Law Hum Behav (2010) 34:489–500
DOI 10.1007/s10979-010-9219-3

ORIGINAL ARTICLE

I Spy with My Little Eye: Jurors' Detection of Internal Validity Threats in Expert Evidence

Bradley D. McAuliff · Tejah D. Duckworth

Published online: 17 February 2010

Abstract This experiment examined whether jury-eligible community members ($N = 223$) were able to detect internally invalid psychological science presented at trial. Participants read a simulated child sexual abuse case in which the defense expert described a study he had conducted on witness memory and suggestibility. We varied the study's internal validity (valid, missing control group, confound, and experimenter bias) and publication status (published, unpublished). Expert evidence quality ratings were higher for the valid versus missing control group version only. Publication increased ratings of defendant guilt when the study was missing a control group. Variations in internal validity did not influence perceptions of child victim credibility or police interview quality. Participants' limited detection of internal validity threats underscores the need to examine the effectiveness of traditional legal safeguards against junk science in court and improve the scientific reasoning ability of lay people and legal professionals.

Keywords Scientific reasoning · Internal validity · Expert testimony · Juror decision-making

Portions of this research were presented at the April 2008 meeting of the Western Psychological Association in Irvine, CA and the March 2009 meeting of the American Psychology-Law Society in San Antonio, TX.

B. D. McAuliff (✉) · T. D. Duckworth
Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8255, USA
e-mail: bradley.mcauliff@csun.edu

The United States legal system has taken serious precautions against the proliferation of junk science in court. In *Daubert v. Merrell Dow Pharmaceuticals, Inc*. (1993), the Supreme Court held that judges must serve as gatekeepers and evaluate the relevance and reliability of expert evidence before admitting it at trial. *Daubert* enumerated four factors that judges should consider when evaluating evidentiary reliability: whether the theory or methodology underlying the expert's opinion is falsifiable, possesses a known or potential rate of error, has been peer-reviewed and published, and is generally accepted within the relevant scientific community. Judges essentially are expected to evaluate the scientific validity of expert evidence when determining its evidentiary reliability.

Daubert requires that judges must be sophisticated consumers of science in order to render effective admissibility decisions regarding expert evidence. However, many law school curricula do not include training on the scientific method (Merlino, Dillehay, Dahir, & Maxwell, 2003) and longitudinal research indicates that students' statistical and methodological reasoning skills do not improve over the course of law school (Lehman, Lempert, & Nisbett, 1988). Based on this lack of training, it is not surprising that 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, 2000). When asked to describe how they would apply the *Daubert* factors in determining the admissibility of expert testimony, only 5% of responding state court judges demonstrated a clear understanding of falsifiability and only 4% sufficiently understood the concept of error rate in a national survey by Gatowski et al. A second study found that variations in the internal validity of an expert's research did not influence judges' simulated admissibility decisions (Kovera &

McAuliff, 2000). Judges in that study read a brief fact pattern of a hostile work environment case and a description of the expert's testimony that the plaintiff wished to present. When asked whether they would admit the expert testimony, judges were no more likely to admit a valid study (17%) than they were to admit a study that lacked a control group (11%), contained a confound (17%), or included the potential for experimenter bias (24%).

Despite these limitations, judges overwhelmingly support their role as gatekeepers (Gatowski et al., 2001) and believe that they are better able to evaluate scientific evidence than jurors (Kovera & McAuliff, 2000). In fact, judges appear quite confident in their gatekeeping abilities. Nearly 80% of judges responding to a three-state survey indicated that expert testimony was rarely too technical for them to understand (Shuman, Whitaker, & Champagne, 1994). Moreover, even though the vast majority of judges in the Gatowski et al. study were unable to demonstrate a clear understanding of the falsifiability and error rate concepts, only a small percentage asked interviewers for a definition or further explanation. When judges fail to recognize their limited scientific literacy, they may admit invalid expert testimony. As such, the ability of attorneys and jurors to evaluate scientific validity effectively becomes increasingly critical.

Attorneys in particular must be able to identify invalid research, file motions to exclude that research, and successfully argue the basis for their motions before the court (Kovera, Russano, & McAuliff, 2002). If a motion is denied and the evidence is admitted, attorneys must understand factors affecting scientific validity in order to cross-examine the expert witness effectively. Unfortunately, recent research suggests that attorneys may be no more adept at identifying flawed psychological science than judges (Kovera & McAuliff, 2009). In that study, attorneys read a basic fact pattern and proposed expert testimony in which the internal validity (valid, missing control group) and general acceptance (peer-reviewed, published, and generally accepted in the scientific community or not) of a psychological study were varied. Internal validity did not affect attorneys' decisions to file a motion to exclude the expert's testimony or their ratings of the study's scientific reliability, and none of the manipulated variables influenced attorneys' ratings of whether they would consult their own expert in preparation for trial in the simulated case. In contrast, general acceptance did affect attorneys' ratings of the expert evidence. Attorneys found the generally accepted study to be more scientifically reliable than the study that was not generally accepted.

If judges and attorneys are unable to identify internal validity threats in psychological science as past research suggests, jurors must rely on their own determinations of expert evidence quality to make decisions at trial. Can jurors detect internally invalid research even if judges and attorneys cannot?

Jurors' Ability to Evaluate Psychological Science

Basic psychological research has identified deficits in lay people's understanding and use of statistical, probabilistic, and methodological information (Fong, Krantz, & Nisbett, 1986; Hamill, Wilson, & Nisbett, 1980; Kahneman & Tversky, 1973). More recently, psychologists have questioned whether these findings generalize to more applied settings, such as the legal system, by examining whether jurors can identify internal validity threats in expert evidence (Levett & Kovera, 2008; McAuliff & Kovera, 2008; McAuliff, Kovera, & Nunez, 2009). These studies have focused largely on one of the most fundamental threats to internal validity: a missing control group. McAuliff and Kovera (2008) varied the presence of a control group and publication status of a study that was described by the plaintiff's expert in a simulated hostile work environment case. Mock 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. Similar findings were observed by Levett and Kovera (2008); however, differences in mock jurors' ratings of scientific quality for the valid versus missing control group study emerged only when they received opposing expert testimony that focused on the study's methodology and those differences were only marginally significant. Based on these studies, it appears that jurors may be able to detect internally invalid research only under special circumstances (i.e., when they are high in NC or when methodologically focused testimony is provided by an opposing expert).

Only one study to date has focused on internal validity threats that are more methodologically sophisticated in nature than a missing control group. McAuliff et al. (2009) presented jurors expert testimony on behalf of the plaintiff in a simulated hostile work environment case. The expert's study examined the effects of viewing sexualized advertisements on men's behavior toward women. Within that testimony McAuliff and colleagues varied whether the expert's study was valid, missing a control group, contained a confound, or included the potential for experimenter bias. Those variables were operationalized in the expert's study by having male participants randomly assigned to a sexualized or nonsexualized advertisement condition versus a sexualized advertisement condition only (missing control group). After viewing the advertisements,

male participants were interviewed by a female confederate who either knew what experimental condition the men had been assigned to (experimenter bias) or not. A second female confederate was added to the expert's study to create a confound by having her interview only men who had seen the sexualized advertisements and having the other female confederate interview only men who had seen the nonsexualized advertisements. The valid version of the study contained none of the internal validity threats. Jury-eligible community members participating in McAuliff et al.'s experiment rated the quality of the expert evidence and credibility of the expert higher for the valid versus missing control group version of the study only. No differences emerged on verdict as a function of internal validity. These results suggest that jurors may be sensitive to basic, but not more complex, threats to internal validity.

The goal of the present study is to advance the scientific literature on jurors' ability to detect internally invalid psychological science presented by an expert at trial. The results to date are mixed, with one experiment indicating that jurors are sensitive to the lack of an appropriate control group and two others indicating this sensitivity only occurs under special circumstances. To achieve this goal, we examined whether jury-eligible community members were able to detect basic and sophisticated threats to internal validity in a simulated criminal child sexual abuse case that included defense expert testimony on witness memory and suggestibility. As described earlier, almost all of the research in this area has focused solely on the missing control group threat to internal validity using a simulated civil hostile work environment case in which the expert testified for the plaintiff about issues and research related to sexual harassment. It remains to be seen whether the findings observed in earlier studies generalize to other settings, stimulus materials, and operationalizations of the key independent variables.

Dual-Process Models of Persuasion and Juror Decision-Making

Two information processing models from the social-cognitive literature on persuasion provide a much needed theoretical framework to predict how jurors make decisions when confronting psychological science in court. These models are the heuristic–systematic model (HSM; Chaiken, 1980; Chaiken, Liberman, & Eagly, 1989) and the elaboration likelihood model (ELM; Petty & Cacioppo, 1986). Both models propose that people seek to hold correct attitudes and are willing to engage in varying levels of cognitive effort (i.e., information processing) to satisfy this goal. Systematic (HSM) or central (ELM) processing is characterized by a high level of cognitive effort and entails careful scrutiny of persuasive message content. When engaging in systematic processing, people evaluate the quality of the arguments presented in the persuasive message. Systematic/central processors are more likely to adopt the position advocated in the persuasive message when it contains valid, high-quality arguments than when it does not (Petty, Cacioppo, & Goldman, 1981). In contrast, individuals who engage in heuristic (HSM) or peripheral (ELM) processing do not scrutinize the quality of the persuasive arguments. Instead, they rely on more superficial mental shortcuts or decision rules to evaluate a persuasive message. Certain cues associated with the persuasive message (e.g., length or number of arguments; Petty & Cacioppo, 1984), its source (e.g., expertise, likability, physical attractiveness; Chaiken & Maheswaran, 1994), and the audience (e.g., positive or negative audience reactions; Axsom, Yates, & Chaiken, 1987) may affect message evaluation in heuristic/peripheral processing.

Systematic Versus Heuristic Processing of Psychological Science

According to the HSM and ELM, two factors that moderate information processing are ability and motivation (Chaiken, 1980; Petty & Cacioppo, 1986). Systematic/central processing requires that an individual is both *able* and *motivated* to scrutinize the quality of arguments contained in the persuasive message. When ability or motivation is low, that individual is likely to rely on heuristic/peripheral processing to make message-related judgments. Factors shown to influence one's *ability* to process a persuasive message systematically/centrally include information complexity, prior knowledge or experience, distraction, and repetition (Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1986; Ratneshwar & Chaiken, 1991). Factors shown to affect one's *motivation* to process systematically/centrally include personal relevance, personal responsibility, and NC (Cacioppo, Petty, Feinstein, & Jarvis, 1996; Chaiken et al., 1989; Petty et al., 1981). As a juror, it is one's civic duty to be motivated to evaluate evidence in a thoughtful, considerate manner; however, this motivation does not necessarily imply that jurors are in fact able to do so.

If ability or motivation to process systematically/centrally is low, an individual is more likely to engage in heuristic/peripheral processing than when ability or motivation is high. One heuristic jurors might use when evaluating expert evidence is that "consensus implies correctness." When processing persuasive messages, people rely on others' evaluations of message quality (Axsom et al., 1987) and consensus information can influence people's judgments of message quality under conditions that produce heuristic processing (Maheswaran & Chaiken,

1991; Tversky & Kahneman, 1974). In the legal domain, jurors may rely on information about a study's general acceptance within the scientific community when evaluating its quality. Jurors may reason that research is methodologically sound if it has been published in a peer-reviewed journal and therefore has been evaluated favorably by qualified members of the relevant scientific community. In contrast, jurors may view research that has not been published or generally accepted negatively. Although this heuristic may lead jurors to make reasonable decisions about psychological science most of the time, jurors may be led astray by evidence of general acceptance in some instances (see Kassin, Ellsworth, & Smith, 1989, for an example of how the general acceptance of a phenomenon—the reliability of showups—was not supported by research).

Two recent experiments varied the general acceptance/publication status of an expert's study to determine its influence on mock jurors' decisions in a simulated hostile work environment case (Kovera, McAuliff, & Hebert, 1999; McAuliff & Kovera, 2008). Mock jurors used publication status as a heuristic cue to evaluate study quality and expert trustworthiness in the Kovera et al. study; however, McAuliff and Kovera (2008) did not replicate these effects. Neither verdicts nor perceptions of expert credibility differed as a function of publication status in either study. Hence, consensus information in the form of published research may affect some trial-related judgments, but not others.

Overview and Hypotheses

The present experiment examined two questions related to jurors' ability to detect internal validity threats in psychological science presented by an expert at trial. Can jurors distinguish between an internally valid study and one that contains a missing control group, confound, or experimenter bias? If they cannot, do jurors instead rely on heuristic cues involving the study's publication status when evaluating its quality? We sought to answer these research questions by presenting jury-eligible community members a simulated child sexual abuse case that contained defense expert testimony about a witness suggestibility study he had conducted. We varied the internal validity (valid, missing control group, confound, and experimenter bias) and publication status (published in a peer-reviewed journal or not) of the expert's study to explore the potential effects of these variables on mock jurors' evaluations of expert evidence quality and other trial-related judgments.

We generated two hypotheses based on our research questions and previous studies of jurors' scientific reasoning ability. Consistent with McAuliff et al. (2009) and McAuliff and Kovera (2008), we predicted that mock jurors would be sensitive to the absence of a control group 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 with mock jurors rating the expert evidence quality of the valid version higher than the missing control group version, but no different from the confound and experimenter bias versions. Second, given their limited ability to process the more sophisticated internal validity threats of a confound and experimenter bias (McAuliff et al., 2009), we predicted that mock jurors in those experimental conditions would rely on the study's publication status as a heuristic cue to evaluate expert evidence quality. This hypothesized finding would be consistent with the findings of Kovera et al. (1999). Support for this hypothesis would consist of a statistically significant interaction between the study's internal validity and publication status. Mock jurors' expert evidence quality ratings should be higher for the published versus unpublished version of the expert's study when it contains a confound or experimenter bias. No such differences should emerge for the valid or missing control group versions.

Method

Participants

Two hundred and twenty-three community members residing in southern California participated in our study in exchange for $10.00. 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 old, able to understand English, and who has not been convicted of a felony (*California Code of Civil Procedure*, §203).

On average participants were 34 years old and female (51%). Seventy-one percent had never served on a jury, 79% had never been involved in legal proceedings, 75% had never been the victim of sexual abuse, and 93% had never been wrongly accused of sexual abuse. Members of various ethnic groups participated including: South/Central American, Hispanics, Mexicans (34%), Caucasians (27%), Asians (16%), Black, nonHispanic (8%), Middle Eastern (8%), Native Americans (2%), and Others (5%). A slight majority of participants did not have children (53%).

Trial Stimulus

Participants read a 16 page summary of a simulated child sexual abuse case in which the victim alleged inappropriate sexual touching by her stepfather. Details of the abuse were derived from an actual case (*United States v. LeBlanc*, 2002). The summary included opening statements and closing arguments from both attorneys, direct- and cross-examined testimony from the child victim, police officer, defendant and expert, and standard California judicial instructions.

The victim was a 10-year-old girl who testified that on the afternoon in question she went into the living room after awaking from a nap and sat on the couch where her stepfather was watching television. The victim described how the stepfather forced her to touch his penis while they were home alone. She explained that the defendant jumped up and ran out of the room when he heard his wife (the victim's mother) come in the front door of the house. On cross-examination, the victim admitted that she missed her real dad and she wished her parents could get back together. She stated how she remembered the details of the event pretty well, but was confused by answering all the questions from her mom, the police officer, and in court. A police officer also testified for the prosecution. She described how she asked the victim open- and close-ended questions regarding what the accused said and did, how long the victim knew her stepfather, if he had been abusive in the past, and the television show he was watching on the night in question. The officer acknowledged during cross-examination that she had not received any specialized training for interviewing alleged victims of sexual abuse and that interviews were not routinely recorded in her precinct.

The defendant testified that he became aroused while watching a television program and began to masturbate, but was interrupted when his stepdaughter (who he thought was taking a nap) entered the room and sat down on the couch next to him. He described how he quickly covered himself up and while yelling at her to leave the room, he heard his wife at the front door. On cross-examination, the defendant stated that he was surprised about the allegations because the young girl had never lied about him before. A cognitive psychologist testified as an expert witness for the defense also.

Experimental Manipulations

The defense expert witness explained that the majority of his research had focused on the suggestibility of witness memory. The expert described his most recent study in which 200 10-year-old children participated in a wellness exam administered by a medical student who measured each child's pulse, blood pressure, heart/respiration rate, temperature, and spinal curvature. He explained that these tasks were chosen because of their similarity to behaviors that often accompany child sexual abuse (e.g., asking child to remove clothes, adult/child physical contact). After the wellness exam, children were questioned about details of the event by a research assistant posing as a nurse. The expert concluded his testimony by noting that police interviews sometimes contain suggestive or misleading information because officers who were not present during the alleged event must question witnesses. He reported that he had reviewed a transcript of the officer's interview of the victim and that it did contain suggestive questions similar to those included in his research. Within the expert's description of his study, we manipulated its internal validity and publication status.

Design

This study used a 4 Internal Validity (Valid, Missing control group, Confound, and Experimenter bias) $\times$ 2 Publication Status (Published in a peer-reviewed journal or not) 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 publication status. With the exception of these manipulations, all information presented in the different versions of the trial stimulus was identical.

Internal Validity

Valid. The first version of the study contained no internal validity threats. The research assistant who posed as the nurse interviewed children about their memories of the exam. She knew nothing about the purpose of the experiment or its predicted results—only that her role was to interview the children. The research assistant asked half of the children a series of neutral questions that contained no misleading information and she asked the remaining children a series of suggestive questions that contained misleading information. Both interviews contained an equal number of questions focusing on central details (e.g., what the medical student did) and peripheral details (e.g., the appearance of the exam room). The order and focus (central, peripheral) of the questions were counterbalanced across conditions.

Missing Control Group. The missing control group condition was the same as the valid condition except that no comparison group of nonmisleading questions was included. Children were asked suggestive questions only.

Confound. This condition introduced a confound to the internally valid version of the expert's study: The research assistant asked half of the children a series of *neutral* questions that contained no misleading information

and focused exclusively on *central details* of the exam. She asked the remaining children a series of *suggestive* questions that contained misleading information and focused exclusively on *peripheral details*. Thus, question type (neutral, misleading) and focus (central, peripheral) were confounded. If differences between the two groups emerged, the experimenter would not know whether they were the result of the type of question asked or its focus. Except for the confound, this version of the expert's testimony was the same as the valid condition.

Experimenter Bias. The experimenter bias condition was identical to the valid condition except that the research assistant was informed about the purpose of the experiment and its predicted findings prior to interviewing the children.

Publication Status

We also manipulated whether the study had been published in a peer-reviewed journal as a potential heuristic cue to its internal validity. In the published condition, the expert reported that his findings had been published in a prestigious journal after being favorably reviewed by other psychologists in the field. In the unpublished condition, the expert reported that he had just recently completed the study and therefore it had not been reviewed by other experts in the field or published in a scientific journal.

Dependent Measures

Participants decided whether the prosecution had demonstrated beyond a reasonable doubt that the defendant had committed a "lewd and lascivious act" against his stepdaughter in violation of California law. Participants rendered their decisions using a dichotomous verdict variable (guilty, not guilty). Then participants rated the defendant's guilt on a 7-point Likert-type scale (1 = Certainly Innocent, 7 = Certainly Guilty).

Participants rated the quality of the expert's evidence based on a series of 7-point Likert-type scales where 1 = Strongly Disagree and 7 = Strongly Agree. Expert evidence quality was measured by whether the expert's research was perceived as being valid, reliable, based on good scientific principles, and included appropriate measures of witness suggestibility. Participants' ratings were averaged across these four items to form a single composite measure of expert evidence quality (Cronbach's alpha = .81).

Participants rated the quality of the child victim's testimony on a series of 7-point Likert-type scales. These items measured participants' perceptions of child victim accuracy, reliability, credibility, suggestibility (R), honesty, truthfulness, and motivation to lie (R). Items followed by (R) were recoded so that smaller numbers represented more negative evaluations of the child victim and larger numbers represented more positive evaluations. Participants' ratings were averaged across these seven items to form a single composite measure of child victim credibility (Cronbach's alpha = .90).

Participants rated the quality of the police officers' interview on a series of 7-point Likert-type scales. These items measured participants' perceptions of whether the police officer's interview of the child victim was fair, biased (R), suggestible (R), and successful at obtaining the truth. Participants' ratings were averaged across these four items to form a single composite measure of police interview quality (Cronbach's alpha = .77).

Four additional items served as manipulation checks for the internal validity and publication status variables. With respect to internal validity, participants responded to three forced-choice questions asking what type of questions the expert included in her study (suggestive only versus both suggestive and nonsuggestive), which of two statements best described the experiment ("Only children in the suggestive interview condition were asked about peripheral details of the exam" versus "All children were asked questions focusing on central and peripheral details of the exam"), and whether the research assistant knew the experiment's purpose and expected findings (yes or no). The publication status manipulation check asked participants to indicate whether the study had been published in a peer-reviewed journal or not.

Mock jurors concluded their participation by providing demographic information about their age, gender, racial/ethnic identity, jury eligibility, history of jury service, previous involvement in legal proceedings (civil or criminal, plaintiff or defendant), number of children and their ages, and the estimated frequency of interaction with children on a daily basis.

Procedure

Participants were tested individually. Upon arrival, they read standardized instructions and an informed consent sheet describing what the experiment would entail. After providing informed consent, participants received the trial stimulus and dependent measures. Upon completion, participants were asked if they had any questions, debriefed, and paid $10.00.

Results

Manipulation Checks

Mock jurors were sensitive to the internal validity and publication status manipulations included in our study.

Missing Control Group. Participants who read the missing control group version of the study were more likely to correctly report that children only received suggestive questions (93%) than that children had received both suggestive and nonsuggestive questions (7%), χ^2 (1, $N = 59$) = 44.09, $p < .05$, $\Phi = .86$.

Confound. Participants who read the confound version of the study were more likely to correctly report that only children in the suggestive interview condition were asked about peripheral details of the exam (86%) than that all children were questioned about both central and peripheral details (14%), χ^2 (1, $N = 58$) = 30.41, $p < .001$, $\Phi = .72$.

Experimenter Bias. Participants who read the experimenter bias version of the study were more likely to correctly report that the research assistant knew the purpose of the experiment and its predicted findings (90%) than that she did not know (10%), χ^2 (1, $N = 58$) = 36.48, $p < .001$, $\Phi = .79$.

Publication Status. Participants who were informed that the expert's study had been published in a peer-reviewed journal were more likely to correctly report that the study had been published (92%) than were participants in the unpublished study condition (8%), χ^2 (1, $N = 113$) = 79.87, $p < .001$, $\Phi = .84$.

Data Analytic Strategy and Primary Results

We began by subjecting mock jurors' dichotomous verdicts to a logistic regression. We regressed verdict on internal validity, publication status, and the interaction of these two variables. Neither the main effects nor interaction were statistically significant, χ^2 (3, $N = 223$) = 1.57, $p = .66$, $\Phi = .08$ (see Table 1 for means).

Next we subjected the data from the defendant guilt, expert evidence quality, child victim credibility, and police interview quality continuous measures to a multivariate analysis of variance (MANOVA). We used the Pillai's Trace criterion multivariate statistic to test the significance of all main effects and interactions. A 4 Internal Validity × 2 Publication Status MANOVA revealed a statistically significant main effect for the study's internal validity, Mult. $F(12, 627) = 2.68$, $p = .002$, partial $\eta^2 = .05$. Results also revealed a statistically significant interaction effect between the study's internal validity and publication status, $F(12, 627) = 1.97$, $p = .03$, partial $\eta^2 = .04$. The publication status main effect was not statistically significant.

We followed up the significant multivariate effects using univariate F-tests for each of the four dependent measures and Tukey's HSD comparisons when appropriate. Only the tests for expert evidence quality and defendant guilt reached traditional levels of statistical significance (see Table 2 for results).

Expert Evidence Quality. The main effect of internal validity on expert evidence quality was statistically significant (see Table 2 for means). Mock jurors rated the valid version of the expert's study higher in evidence quality than the missing control group version only. Evidence quality ratings for the confound and experimenter bias versions did not differ from each other or any of the other conditions. Neither the publication status main effect nor the interaction effect was statistically significant.

Defendant Guilt. The interaction between internal validity and publication status on defendant guilt was statistically significant, $F(3, 210) = 3.24$, $p = .02$, partial $\eta^2 = .04$ (see Table 3 for means). Mock jurors in the missing control group condition believed the defendant was more guilty when the expert's study was published versus unpublished. The difference between the published and unpublished study was not statistically significant for any of the other internal validity conditions, nor was either main effect.

We reran the MANOVA and univariate F-tests using only participants who answered the internal validity and publication status manipulation checks correctly. The

Table 1 Means and standard deviations for effects of internal validity and publication status on verdict

	Means (SD)				Total
	Internal validity				
	Valid	Missing control	Confound	Experimenter bias	
Publication status					
Published	.62 (.50)	.84 (.37)	.59 (.50)	.59 (.50)	.66 (.48)
Unpublished	.67 (.48)	.61 (.50)	.50 (.51)	.75 (.44)	.63 (.48)
Total	.65 (.48)	.73 (.45)	.55 (.50)	.67 (.48)	

Table 2 Means and univariate main effects of internal validity on defendant guilt, expert evidence quality, child victim credibility, and police interview quality

Dependent measure	Means (SD)				Univariate effect of internal validity			
	Valid	Missing control	Confound	Experimenter bias	F	df	p	$\partial\eta^2$
Defendant guilt	4.92 (1.77)	5.51 (1.76)	5.04 (1.56)	5.27 (1.72)	1.11	3, 210	.35	.02
Expert evidence quality	4.85 (1.03)[a]	3.77 (1.43)[a]	4.28 (1.41)	4.32 (1.23)	5.68	3, 210	.001	.08
Child victim credibility	4.93 (1.42)	4.89 (1.18)	4.56 (1.24)	4.93 (1.23)	1.12	3, 210	.34	.02
Police interview quality	4.58 (1.30)	4.62 (1.15)	4.12 (1.14)	4.46 (1.28)	1.89	3, 210	.13	.03

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

Expert evidence quality, child victim credibility, and police interview quality were all composite variables

Table 3 Means for internal validity × publication status interaction effect on defendant guilt

Publication status	Means (SD)			
	Valid	Missing control	Confound	Experimenter bias
Published	5.00 (1.95)	6.13 (1.46)[a]	4.90 (1.70)	5.04 (1.88)
Unpublished	4.85 (1.03)	4.81 (1.84)[a]	5.19 (1.39)	5.50 (1.55)

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

results and pattern of effects were consistent with those obtained when the entire participant sample was included.

Discussion

We designed the present study to examine two research questions involving jurors' sensitivity to internally invalid psychological science presented by an expert at trial. Can jurors identify an internally invalid study? And if they cannot, do they instead rely on other heuristic cues associated with the expert's research when judging its quality? We discuss the specific results relevant to each hypothesis and conclude by considering the limitations of our work as well as its implications for the scientific reasoning literature and trials containing psychological science presented by experts.

Can Jurors Detect Internal Validity Threats?

Consistent with our first hypothesis, mock jurors in our study were able to detect the missing control group threat to internal validity but not the confound and experimenter bias threats. These findings mirror those observed by McAuliff et al. (2009) and provide support for the generalizability of jurors' sensitivity to missing control group information across different trials (civil hostile work environment case, criminal child sexual abuse case), sources of expert evidence (plaintiff, defendant), nature of the expert evidence (factors that increase sexual harassment in the workplace, factors that increase witness suggestibility), and operationalizations of missing control group information (absence of men being exposed to nonsexualized advertisements, absence of children receiving nonsuggestive questions). At the same time, these findings are slightly at odds with previous research indicating that jurors are only sensitive to missing control group information under special circumstances such as when they are high in NC (McAuliff & Kovera, 2008) or when they receive methodologically oriented testimony from an opposing expert (Levett & Kovera, 2008).

Why did mock jurors recognize the absence of control group information in our study even when these "special circumstances" did not exist? First, with respect to McAuliff and Kovera (2008), mock jurors in our study did not complete the NC scale, so we do not know whether their levels of NC were similar to or different from mock jurors in the earlier study. Even if we had included this scale, high versus low NC is determined by performing a median-split on the observed NC scores. As such, actual NC scores and the corresponding high/low classification can vary considerably across different samples. This is not problematic when two samples of NC scores are normally distributed; however, if one sample of scores is either positively or negatively skewed and the other is not, any comparison of high/low NC between the two samples is much less meaningful.

It is more difficult to reconcile the finding that our mock jurors were sensitive to the absence of control group information whereas mock jurors in Levett and Kovera's (2008) experiment were not unless they received additional testimony from an opposing expert who highlighted the importance of a control group in an internally valid study. The nature and format of the written child sexual abuse trial stimulus used in both experiments was highly similar, as was the operationalization of the missing control group threat to internal validity and the community member population sampled, so these features provide little insight into why the inconsistent results may have emerged. What

was different across studies, however, were the exact details of the case, the number and type of witnesses who testified, the actual suggestibility experiment described by the expert, and the number and type of items used to measure expert evidence quality. Perhaps one or more of these differences influenced jurors' sensitivity to the internal validity of the expert's study.

It is also crucial to realize that even though jurors were sensitive to the missing control group threat to internal validity, this sensitivity did not bleed into jurors' ratings of child victim credibility or police interview quality. No differences emerged on these dependent measures as a function of the internal validity manipulation. This is reassuring because recall that the description of the child's testimony and police interview were held constant across all experimental conditions. If differences would have emerged, this could be evidence of a skepticism or confusion effect for the expert's testimony (Cutler, Penrod, & Dexter, 1989). In other words, jurors would have used expert evidence quality as a proxy to differentially evaluate child victim credibility and police interview quality even though these trial features were identical in all conditions. Such an effect would be undesirable and could support a motion to exclude the expert's testimony under Federal Rules of Evidence Rule 403, which states that even relevant evidence can be excluded if its probative value is "substantially outweighed" by danger of unfair prejudice or confusion of the issues.

Beyond the missing control group issue, mock jurors were insensitive to the confound and experimenter bias threats to internal validity across all four dependent measures. These results are consistent with previous work by McAuliff et al. (2009) and Kovera et al. (1999). Before considering potential reasons why mock jurors struggle with these internal validity threats, we must rule out alternative explanations for the null effects. First, recall that manipulation checks demonstrated that mock jurors recognized that children in the suggestive interview condition were questioned about peripheral details of the event and that children in the nonsuggestive interview condition were questioned about central event details (confound). Similarly, mock jurors correctly identified whether the research assistant who interviewed children knew in advance the purpose of the experiment and its predicted findings (experimenter bias). The statistically significant effect for the missing control group version of the expert's study also helps eliminate the possibilities of insufficient power and insensitive dependent measures to detect differences for the other versions. Indeed, additional calculations confirm that our study had adequate power (.88) to detect a medium-sized effect given the number of participants in our study and $\alpha = .05$. What seems more likely is that, similar to arguments made by McAuliff et al. (2009), internal validity threats in the form of confounds and experimenter bias may require a more sophisticated knowledge of research methodology because they are inherently more difficult to comprehend compared to the absence of control group information. As such, lay people may require more specialized training to understand these methodologically sophisticated internal validity threats and to identify them in real-world contexts.

Did Jurors Who Failed to Detect Certain Internal Validity Threats Instead Rely on the Consensus Heuristic to Render Decisions?

Our second hypothesis predicted that mock jurors who did not systematically/centrally process the expert evidence would rate it's quality higher when the study was published in a peer-reviewed journal than when it was not. Our data only partially supported this hypothesis and not in the manner that we predicted. Publication status did affect mock jurors' ratings of defendant guilt, but only in the missing control group condition. Mock jurors in this condition provided higher ratings of defendant guilt when the expert's study was published versus unpublished whereas publication status did not affect jurors' ratings in any of the remaining internal validity conditions.

This interaction effect is noteworthy in two respects. Although we predicted that mock jurors would attend to publication status only when they were unable to process systematically/centrally, both the HSM (Chaiken et al., 1989) and ELM (Petty, Kasmer, Haugtvedt, & Cacioppo, 1987) propose that systematic and heuristic processing may occur *simultaneously*. Under such conditions, the two processing modes can attenuate or bolster one another depending on the congruency between the heuristic cue and argument quality (Maheswaran, Mackie, & Chaiken, 1992). Mock jurors in our study relied on both the publication status of the expert's study (heuristic cue) *and* its internal validity (argument quality). Specifically, participants in the missing control group condition rated defendant guilt higher when the expert's study had been published versus unpublished. Although this may seem counterintuitive, keep in mind that the expert testified on behalf of the defense, so a positive effect of expert testimony would result in decreased perceptions of guilt and a negative effect would result in increased perceptions of guilt. Having an expert who had published an internally invalid study testify on behalf of the defendant actually hurt, and not helped, his case. In considering why this effect only was present for the missing control group condition, we must keep in mind that mock jurors' evidence quality ratings revealed that they were only sensitive to this single internal validity threat. Given their apparent inability to systematically process the confound and

experimenter bias threats to internal validity, the only conditions in which both heuristic and systematic processing could co-occur was the published versus unpublished missing control group study. That said, this interaction effect does not appear to be very robust because it only emerged on the defendant guilt dependent measure.

The question still remains, however, why mock jurors in the confound and experimenter bias conditions did not rely on the heuristic cue of publication status when they failed to process systematically. Although inconsistent with the results of Kovera et al. (1999), mock jurors in another study also failed to rely on publication status as hypothesized (McAuliff & Kovera, 2008). Several potential explanations for these contradictory findings exist, all of which are admittedly post-hoc and speculative. It is possible, for example, that mock jurors who were unable to process systematically relied on other heuristics or simplified decision-making strategies to evaluate message quality. One common heuristic from the basic social psychological literature involves source expertise (Chaiken & Maheswaran, 1994; Petty et al., 1981). This heuristic may be particularly common in legal settings where experts must meet certain evidentiary rules or criteria before giving testimony in court. It seems unlikely that jurors will have a sophisticated understanding of these rules/criteria; however, they may intuit that the judge allowed the expert to testify because he or she has something relevant and reliable to say. As a result, jurors may defer to the psychologists' expertise and take the testimony at face value instead of systematically scrutinizing evidence quality (Cooper, Bennett, & Sukel, 1996; Cooper & Neuhaus, 2000).

It is also possible that participants in our study attended to publication status, but simply made mistakes when evaluating it or used the information in ways we did not anticipate. As one reviewer noted, the unpublished study condition explained that the expert had just completed the research and therefore it had not been peer-reviewed or published yet. Perhaps mock jurors' evaluations of the unpublished study were more favorable because they believed the research was brand new and more cutting edge than the published study. If so, this may have neutralized differences between the published and unpublished conditions that would have emerged had we not provided the "just completed" explanation. Finally, jurors may have attended to publication status and simply chose to disregard the information altogether or relied more heavily on other features of the expert testimony to evaluate its quality.

Limitations and Research Implications

Certain limitations must be kept in mind when considering whether our findings generalize to actual trials containing psychological science. Mock jurors in our study read a written summary of a child sexual abuse case, which undoubtedly lacked the look and feel of an actual courtroom trial. The written stimulus also afforded jurors more control over the rate and degree to which they processed the information presented. Actual jurors cannot control how quickly witnesses provide evidence, and unless jurors take notes or ask for the court reporter's transcript, they cannot reread the exact courtroom testimony when rendering decisions. Mock jurors' ability to do so in our experiment may have enhanced juror sensitivity to expert evidence quality compared to more ecologically valid simulations or actual trials. Although possible, this seems unlikely in light of recent research by Pezdek, Avila-Mora, and Sperry (2009). Those researchers varied the presence of eyewitness expert testimony and its presentation modality (written versus videotaped) and observed no significant interaction effects between those variables on jurors' trial-related decisions. The pattern of results essentially was consistent across the written and videotape trial conditions. Even so, we must keep in mind that simulated trial decisions do not carry the real-world consequences of those made in actual cases. Such differences could affect real jurors' motivation and willingness to engage in systematic processing in ways that did not occur in our trial simulation.

With respect to the jury-eligible community members who participated in our study, it is possible that our recruitment methods (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 was more representative of actual jurors and their ability to detect internally invalid research than the college student samples typically used in jury decision-making research. Our study also did not include jury deliberations; mock jurors rendered verdicts independently of one another after reading the case materials. Whether the deliberation process enhances or impedes jurors' individual and collective sensitivity to internally invalid psychological science presented at trial remains an unanswered empirical question. Previous research suggests, however, that jurors' individual verdicts do not vary pre- and post-deliberation (Hastie, Penrod, & Pennington, 1983) and that jurors' discussion of the expert or the expert's testimony during deliberations is quite uncommon (Kovera, Gresham, Borgida, Gray, & Regan, 1997).

One final limitation should be kept in mind when considering the implications of our study. The internal validity manipulation resulted in differences that were statistically significant but relatively small in size with partial

η^2s $\leq$ .08. These small effects raise the issue of practical versus statistical significance, especially when we consider that mock jurors' dichotomous verdicts were unaffected by the internal validity manipulation. Our data cannot speak to whether the statistically significant differences observed in this experiment will result in practical, meaningful differences in actual trials.

Despite these limitations, the present study contributes to the scientific reasoning literature by expanding our understanding of juror decision-making in cases involving psychological science and expert testimony. Our results indicate that although jurors may be capable of identifying a missing control group, they struggle with more complex internal validity threats such as a confound and experimenter bias. As such, the role of traditional legal safeguards against junk science in court such as cross-examination, opposing expert testimony, and judicial instructions become increasingly important. Continued research is needed to examine the effectiveness of these safeguards and to develop new strategies for improving the ability of judges, attorneys, and jurors to identify internally invalid psychological science in court.

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 grants from the National Institute of Mental Health (#T34-20023) and the California Endowment to the second author.

References

Axsom, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of Personality and Social Psychology, 53*, 30–40.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in the need for cognition. *Psychological Bulletin, 119*, 197–253.

California Code of Civil Procedure, §203. Retrieved from www.leginfo.ca.gov on June 9, 2009.

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

Chaiken, S., Liberman, A., & Eagly, A. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–251). New York: Guilford Press.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology, 66*, 460–473.

Cooper, J., Bennett, E. A., & Sukel, H. L. (1996). Complex scientific testimony: How do jurors make decisions? *Law and Human Behavior, 20*, 379–394.

Cooper, J., & Neuhaus, I. M. (2000). The "hired gun" effect: Assessing the effect of pay, frequency of testifying, and credentials on the perception of expert testimony. *Law and Human Behavior, 24*, 149–171.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1989). The eyewitness, the expert psychologist, and the jury. *Law and Human Behavior, 13*, 311–332.

Daubert v. Merrell Dow Pharmaceuticals Inc., 113 S.Ct. 2786 (1993).

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292.

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-*Daubert* world. *Law and Human Behavior, 25*, 433–458.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology, 39*, 578–589.

Hastie, R., Penrod, S. D., & Pennington, N. (1983). *Inside the jury*. Cambridge, MA: Harvard University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist, 44*, 1089–1098.

Kovera, M. B., Gresham, A. W., Borgida, E., Gray, E., & Regan, P. C. (1997). Does expert testimony inform or influence juror decision-making? A social cognitive analysis. *Journal of Applied Psychology, 82*, 178–191.

Kovera, M. B., & McAuliff, B. D. (2000). The effects of peer review and evidence quality on judge evaluations of psychological science: Are judges effective gatekeepers? *Journal of Applied Psychology, 85*, 574–586.

Kovera, M. B., & McAuliff, B. D. (2009). Attorneys' evaluations of psychological science: Does evidence quality matter? (manuscript under review).

Kovera, M. B., McAuliff, B. D., & Hebert, K. S. (1999). Reasoning about scientific evidence: Effects of juror gender and evidence quality on juror decisions in a hostile work environment case. *Journal of Applied Psychology, 84*, 362–375.

Kovera, M. B., Russano, M. B., & McAuliff, B. D. (2002). 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, 8*, 180–200.

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. *American Psychologist, 43*, 431–442.

Levett, L. M., & Kovera, M. B. (2008). The effectiveness of opposing expert witnesses for educating jurors about unreliable expert evidence. *Law and Human Behavior, 32*, 363–374.

Maheswaran, D., & Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology, 61*, 13–25.

Maheswaran, D., Mackie, D. M., & Chaiken, S. (1992). Brand name as a heuristic cue: The effects of task importance and expectancy confirmation on consumer judgments. *Journal of Consumer Psychology, 1*, 317–336.

McAuliff, B. D., & Kovera, M. B. (2008). Juror need for cognition and sensitivity to methodological flaws in expert evidence. *Journal of Applied Social Psychology, 38*, 385–408.

McAuliff, B. D., Kovera, M. B., & Nunez, G. (2009). Can jurors recognize missing control groups, confounds, and experimenter

bias in psychological science? *Law and Human Behavior, 33*, 247–257.

Merlino, M. L., Dillehay, R. C., Dahir, V., & Maxwell, D. (2003). Science education for judges: What, where, and by whom? *Judicature, 86*, 210–213.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46*, 69–81.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–203). New York: Academic Press.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*, 847–855.

Petty, R. E., Kasmer, J. A., Haugtvedt, C. P., & Cacioppo, J. T. (1987). Source and message factors in persuasion: A reply to Stiff's critique of the elaboration likelihood model. *Communication Monographs, 54*, 233–249.

Pezdek, K., Avila-Mora, E., & Sperry, K. (2009). Does trial presentation medium matter in jury simulation research? Evaluating the effectiveness of eyewitness expert testimony. *Applied Cognitive Psychology*. www.interscience.wiley.com. Accessed 11 June 2009. doi:10.1002/acp.1578.

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

Shuman, D. W., Whitaker, E., & Champagne, A. (1994). An empirical examination of the use of expert witnesses in the courts—Part two: A three city study. *Jurimetrics, 34*, 193–208.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

United States of America v. Daniel George LeBlanc, 45 Fed. Appx. 393 (2002).