internal validity. We can expect jurors who are presented a missing control group to make better decisions when relying on their evaluations of expert evidence quality. On the other hand, the fact that jurors' evaluations of expert evidence quality were positively related to their verdicts is discouraging because jurors were insensitive to the presence of a confound or experimenter bias in the expert's research. In other words, jurors' evaluations of expert evidence quality were imperfect but they still relied on those evaluations when rendering verdicts.

Limitations and Research Implications

Certain limitations must be considered before generalizing our findings to cases in which jurors are asked to reason about scientific evidence. The written trial summary, although detailed and realistic, constituted a relatively impoverished stimulus compared to the courtroom experience of jurors in a real case. It lacked certain source- (e.g., expert credentials) and audience-related cues (e.g., reactions of other jurors) that might interact with internal validity manipulations to affect jurors' decisions. Also, jurors rendered liability verdicts independently of one another. Previous research suggests, however, that jurors' pre- and post-deliberation verdicts often do not differ (Hastie, Penrod, & Pennington, 1983) and that jurors rarely discuss the expert or the expert's testimony during deliberations (Kovera, Gresham, Borgida, Gray, & Regan, 1997). It is also possible that our recruitment methods (i.e., student “word of mouth” and flyers placed in the local community) may have

inadvertently resulted in a sample with different demographic characteristics than if we had included citizens actually reporting for jury duty as did McAuliff and Kovera (2008). Despite this possibility, we are confident that our study's sample is more representative of actual jurors and their ability to identify flawed research than the college student samples typically used in jury decision-making research.

One final limitation should be kept in mind when considering the implications of our findings. Our internal validity manipulation resulted in differences that were statistically significant but relatively small in size with partial η^2s ≤ .07. These small effects raise the issue of practical versus statistical significance, especially when we consider that jurors' verdicts were unaffected by the internal validity manipulation. In other words, despite the statistically significant effects of internal validity on jurors' evaluations of expert evidence quality and expert credibility, one could argue these differences were not practically significant because jurors' verdicts were unaffected. Although this may be true, future research is necessary to examine whether the internal validity effects observed in our civil trial simulation emerge and are more practically significant in a criminal case in which the burden of proof is much higher and the verdict distribution shifts as a result. Invalid versus valid expert evidence may be enough to tip the scales of justice in favor of a defendant in a case when jurors' must be convinced of guilt “beyond a reasonable doubt.”

Despite these limitations, the present study contributes to the scientific reasoning literature and to the growing body of empirical research on juror decision-making in complex cases involving expert testimony on statistical, probabilistic, and experimental findings. Consistent with earlier basic and applied research, our results indicate that jurors may be unable to evaluate statistical and methodological issues in a sophisticated manner. Our study also has practical implications for trials containing psychological science. Recall that the *Daubert* decision placed judges in a gatekeeping role in which they are the arbiters of whether psychological science is admitted at trial. If judges are unable to differentiate between valid and junk science (and research by Kovera & McAuliff, 2000a and Gatowski et al., 2001 suggests that this may be the case), it is likely that invalid research will be admitted at trial. Given the inability of jurors to detect flawed psychological science in our study, the role of attorneys in identifying invalid science becomes increasing important. In other words, if judges admit flawed testimony and jurors tend to believe it, attorneys must first recognize the junk science themselves and then use the traditional evidentiary safeguards enumerated by the *Daubert* Court. Future research addressing the effectiveness of these safeguards would be extremely worthwhile from both a social scientific and legal perspective. Does a cross-examination that focuses on internal validity issues help jurors' identify common threats to internal validity such as those included in our study? Can judicial instructions on the burden of proof that emphasize jurors' accountability sufficiently motivate jurors to evaluate expert evidence effectively?

Because the success of these evidentiary safeguards hinges on attorneys' ability to recognize flawed research and preliminary research by Kovera and McAuliff (2000b) suggests that they are no better doing so than judges or jurors, training programs on statistics and research methodology for the judiciary and bar become increasingly important. Future research aimed at developing new programs or evaluating those already in place is greatly needed if we genuinely desire to help the legal system better accommodate jurors' reasoning skills in trials containing psychological science.

Acknowledgments

This research was supported in part by grants from the College of Behavioral and Social Sciences and the Office of Research and Sponsored Projects at California State University, Northridge to the first author and a grant from the National Science Foundation (SBE #9711225) to the second author. Portions of this research were presented at the

March 2005 meeting of the American Psychology-Law Society in La Jolla, CA. We would like to thank Karla Gonzalez, Alicia Samis, Jose Vargas, and Yoshino Zenta for their assistance with data collection and entry.

REFERENCES

Brewer, MB. Research design and issues of validity. In: Reis, HT.; Judd, CM., editors. Handbook of research methods in social and personality psychology. Cambridge University Press; New York: 2000. p. 3-16.

Cacioppo JT, Petty RE. The need for cognition. Journal of Personality and Social Psychology 1982;42:116–131.

Cacioppo JT, Petty RE, Feinstein JA, Jarvis WBG. Dispositional differences in cognitive motivation: The life and times of individuals varying in the need for cognition. Psychological Bulletin 1996;119:197–253.

Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology 1980;39:752–766.

Chaiken, S.; Liberman, A.; Eagly, A. Heuristic and systematic information processing within and beyond the persuasion context. In: Uleman, JS.; Bargh, JA., editors. Unintended thought. Guilford Press; New York: 1989. p. 212-251.

Chaiken S, Maheswaran D. Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. Journal of Personality and Social Psychology 1994;66:460–473. [PubMed: 8169760]

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. Lawrence Erlbaum; Hillsdale, NJ: 1988.

Daubert. Merrell Dow Pharmaceuticals Inc.. 113 S.Ct. 2786. 1993.

Faigman DL, Baglioni AJ. Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence. Law and Human Behavior 1988;12:1–17.

Fong GT, Krantz DH, Nisbett RE. The effects of statistical training on thinking about everyday problems. Cognitive Psychology 1986;18:253–292.

Gatowski SI, Dobbin SA, Richardson JT, Ginsburg GP, Merlino ML, Dahir V. Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-*Daubert* world. Law and Human Behavior 2001;25:433–458. [PubMed: 11688367]

General Electric Co.. et al. Joiner et ux.. 522 U.S. 136. 1997.

Goodman J. Jurors' comprehension and assessment of probabilistic evidence. American Journal of Trial Advocacy 1992;16:361–389.

Groscup JL, Penrod SD, Studebaker CA, Huss MT, O'Neil KM. The effects of *Daubert* on the admissibility of expert testimony in state and federal criminal cases. Psychology, Public Policy, and Law 2002;8:339–372.

Hamill R, Wilson TD, Nisbett RE. Insensitivity to sample bias: Generalizing from atypical cases. Journal of Personality and Social Psychology 1980;39:578–589.

Hastie, R.; Penrod, SD.; Pennington, N. Inside the jury. Harvard University Press; Cambridge, MA: 1983.

Kahneman D, Tversky A. On the psychology of prediction. Psychological Review 1973;80:237–251.

Kaye DH, Koehler JJ. Can jurors understand probabilistic evidence? Journal of the Royal Statistical Society Series A 1991;154(Pt 1):75–81.

Kovera MB, Gresham AW, Borgida E, Gray E, Regan PC. Does expert testimony inform or influence juror decision-making? A social cognitive analysis. Journal of Applied Psychology 1997;82:178–191. [PubMed: 9119796]

Kovera MB, McAuliff BD. The effects of peer review and evidence quality on judge evaluations of psychological science: Are judges effective gatekeepers? Journal of Applied Psychology 2000a; 85:574–586. [PubMed: 10948802]

Kovera, MB.; McAuliff, BD. Attorneys' evaluations of psychological science: Does evidence quality matter?. In: Kovera, M.; McAuliff, B., editors. Judge, attorney, and juror decisions about scientific and statistical evidence; Symposium conducted at the annual meeting of the American Psychology-Law Society; New Orleans, LA. 2000b.

Kovera MB, McAuliff BD, Hebert KS. Reasoning about scientific evidence: Effects of juror gender and evidence quality on juror decisions in a hostile work environment case. Journal of Applied Psychology 1999;84:362–375. [PubMed: 10380417]

Kovera MB, Russano MB, McAuliff BD. Assessment of the commonsense psychology underlying *Daubert*: Legal decision makers' abilities to evaluate expert evidence in hostile work environment cases. Psychology, Public Policy, and Law 2002;8:180–200.

Kumho Tire Co., Ltd.. et al. Carmichael. et al. 526 U.S. 137. 1999.

Levett LM, Kovera MB. The effectiveness of opposing expert witnesses for educating jurors about unreliable expert evidence. Law and Human Behavior. 2007 http://www.springerlink.com. doi: 10.1007/s10979-007-9113-9.

McAuliff BD, Kovera MB. Juror need for cognition and sensitivity to methodological flaws in expert evidence. Journal of Applied Social Psychology 2008;38:385–408.

Mill D, Gray T, Mandel DR. Influence of research methods and statistics courses on everyday reasoning, critical abilities, and belief in unsubstantiated phenomena. Canadian Journal of Behavioural Science 1994;26:246–258.

Niedermeier KE, Kerr NL, Messé LA. Jurors' use of naked statistical evidence: Exploring bases and implications of the Wells effect. Journal of Personality and Social Psychology 1999;76:533–542.

Petty RE, Cacioppo JT. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. Journal of Personality and Social Psychology 1984;46:69–81.

Petty, RE.; Cacioppo, JT. The elaboration likelihood model of persuasion. In: Berkowitz, L., editor. Advances in experimental social psychology. Vol. Vol. 19. Academic Press; New York: 1986. p. 123-203.

Petty RE, Cacioppo JT, Goldman R. Personal involvement as a determinant of argument based persuasion. Journal of Personality and Social Psychology 1981;41:847–855.

Ratneshwar S, Chaiken S. Comprehension's role in persuasion: The case of its moderating effect on the persuasive impact of source cues. Journal of Consumer Research 1991;18:52–62.

Robinson. Jacksonville Shipyards Inc.. 760 F.Supp. 1486. 1991. M.D. Fla.

Rudman LA, Borgida E. The afterglow of construct accessibility: The behavioral consequences of priming men to view women as sexual objects. Journal of Experimental Social Psychology 1995;31:493–517.

Schklar J, Diamond SS. Juror reactions to DNA evidence: Errors and expectancies. Law and Human Behavior 1999;23:159–184.

Smith BC, Penrod SD, Otto AL, Park RC. Jurors' use of probabilistic evidence. Law and Human Behavior 1996;20:49–82.

Thompson WC, Schumann EL. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. Law and Human Behavior 1987;11:167–187.

Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. Science 1974;185:1124–1131. [PubMed: 17835457]

Wells GL. Naked statistical evidence of liability: Is subjective probability enough? Journal of Personality and Social Psychology 1992;62:739–752.

Table 1

Individual items used to create “expert evidence quality” composite variable

1. Dr. Johnson's research was based on good scientific principles.
2. Dr. Johnson's study contained appropriate measures of sexual harassment.
3. Dr. Johnson's testimony was helpful in reaching my verdict.
4. Dr. Johnson's study was scientifically valid.
5. Dr. Johnson used appropriate scientific procedures in her study.
6. The findings from Dr. Johnson's study can be used to understand what occurred in the garage at Sunshine Trucking Company.
7. Dr. Johnson did not use valid measures of sexual harassment in her study. (R)
8. The findings from Dr. Johnson's study cannot be used to understand why sexual harassment occurs in actual work settings. (R)
9. The scientific evidence on sexual harassment that I heard in this trial is reliable.

Note: Items followed by (R) were reverse scored

Table 2

Means and univariate effects of internal validity on expert evidence quality, expert credibility, and plaintiff credibility

Dependent measure	Means (SD)				Univariate effect of internal validity			
	Valid	No control	Confound	Experimenter bias	F	df	p	η^2
Expert evidence quality	4.55 (1.07)[a]	3.86 (1.11)[a]	4.29 (0.90)	4.13 (1.19)	4.01	3, 239	0.02	0.07
Expert credibility	5.52 (0.92)[b]	5.01 (1.04)[b]	5.37 (0.99)	5.17 (0.96)	2.95	3, 239	0.05	0.06
Plaintiff credibility	3.89 (0.89)	3.79 (0.98)	3.74 (0.98)	3.88 (1.07)	0.38	3, 239	0.50	0.03

Notes: Differences between means sharing the same superscript within each row were statistically significant at $p \leq .05$

Expert evidence quality, expert credibility, plaintiff credibility were all composite variables

η^2 = partial η^2

Table 3

Means, standard deviations, and correlations among the dependent measures

	M	SD	1	2	3	4
1. Verdict	0.37	0.48	-			
2. Expert evidence quality	4.22	1.09	0.36*	-		
3. Expert credibility	5.28	1.00	0.04	0.37*	-	
4. Plaintiff credibility	3.83	0.97	0.63*	0.43*	0.11	-

Notes:

Verdict = percentage of judgments in favor of plaintiff

Expert evidence quality, expert credibility, plaintiff credibility were all composite variables

* correlation significant at $p \leq .05$

Law Hum Behav. Author manuscript; available in PMC 2010 April 28.