

The Effects of Peer Review and Evidence Quality on Judge Evaluations of Psychological Science: Are Judges Effective Gatekeepers?

Margaret Bull Kovera and Bradley D. McAuliff
Florida International University

Scientifically trained and untrained judges read descriptions of an expert's research in which the peer review status and internal validity were manipulated. Seventeen percent of the judges said they would admit the expert evidence, irrespective of its internal validity. Publication in a peer-reviewed journal also had no effect on judges' decisions. Training interacted with the internal validity manipulation. Scientifically trained judges rated valid evidence more positively than did untrained judges. Untrained judges rated a study with a confound more positively than did trained judges. Training did not affect judge evaluations of studies with a missing control group or potential experimenter bias. Admissibility decisions were correlated with judges' perceptions of the study's validity, jurors' ability to evaluate scientific evidence, and the effectiveness of cross-examination and opposing experts to highlight flaws in scientific methodology.

Courts have been concerned with the proliferation of "junk science" in the courtroom since a Harvard psychologist tried to introduce evidence from a crude precursor of the polygraph in the murder case against James Frye (*Frye v. United States*, 1923). The "general acceptance" standard adopted in this case governed the admissibility of scientific evidence in the federal (and many state) courts for many years. Under this standard, scientific evidence must be generally accepted within the relevant scientific community to be admitted at trial. However, a recent Supreme Court decision changed admissibility standards in the federal courts. In *Daubert v. Merrell Dow Pharmaceuticals* (1993), the U.S. Supreme Court held that the Federal Rules of Evidence (1975) superseded the Frye decision. General acceptance of scientific findings was no longer an absolute prerequisite for the admissibility of expert testimony. Recognizing that it takes considerable time for even high quality research to become generally accepted in the scientific community, the Court decided that scientific evidence must be both relevant and reliable. Specifically, the Court in *Daubert* ruled that expert evidence under Federal Rule 402—like other forms of evidence—must be relevant to a material fact at issue in the case at hand to be admissible. Relevance is estab-

lished if the expert evidence has probative value (i.e., makes some material fact more or less likely).

The Court also ruled that admissible scientific evidence must be reliable. In *Daubert*, the justices proffered a nonexclusive list of factors to consider in determining the reliability of scientific evidence. They argued that judges should consider whether experts used the scientific method of hypothesis testing to develop their theory or technique. The Court suggested that judges also consider whether the theory or technique had been peer reviewed and published, has a known or potential rate of error, and is generally accepted within the relevant scientific community. Although general acceptance is still one criterion that judges may use to evaluate the admissibility of scientific evidence, the focus of the inquiry under *Daubert* is clearly on the underlying scientific assumptions, methods, data, and interpretations of the expert's testimony. In essence, the *Daubert* decision recommends that judges evaluate the validity of scientific evidence when evaluating its evidentiary reliability.

The *Daubert* decision requires judges to become sophisticated consumers of science (Faigman, 1995). Specifically, *Daubert* clarifies the judges' role as that of gatekeepers who are required to admit only scientific evidence that is reliable (Berger, 1994; Fisher, 1994; Simard & Young, 1994; Walker & Monahan, 1996). Some state courts have followed the federal lead and adopted the *Daubert* standard in their states; other states have specifically rejected the Supreme Court decision and continued to rely on the *Frye* test when judging the admissibility of scientific evidence. Irrespective of whether courts use the *Frye* or *Daubert* standard, it is important to know whether judges are likely to admit flawed scientific studies into evidence. Because there is some evidence that many jurors may have difficulty recognizing variations in the methodological quality of psychological science (Kovera, McAuliff, & Hebert, 1999), it is crucial to determine whether jurors are likely to be confronted with the task of differentiating flawed research from valid research.

Margaret Bull Kovera and Bradley D. McAuliff, Department of Psychology, Florida International University.

Portions of this research were presented at the 106th Annual Convention of the American Psychological Association, San Francisco, August 1998, and the International Congress of Applied Psychology, San Francisco, August 1998. This research was supported by National Science Foundation Grant SBE No. 9711225 and a grant in aid from the American Psychology-Law Society. We are grateful for the research assistance of Elizabeth Freeman, Barbara Gonzalez, Monique Harris, Michelle Kvaska, Jennifer Maloney, and Robin Connell White and for the helpful suggestions of Brian L. Cutler and the Honorable Michael Salmon.

Correspondence concerning this article should be addressed to Margaret Bull Kovera, Department of Psychology, Florida International University, North Miami, Florida 33181. Electronic mail may be sent to koveram@fiu.edu.

Evaluating Scientific Evidence

Although there is a growing body of research on the use of expert testimony in jury decisions, we know little about how judges evaluate psychological science or scientific evidence in general. We do know that judges' socio-political attitudes influence their judgments about the relevance of scientific evidence (Redding & Reppucci, 1999). However, we do not know whether the quality of the scientific evidence influences judges' admissibility decisions. Recent research on methodological reasoning skills suggests that judges may not be able to assess scientific validity accurately without additional training. For example, researchers have demonstrated that people have difficulty applying the law of large numbers to everyday events such as social behavior (Jepson, Krantz, & Nisbett, 1983; Nisbett, Fong, Lehman, & Cheng, 1987). Laypersons may also fail to recognize the limitations of studies that do not include an appropriate control group (Lehman, Lempert, & Nisbett, 1988). Similarly, results from a trial-simulation study demonstrated that jurors are unable to differentiate between psychological research that has good construct validity and research that has poor construct validity (Kovera et al., 1999). Judges probably are no more skilled at these evaluations than are laypeople. Lehman and colleagues (1988) demonstrated that legal education does not improve either methodological or statistical reasoning abilities. Thus, the research on methodological reasoning suggests that judges and jurors may lack the ability to recognize psychological evidence that lacks scientific validity; however, it provides little insight into what types of criteria they might use when evaluating psychological science.

A better understanding of scientific reasoning may be achieved by borrowing some ideas from the social-cognitive literature on persuasion. Current persuasion research suggests that people use one of two cognitive processes when evaluating a persuasive message (Chaiken, Liberman, & Eagly, 1989; Petty & Cacioppo, 1986). Sometimes people systematically examine a message, carefully scrutinizing the quality of the arguments made. Systematic processors are more likely to adopt the position advocated in the message if the message contains high quality arguments (e.g., Leippe & Elkin, 1987; Petty, Cacioppo, & Goldman, 1981). At other times, people engage in heuristic (Chaiken, 1980) or peripheral (Petty et al., 1981) processing. People engaged in heuristic processing use cognitive shortcuts or heuristics to determine the persuasiveness of a message rather than attending to the quality of the evidence presented. If someone believes that experts always provide high quality information, they may judge a communication from a highly qualified expert to be more persuasive than a communication from a less qualified expert. This influence of expert qualifications occurs even though the quality of the arguments proffered by the two sources is equivalent (Cooper, Bennett, & Sukel, 1996; Petty et al., 1981). Similarly, consensus information (i.e., information about others' opinions) influences the judgments of people engaged in heuristic processing (Axson, Yates, & Chaiken, 1987; Giner-Sorolla & Chaiken, 1997; Maheswaran & Chaiken, 1991). Thus, a judge engaged in heuristic processing who believes that consensus implies correctness may admit all evidence that has received a favorable peer review and has been published in a professional journal, even if that evidence is fundamentally flawed.

Persuasion research suggests that audience members must be motivated to scrutinize the message (e.g., Chaiken & Maheswaran, 1994; Petty et al., 1981), but they also must be able to scrutinize the message. Even highly motivated people may be unable to attend to argument quality if they lack the requisite knowledge for understanding the communication (Cooper et al., 1996) or if there are situational constraints on their ability to process (Ratneswar & Chaiken, 1991). Thus only when people have both the motivation and the ability to process information will they engage in systematic processing. Because judges' decisions regarding the admissibility of expert evidence are subject to appeal, we assume that judges are motivated to process scientific evidence systematically. However, recall that the body of research on methodological reasoning suggests that judges may lack the ability to reason systematically about scientific evidence.

In sum, synthesis of the research literatures on methodological reasoning and persuasion suggests that decision makers who lack the reasoning abilities needed to make sense of scientific evidence might rely on peripheral cues or heuristics when reasoning about psychological science. For example, judges may use simple heuristics (e.g., a published study is better than an unpublished study) to evaluate the reliability of experimental psychological research. This approach also suggests that if judges lack the ability to reason systematically about the expert evidence, they will be insensitive to the methodological quality of the evidence (e.g., the importance of an appropriate control group).

Increasing Ability to Process Through Scientific Training

Some judges may be better able to identify flawed scientific research than others are. Research on methodological reasoning suggests that training in the scientific method can improve individuals' reasoning abilities. For example, Lehman et al. (1988) demonstrated that graduate students in disciplines emphasizing methodological reasoning (e.g., psychology) were more likely to apply these reasoning skills to everyday problems than were graduate student in other disciplines (e.g., law). In addition, Fong, Krantz, and Nisbett (1986) found that people given brief training in methodological reasoning provided more scientifically sophisticated answers to a series of real-world problems. If judges lack the ability to reason about scientific evidence, perhaps they can be trained to reason systematically about factors associated with scientific validity when evaluating expert evidence. If training can enable judges to reason in a methodologically sophisticated manner, then they should be able to scrutinize the quality of expert evidence more systematically and thus make better informed decisions.

There are many ways in which judges may be trained to think scientifically. For example, judges who studied the scientific method during college or graduate school may be better able to differentiate high quality research from low quality research. Continuing legal education and judicial conferences also provide judges with the opportunity to receive training in the scientific method. The Federal Judicial Center (1995) made a formal attempt to train federal judges in the scientific method when it published the *Reference Manual on Scientific Evidence* and sent copies of this manual to all federal judges (Walker & Monahan, 1996). Among other topics, this guide provides information on statistical inference, survey research, and multiple regression analysis. Be-

cause this reference manual provides instruction in methodology and abstract rules of statistics, it may provide judges with the skills that they need to be effective gatekeepers. However, state court judges may not have access to this publication or may not know of its existence. Whether these modes of training provide judges with a sufficient understanding of the scientific method and the ability to apply this knowledge to their judicial decision making remains an unanswered empirical question.

Judges' Beliefs About Junk Science in the Courtroom

As noted previously, psychological research demonstrates that judges may be ill equipped to reason about statistical and methodological issues (Lehman et al., 1988). Thus, some judges may admit flawed scientific evidence or may bar the admission of valid research because they are unable to recognize methodological flaws in scientific research. Even if judges can differentiate between valid and less valid expert evidence, they may choose to admit some less valid evidence because of their beliefs that jurors are capable of determining the appropriate weight of the evidence. Some recent research indicates that many jurors may have difficulty assessing the quality of psychological science (Kovera et al., 1999). Therefore, it is important to determine whether judges have an unjustified confidence in jurors' ability to differentiate good psychological science from junk science, because their beliefs may influence their admissibility decisions.

Even if judges do admit flawed psychological science into evidence, safeguards are present in criminal and civil procedure that may help jurors identify flaws in scientific research. According to the *Daubert* decision, doubts about whether the expert testimony will help the jury should be resolved in favor of admissibility. The assumption underlying this decision is that traditional legal safeguards such as cross-examination or opposing experts educate the jury about the reliability of the scientific evidence. In contrast to the reasoning provided by the Supreme Court Justices in *Daubert*, trial simulation research has questioned the efficacy of both cross-examination (Kovera, Levy, Borgida, & Penrod, 1994; Kovera et al., 1999) and opposing experts (Cutler & Penrod, 1995). However, judges may believe that these safeguards assist jurors in assigning appropriate weight to scientific evidence. Because judges' faith in the effectiveness of these safeguards may influence the likelihood that they will admit scientific evidence, it is important to examine judges' beliefs about the efficacy of cross-examination and opposing experts.

Overview

In sum, we designed the present research to answer several related questions. First, do judges rely on information regarding peer review when making decisions about the admissibility of psychological science, especially when they lack scientific training? Second, are scientifically trained judges more sensitive to variations in the internal validity of psychological science than are judges who lack scientific training? Finally, what are judges' beliefs about the abilities of judges, attorneys, and jurors to identify flaws in scientific research? Do judges believe that legal safeguards such as cross-examination and opposing experts can educate jurors about scientific evidence? Are these beliefs related to their admissibility decisions?

To examine these questions, we surveyed state circuit court judges who are in the position of evaluating the admissibility of scientific evidence, including psychological evidence. At the beginning of the survey, we provided a brief description of a hostile work environment case and a more detailed description of the expert testimony that the plaintiff wished to present at trial. The expert testimony consisted of general information on gender stereotyping and an explanation of a relevant study conducted by the expert. Within the context of this expert testimony, we manipulated two variables: the peer-review status and the internal validity of the expert's study. We also measured one individual difference variable: judges' scientific training.

In half the surveys, the study had been subjected to peer review and had been published in a reputable psychology journal. In the remaining surveys, the study had not been peer reviewed. We predicted that in the absence of scientific training, judges would rely on this information about peer review when making their admissibility decisions. Thus, untrained judges would admit the peer-reviewed study more frequently than the study that had not been peer reviewed, irrespective of the methodological quality of the study. In accordance with dual-process models of persuasion, we predicted that this variable would not affect the admissibility decisions of judges who had been trained in the scientific method. This pattern of results would be demonstrated by a significant interaction between peer-review status and scientific training.

We also manipulated the methodological quality of the expert's study by varying its internal validity. In one description of the expert evidence, the study contained appropriate controls and lacked any methodological flaws. Other versions of the expert's study had the potential for experimenter expectancies to bias the results, lacked a control group, or contained a confound. The research on dual-process models of persuasion led us to predict that untrained judges would be unable to differentiate valid research from flawed research. However, we predicted that judges trained in the scientific method would rate the valid study more positively than the flawed studies. This pattern of results would be demonstrated by a significant interaction between internal validity and scientific training.

Finally, we asked our participants to assess the abilities of judges, attorneys, and jurors to identify flaws in scientific research and the effectiveness of cross-examination and opposing experts in educating jurors about scientific evidence. We predicted that judges would report moderate confidence in judges' and attorneys' abilities to evaluate scientific evidence. We also predicted that judges would express less confidence in jurors' abilities than in the abilities of judges and attorneys. In addition, we believed that judges' beliefs about juror abilities would be related to their admissibility decisions; judges should be more likely to admit the psychological science if they believed that jurors were able to identify flaws in the research. We also predicted that judges' beliefs about attorneys' abilities would be related to their admissibility decisions, because attorneys' abilities to recognize flawed psychological research should affect their ability to effectively cross-examine an expert witness. Conversely, we predicted that judges' admissibility decisions would be unrelated to their assessments of the abilities of judges to reason about science. Although we made no predictions about whether judges would view cross-examination or opposing experts as a more effective method for educating jurors about scientific evidence, we did predict that

judges' beliefs about the effectiveness of these safeguards would be related to their admissibility decisions; judges should be more likely to admit expert evidence if they believe there are effective safeguards to protect jurors from being overly influenced by junk science.

Method

Respondents

We sent letters to 400 Florida circuit court judges, soliciting their participation in a study of judges' evaluations of social scientific evidence. A survey and two stamped, addressed envelopes accompanied each letter. We mailed letters and surveys to all criminal, civil, family, and juvenile court judges ($n = 246$). The remaining 154 judges in the sample were randomly selected from a list of all Florida judges who appeared to have general trial court assignments (e.g., they preside over all types of trials). We later learned that 17 of these judges were assigned to drug courts. Because drug court judges do not hear scientific testimony, these 17 judges were removed from the sample. In addition, three surveys were returned because the judges were no longer in office. Thus, our final sample consisted of 380 judges who evaluate scientific testimony.

We chose this particular group of judges for a variety of reasons. Many Florida criminal court judges have experience presiding over civil cases because they rotate between the civil and criminal benches regularly. Therefore, we chose not to limit our population to civil court judges (i.e., the type of judge who would preside over a sexual harassment case). In addition, research suggests that the ability to reason about scientific methodology is not context specific (Nisbett, 1993). Therefore, judges who can reason about scientific evidence in a sexual harassment case should be able to reason about similar methodological issues in a criminal case containing eyewitness testimony. We chose to study Florida judges because we believed that we would have a better response rate from judges residing within our state than from judges within the federal system; state judges might feel some obligation to return a survey from a university in their state. Moreover, Florida's evidentiary rules are patterned after the Federal Rules of Evidence. Therefore, although the Florida Supreme Court has explicitly rejected the *Daubert* decision in favor of *Frye*, judges may still attend to the criteria of relevance and reliability mentioned in *Daubert* and in Florida's evidentiary rules.

We adapted several procedures from the total design method for survey research (Dillman, 1978; Sallant & Dillman, 1994) to increase the response rate. One week after the survey was first mailed, we sent the judges a postcard to remind them to complete the survey. Four weeks after the initial mailing, we sent a second letter, which emphasized the importance of receiving completed surveys from as many judges as possible. A replacement survey was also included in this mailing. Eight weeks after the initial mailing, a well-known judge sent a letter to the judges, explaining the importance of the survey and encouraging them to respond. One week later, we sent a final letter to the judges, accompanied by another copy of the survey.

To increase the response rate further, we offered judges \$20 in return for their participation. A payment form was included with the solicitation letter and the survey. This form asked the judges to provide information that enabled us to compensate them for their participation (e.g., social security number). If the judges wished to receive payment for their participation, they returned the completed payment form in the extra envelope that we provided. We assured judges that it would be impossible to match responses to the individual who provided them. Because of our repeated attempts to solicit participation, we received 144 completed surveys for a response rate of 38%. Only 13 judges (3%) actively refused to participate in the survey.

To assess the representativeness of our sample, we compared the demographics of our sample with the entire population of Florida state judges.

These analyses revealed that our sample did not significantly differ from the population in gender, $\chi^2(1, N = 889) < 1$, *ns*, race, $\chi^2(3, N = 887) = 3.53$, *ns*, age, $F(1, 885) < 1$, *ns*, or years on the bench, $F(1, 884) = 1.67$, *ns*.

Survey

We asked judges to role play that they were presiding over a civil case in which the plaintiff claimed that she had been the victim of sexual harassment in a hostile work environment. The survey contained a description of the fact pattern in a hostile work environment case, a description of expert testimony the plaintiff wished to proffer, and a standard defense motion to bar the admission of the expert testimony. The remainder of the survey assessed judges' evaluations of the expert testimony; judges' beliefs about the abilities of judges, attorneys, and jurors to evaluate scientific evidence and the effectiveness of the cross-examination and opposing expert safeguards; and basic demographic information.

Description of the Fact Pattern and Proposed Expert Evidence

We based the details of this civil case on the fact pattern in *Huffman v. Pepsi-Cola Bottling Co.* (1994). Specifically, the plaintiff claimed that sexually suggestive photos were displayed in the workplace, computer games with sexual content were present on the company computers, and that she frequently had been the target of unwelcome sexual behavior from her male coworkers. She alleged that these experiences created a hostile work environment.

The judges also read a detailed description of expert testimony on gender stereotyping and sexual harassment that the plaintiff wished to present at trial. In this testimony, the expert planned to discuss the conditions under which gender stereotyping is likely to occur (e.g., rarity of women, paucity of information, ambiguity of evaluation criteria, sexualized environment; see American Psychological Association, 1991, for a review). The expert also wished to explain how a sexualized working environment can lead to an increased likelihood of sexual harassment by describing a study she had conducted for the trial. This study, which we based on a study conducted by Rudman and Borgida (1995), demonstrated that viewing sexualized advertisements caused men to rely on gender stereotypes when thinking about, evaluating, and behaving toward women. This particular study was chosen because it has already been admitted at trial in at least two states and because there was independent evidence that an expert sample judged this study to be internally valid. Specifically, the Society for the Psychological Study of Social Issues (Division 9 of the American Psychological Association) awarded this study the Gordon Allport Prize for the best paper published on intergroup relations during the year in which it was published.

It is within the description of Rudman and Borgida's study that we manipulated the peer-review status and the internal validity of the scientific evidence. We manipulated internal validity by varying whether there was the potential for the confederate to bias the results of the study, whether the study had an appropriate control group, or whether there was a confound in the experimental design. We manipulated these aspects of a study's internal validity because they directly correspond to some of the classic internal validity threats discussed by Cook and Campbell (1979).

Peer Review Manipulation

In the peer-reviewed condition, the expert's findings had been well received by the psychological community. Her study had been published in a peer-reviewed journal and had been described in several psychology textbooks. In the not-peer-reviewed condition, the expert had just completed her study, so it had not yet been published in a peer-reviewed journal, nor had it been described in any texts. When this type of information was embedded in a richer, more complex stimulus such as a

videotaped trial (Kovera et al., 1999) or a lengthy written trial summary (McAuliff & Kovera, 1999a, 1999b) in other studies, manipulation checks revealed that participants noticed the manipulation. In addition, jurors' evaluations of expert evidence have been influenced by this type of manipulation (Kovera et al., 1999).

Internal Validity Manipulation

Valid expert testimony. In the internally valid version of the expert's study, 300 employees at a trucking company—who believed they were taking part in a marketing study—evaluated the persuasiveness of 20 television commercials. Half of the men viewed commercials that used scantily clad women and sexual innuendo to sell products. The remaining men viewed commercials for the same products; however, these commercials focused on the attributes of the advertised product (i.e., a control group was present). Subsequently, the men interviewed a woman for the job of an undergraduate research assistant. The same woman acted as the research assistant in both conditions and was trained to act exactly the same way in every interaction (i.e., no confound was present). The research assistant was blind to each participant's condition while she was interacting with him and while rating his behavior (i.e., no experimenter bias).

The research assistant rated the men who viewed sexualized ads to be more sexually motivated than were men who viewed nonsexual ads. The study also demonstrated that men who watched the sexualized commercials were more likely to ask the interviewee sexually inappropriate questions, to sit closer to her during the interview, and to evaluate her capabilities more negatively than were men who watched nonsexual commercials. The expert concluded that exposure to sexually suggestive material increases the likelihood that men will sexually harass female coworkers.

Experimenter-bias condition. In the experimenter-bias condition, the expert described the basic experiment with one alteration. In this version of the study, the confederate knew which advertisements the participants had viewed in the previous phase of the study. That is, the confederate was not blind to the participants' conditions; therefore, she unknowingly may have behaved differently toward the men in different conditions, eliciting the differences in the men's behavior. In addition, her knowledge of the experimental hypotheses may have biased her ratings of the men's behavior.

No-control-group condition. In the no-control-group condition, the expert again described the basic study with one alteration: It lacked the appropriate control group (i.e., advertisements without scantily clad women). In this version of the study, all of the participants viewed commercials that used scantily clad women and sexual innuendo to sell their products. The expert explained that the research assistant judged the men to be sexually motivated, the men were likely to ask the interviewee sexually inappropriate questions, the men sat close to her during the interview, and the men evaluated the assistant's capabilities negatively.

Confound condition. In the confound condition, the expert described the basic experiment with a different alteration. In this version of the study, two confederates played the role of the research assistant: One confederate played the research assistant when the men were shown the sexually suggestive advertisements, and the other confederate played the research assistant when the men were shown the nonsexual advertisements. Neither confederate was aware of the study hypotheses nor were they aware which commercials the men with whom they interacted had seen. Although experimenter bias was not possible in this condition, it is possible that the research assistant who was paired with the sexually suggestive advertisements was more attractive than the other confederate or in some other way elicited more sexually inappropriate behavior from the men with whom she interacted than did the other confederate.

Because we believed keeping the survey short would increase the number of judges who responded, we did not include survey items that would serve as manipulation checks. We have conducted other studies that have used the same description of the expert evidence and the same manipulations of internal validity embedded in a 15-page written trial

summary (McAuliff & Kovera, 1999a, 1999b). Manipulation checks in these studies suggest that participants do notice these methodological differences. Given these findings, we believe that the same manipulations in this more impoverished context were strong enough to produce effects on judges' decisions if this type of methodological information is something they consider when making admissibility decisions.

Individual Differences in Scientific Training

We asked the judges several questions to assess their level of education or training on scientific issues. They provided information about their undergraduate major field of study, whether they completed a science course during their graduate training, and whether they had received any continuing legal education on scientific methods. They also indicated whether they had seen the Federal Judicial Center's *Reference Manual on Scientific Evidence* and rated their familiarity with its contents. We divided the judges into two groups. We categorized those who reported receiving an undergraduate major in the natural sciences or psychology, graduate training in science, or continuing legal education in the scientific method as scientifically trained judges. Many of the trained judges had training from multiple sources. We categorized judges who lacked this kind of scientific background as scientifically untrained. Because so few judges were familiar with the *Reference Manual*, and because all of those who were had some scientific training, we did not use it as a criterion for categorizing judge education. This information allowed us to examine whether scientific training increases judges' abilities to make sophisticated judgments about scientific validity.

Survey Items

The judges read the case materials and then provided their opinions regarding the admissibility of the proposed expert evidence. The judges rendered a dichotomous judgment regarding the admissibility of the evidence; that is, they indicated whether they would admit the proposed expert testimony in the hypothetical hostile work environment case. We also asked them to provide justification for their admissibility decisions. We were interested in examining the frequency with which they provided legal or scientific validity justifications for their decisions. Therefore, two independent raters coded the judges' free responses to determine whether the judges gave particular legal justifications (i.e., general acceptance, relevance, prejudicial, expert qualifications, common understanding of the jury, helpfulness), scientific validity justifications (i.e., general statements about validity, external and ecological validity, construct validity, internal validity), or other types of justification.

The formula used for calculating concordance estimates was $C = 2(C_{1,2}) / (C_1 + C_2)$, where C = concordance for judge justification, $C_{1,2}$ = number of identical categories assigned by both coders, and C_1 and C_2 = total number of categories assigned by the first and second coders, respectively. Strict agreement criteria were used in computing concordance rates. That is, if one rater coded a phrase using a superordinate category label (e.g., legal justification) and the other rater coded the same phrase using a more specific subordinate category label (e.g., common knowledge of the jury), codes were classified as identical. However, if the two raters chose different specific subordinate category labels for the same phrase (e.g., ecological validity of laboratory setting and ecological validity of experimental procedure for the first and second coders, respectively), these codes were classified as nonidentical, despite the fact that both codes could be subsumed under the same superordinate category (e.g., ecological validity). Codes were also considered nonidentical if the first rater coded a phrase and the second rater did not. Overall, the concordance rate for the two coders was .84; disagreements were resolved through discussion.

Judges also rated the extent to which the proposed expert testimony was relevant, probative, prejudicial, reliable, and scientifically valid on 7-point Likert-type scales. The data were recorded so that higher numbers always

indicated more positive evaluations (e.g., 1 = *not at all relevant*, 7 = *very relevant*). The items tapping judges' opinions about the legal admissibility of the expert evidence (i.e., relevant, probative, prejudicial) were averaged to create a legal admissibility scale (Cronbach's $\alpha = .73$), with higher numbers indicating that the testimony was more likely to be admissible. This multiple-level measure of admissibility provides a more sensitive measure of the effects of variability in scientific validity on judges' decisions to admit expert evidence than does the dichotomous admissibility decision. Two items assessing the study's evidentiary reliability and scientific validity were averaged to create an index of the study's validity (Cronbach's $\alpha = .91$), with higher numbers indicating that the testimony was more valid.

We also asked judges a series of questions to assess their beliefs about the ability of legal decision makers and attorneys to identify flaws in scientific research. First, judges rated the extent to which they agreed with the proposition that jurors are able to identify flaws in scientific research. They also rated the extent to which they agreed with similar statements about judges and attorneys. Second, judges rated the extent to which they agreed with statements that cross-examination and opposing experts are effective methods of helping jurors to identify flaws in expert evidence. They made all of these ratings on 7-point Likert-type scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

In addition to these ratings, the judges provided demographic information such as their age, ethnicity, and number of years on the bench. They also described the types of expert testimony that had been proffered in their courts during the past year. These descriptions were coded by two independent raters into the following categories of expert testimony: accident and traffic reconstruction, ballistics, chemical and drug testing, economic, identification (e.g., DNA, fingerprint, serology), medical, psychological (e.g., child abuse, competency, custody, insanity), or other. Concordance between the two raters was high (.92); disagreements were resolved through discussion.

Results

Respondent Demographics

Demographic information for our respondents is presented in Table 1. Most of our sample was male and Caucasian. Seven percent of the sample was Hispanic, and 5% was Black (e.g., Haitian, Jamaican, or African American). Less than 1% of the sample was either Asian, Pacific Islander, or Native American. The respondents ranged in age from 37 to 68 years old ($M = 51.58$, $SD = 7.04$) and had served anywhere between 1 and 27 years on the bench ($M = 9.61$, $SD = 6.68$). Few judges had received undergraduate or graduate training in the scientific method. However, a majority of the judges reported that they had some form of scientific training through continuing legal education or judicial conferences. Very few of the judges reported being familiar with the Federal Judicial Center's reference manual, which suggests that the important information in this volume has not yet trickled down from the federal to the state courts.

Randomization checks indicated that random assignment of the judges to condition was successful (i.e., demographic variables did not vary as a function of experimental condition). We present cell sizes for the three-way interaction of peer review, internal validity, and scientific training in Table 2. Although some of the cell sizes for the three-way interaction are small, the cell sizes for the predicted two-way interactions are quite respectable, ranging from 23 to 45 for the Peer Review \times Scientific Training interaction and from 13 to 24 for the Internal Validity \times Scientific Training interaction.

Table 1
Demographic Information for Judge Sample

Demographic variable	% of sample
Gender	
Male	19
Female	79
Race/ethnicity	
Caucasian	83
Hispanic	7
Black/African American	5
Asian/Pacific Islander	1
Native American	1
Age ($M = 51.58$, $SD = 7.04$)	
40 and under	5
41–45	13
46–50	32
51–55	25
56–60	9
61–65	9
Over 65	4
Years on the bench ($M = 9.61$, $SD = 6.68$)	
1–5	30
6–10	30
11–15	21
16–20	8
Over 20	8
Undergraduate major	
Science	18
Nonscience	82
Graduate school	
Science	13
Nonscience	86
Continuing legal education	
Yes	53
No	47
Familiar with FJC reference manual ^a	
Yes	6
No	92

Note. $N = 144$. Percentages may not sum to 100 because of missing data. FJC = Federal Judicial Center.

^a Federal Judicial Center (1995). Familiarity was rated on a scale ranging from 1–7, for which 1–3 indicated lower familiarity and 5–7 indicated higher familiarity.

The types of expert testimony heard by judges appear in Table 3. Three quarters of the judges reported that medical evidence had been presented in their courtroom during the past year. Seventy-five percent of the judges also reported hearing psychological testimony. It is interesting to note that none of the judges reported hearing experimental psychological testimony. Instead, they most frequently heard psychologists testify about forensic issues such as competency, custody, and insanity. Over a third of the judges reported that identification evidence had been proffered in their court. The judges also reported hearing testimony about drugs, accident reconstruction, economics, and ballistics with some frequency.

Evaluations of Scientific Evidence

We subjected the judges' admissibility decisions to a 2 (peer review: not peer reviewed vs. peer reviewed) \times 4 (internal validity: valid, experimenter bias, no control group, confound) \times 2 (scientific training: untrained vs. trained) logistic regression. On

Table 2
Cell Sizes for Each of the Experimental Conditions

Condition	Valid	Experimenter bias	Missing control	Confound
Not peer reviewed				
Untrained	5	8	6	4
Trained	12	11	9	12
Peer reviewed				
Untrained	8	8	7	9
Trained	12	10	13	10

average, only 17% of the judges believed that this testimony should be admitted in court. Thus, most judges believed that this expert testimony should not be admitted in court despite the fact that gender stereotyping evidence has been admitted in a highly publicized hostile work environment case tried in the Florida courts (*Robinson v. Jacksonville Shipyards, Inc.*, 1991). Even the internally valid study, which is likely to meet both *Daubert* and *Frye* admissibility criteria, was rarely admitted by judges (17%). Moreover, the judges were no more likely to admit the internally valid study than to admit the confounded study (17%), the study missing a control group (11%), or the study with a confederate who was not blind to condition (24%). The peer-review and internal validity manipulations did not significantly influence the judges' decisions to admit the expert evidence, nor did scientific training. None of the interactions were significant.

It is possible that we were unable to detect the effects of our manipulations on the judges' admissibility decisions because of the dichotomous nature of the measure. Thus, we also subjected the judges' scores on the multiple-level legal admissibility scale and their ratings of the study's validity to a 2 (peer review: not peer reviewed vs. peer reviewed) \times 4 (internal validity: valid, experimenter bias, no control group, confound) \times 2 (scientific training: untrained vs. trained) multivariate analysis of variance. The multivariate test of the Peer Review \times Scientific Training interaction was not significant. However, the multivariate test of the Internal Validity \times Scientific Training interaction was significant, $F(6,$

Table 3
Percentage of Judges Who Had Encountered Different Types of Expert Testimony in the Past Year

Expert testimony type	% judges
Medical evidence	76
Psychological testimony	75
Competency	47
Custody	13
Insanity	11
Child abuse	7
Addiction	5
Identification evidence	40
DNA evidence	27
Fingerprint evidence	23
Serology/blood typing	11
Chemical/drug identification	29
Accident reconstruction	24
Economic (property valuation, etc.)	22
Ballistics	15
Other	29

Table 4
Summary of Univariate ANOVA Results for Judges' Legal Admissibility Ratings

Source	F	df	p	η^2
Peer review (PR)	0.27	1,127	<i>ns</i>	.00
Internal validity (IV)	2.23	3,127	<i>ns</i>	.05
Scientific training (ST)	0.04	1,127	<i>ns</i>	.00
PR \times IV	1.01	3,127	<i>ns</i>	.02
PR \times ST	1.36	1,127	<i>ns</i>	.01
IV \times ST	4.00	3,127	.01	.09
PR \times IV \times ST	0.88	3,127	<i>ns</i>	.02

Note. ANOVA = analysis of variance.

250) = 2.35, $p = .03$. Table 4 contains the results of the follow-up univariate analysis for the judges' ratings of legal admissibility. This analysis revealed a significant Internal Validity \times Training interaction for the judges' ratings of legal admissibility (see Figure 1 for means). Simple effect tests revealed that judges educated in the scientific method gave higher legal admissibility ratings to the internally valid study than did judges who had no scientific training. In contrast, untrained judges gave higher legal admissibility ratings to the study containing a confound than did scientifically trained judges. Level of training did not affect the judges' admissibility ratings for the study with a missing control group or a nonblind research assistant. None of the other multivariate main effects or interactions were significant. Moreover, the univariate analysis of the judges' ratings of study validity revealed no significant main effects or interactions, all F s < 1, all η^2 s < .02.

In some cases, the judges' justifications for their decisions differed depending on whether they chose to admit the expert evidence. Table 5 presents the proportion of judges who provided each type of justification as a function of their admissibility decision. If judges chose to admit the testimony, they were more likely to argue that the evidence would be helpful to a jury and the

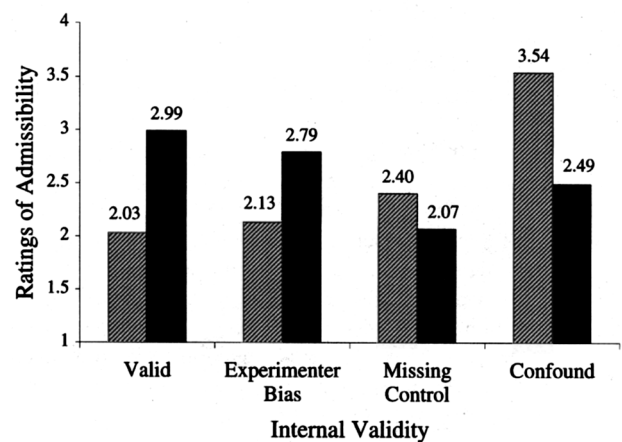


Figure 1. Judges' ratings of admissibility as a function of internal validity and scientific training. Black bars represent judges who received scientific training; hatched bars represent judges who received no scientific training.

Table 5
Proportion of Judges Providing Legal and Scientific Validity Justifications for Their Admissibility Decisions

Justification	Admissibility decision		χ^2	<i>p</i>	<i>r</i>
	Not admit	Admit			
Legal	79.1	86.4	0.61	<i>ns</i>	.07
General acceptance	43.6	9.1	9.30	.01	.27
Relevance	34.5	41.1	0.32	<i>ns</i>	.05
Helpfulness	10.9	31.8	6.50	.01	.22
Common understanding	10.0	18.1	1.22	<i>ns</i>	.10
Probative vs. prejudicial	8.1	4.5	0.35	<i>ns</i>	.05
Qualifications	5.5	22.7	7.16	.01	.23
Scientific validity	66.4	31.8	9.16	.01	.26
General statement re: validity	39.1	18.1	3.50	.10	.16
External/ecological validity	17.3	13.6	0.17	<i>ns</i>	.04
Construct validity	11.8	0.0	2.88	.10	.15
Internal validity	10.9	0.0	2.64	.10	.14

Note. For all chi-squares, degrees of freedom = 1 and *N* = 132.

expert was highly qualified than were judges who did not admit the testimony. If judges decided not to admit the testimony, they were more likely to mention the *Frye* rule (e.g., “[the study] does not meet the *Frye* test”) than were judges who admitted the testimony. Judges were slightly more likely to mention construct validity (e.g., “research assistant who renders a subjective analysis of behavior”) and internal validity (e.g., “two different research assistants . . . which should have remained constant”) or to make a general statement about the study’s validity (e.g., “not a valid study”) if they decided not to admit the testimony than if they decided to admit it. The proportion of judges who mentioned the testimony’s relevance, jurors’ common understanding, the relationship between the testimony’s prejudicial and probative value, and external and ecological validity did not differ as a function of their admissibility decision.

Very few judges mentioned the internal validity threats present in the scientific research they evaluated. Thirteen percent of the judges reading the version of the study in which the research assistant was confounded with experimental condition mentioned the confound as a justification for their admissibility decision. Nine percent of the judges who read a description of the study in which the research assistant was aware of the participants’ experimental condition before interacting with them mentioned the knowledge of the confederate as a potential biasing influence. Finally, only 8% of the judges reading the description of the study in which all the participants watched sexualized commercials noted the missing control group.

In our final analysis of the judges’ justifications for their admissibility decisions, we created a new independent variable, study quality, by collapsing across the three conditions in which judges read descriptions of flawed studies. We also created an overall index of whether the judges made any mention of internal validity threats (i.e., missing control group, confound, or experimenter bias) to the expert’s research. We subjected this index and the frequency of the judges’ other justifications to 2 (peer review: not peer reviewed vs. peer reviewed) \times 2 (study quality: valid vs. invalid) \times 2 (scientific training: untrained vs. trained) analyses of

variance (ANOVAs). Only the analysis of the internal validity index revealed a marginally significant main effect for study quality, $F(1, 126) = 3.50, p = .06, \eta^2 = .03$. Judges were more likely to mention flaws in the expert testimony when such flaws actually existed (12% of the judges who read descriptions of flawed studies noticed the flaw). Only 1 judge (3% of the judges reading the valid study) mistakenly complained that the study did not have the appropriate controls when in fact it had an appropriate control group. No other judges claimed that the study had a particular internal validity threat when it did not. Moreover, the study’s peer review status, its quality, and the judges’ training did not influence the frequency with which judges provided other types of justifications for their admissibility.

Judges’ Ratings of Decision-Maker Ability and Safeguard Effectiveness

We asked the judges how well judges, attorneys, and jurors could identify flaws in scientific research. We subjected their ratings to a 3 (decision maker: judge, attorney, juror) \times 2 (scientific training: untrained vs. trained) repeated measures ANOVA, with decision maker as the within-subject factor. They reported that judges and attorneys were better able to spot flaws in scientific studies than were jurors, $F(2, 280) = 75.06, p < .001, \eta^2 = .35$ (*M*s = 4.52, 4.38, and 3.13 for judges, attorneys, and jurors, respectively). Level of training did not affect the judges’ ratings. They did not express overwhelming confidence in anyone’s ability to judge scientific evidence; their ratings of judge and attorney competence were just above the midpoint of the scale, and their ratings of juror competence were below the scale midpoint.

We also asked the judges whether cross-examination and opposing experts were effective methods for alerting jurors to flaws in an expert’s evidence. We subjected these ratings to a 2 (safeguard: cross-examination vs. opposing expert) \times 2 (scientific training: untrained vs. trained) repeated measures ANOVA, with safeguard as the within-subject factor. The judges reported that cross-examination was more effective than were opposing experts in educating jurors about the flaws in scientific testimony, $F(1, 137) = 15.42, p < .001, \eta^2 = .10$ (*M*s = 5.15 and 4.51 for cross-examination and opposing experts, respectively).

Finally, we examined the correlations among the judges’ beliefs about decision-maker abilities, safeguard effectiveness, and their admissibility decisions. Means, standard deviations, and correlations for all dependent measures are presented in Table 6. Whether measured using a dichotomous measure or a multilevel rating, the judges’ admissibility decisions were positively correlated with their ratings of the study’s validity, jurors’ abilities to identify flaws in research, and the effectiveness of the cross-examination safeguard. The multilevel admissibility ratings also were positively correlated with the judges’ ratings of the effectiveness of the opposing expert safeguard. The judges’ ratings of judges’ abilities to understand scientific evidence, not surprisingly, were strongly correlated with their beliefs about attorneys’ abilities in this area. They also were significantly correlated with their perceptions of jurors’ abilities and with the effectiveness of the cross-examination safeguard. Their beliefs about attorneys’ abilities were positively correlated with their ratings of jurors’ abilities and the effectiveness of both safeguards. They also believed that jurors’ abilities are strongly related to the effectiveness of both the cross-examination

Table 6
Means, Standard Deviations, and Correlations Among the Dependent Measures

	1.	2.	3.	4.	5.	6.	7.	8.
<i>M</i>	0.17	2.56	2.20	4.52	4.38	3.13	5.11	4.52
<i>SD</i>	0.38	1.31	1.24	1.63	1.64	1.60	1.70	2.03
1. Admit?	—	.71	.71	.09	.06	.30	.23	.13
2. AR		—	.67	.12	.13	.33	.25	.18
3. SV			—	.09	.09	.30	.14	.05
4. Judge				—	.85	.47	.25	.14
5. Attorney					—	.51	.33	.17
6. Juror						—	.44	.24
7. C-E							—	.49
8. OPSE								—

Note. Significant correlations are indicated by boldface type. Admit? = dichotomous admissibility decision; AR = multilevel admissibility ratings; SV = study validity ratings; Judge = ratings of judges' abilities to evaluate scientific evidence; Attorney = ratings of attorneys' abilities to evaluate scientific evidence; Juror = ratings of jurors' abilities to evaluate scientific evidence; C-E = ratings of the effectiveness of the cross-examination safeguard; OPSE = ratings of the effectiveness of the opposing expert safeguard.

and opposing expert safeguards. Finally, judges' beliefs about the effectiveness of cross-examination and opposing experts were positively correlated.

Discussion

Judges' Evaluations of Scientific Evidence

It is noteworthy that only 17% of the Florida circuit court judges we surveyed would admit the psychological evidence we described in a hypothetical sexual harassment case. Even more important is the finding that the methodological quality of the psychological research presented to the judges did not influence their admissibility decisions or their evaluations of the study's validity. Although many scholars have expressed concern about the admission of flawed science into evidence (e.g., Black, Ayala, & Saffran-Brinks, 1994; Walker & Monahan, 1996), our research suggests that high quality psychological science may be frequently excluded from evidence. Depriving jurors of the information provided by this valid research may decrease the quality of the decisions that they render. These findings also suggest that judges may sometimes admit flawed psychological research into evidence. Thus, it is important to determine whether jurors can discriminate between flawed and unflawed research with or without the help of cross-examination or opposing expert testimony.

Why were the judges so reluctant to admit the psychological research described in our survey? Judges' evaluations of the scientific validity of the expert's research were positively correlated with their admissibility decisions. Specifically, they were more likely to admit the expert's testimony if they believed that her research was competently conducted. Although it is clear that our internal validity manipulation did not influence the judges' perceptions of study validity, we do not know what factors did influence their evaluations. Perhaps the nature of the experiment described by the expert (i.e., interviewing an undergraduate research assistant) caused the judges to dismiss the expert testimony out of hand because they perceived the task to be dissimilar to the types of interactions prevalent in the plaintiff's work environment. The justifications provided by the judges for their admissibility

decisions suggest that this explanation for the low admission rates is not likely. Recall that we coded the judges' justifications for mentions of external validity issues. We also coded their justifications specifically for mentions of the artificiality of the task or the laboratory setting. Only 16% of the judges objected to the study on any grounds relating to external or ecological validity. Even fewer (less than 3% of the sample) objected to the task or the laboratory setting when justifying their decision to bar the admission of the expert's testimony. In their open-ended comments, many judges remarked about their distrust of experts and of psychological testimony. It is possible that this distrust was the source of the judges' unwillingness to admit the testimony; however, more research is needed to confirm this possibility.

Contrary to our predictions, the peer-review status of the expert's study did not interact with judges' scientific training to influence admissibility decisions or other evaluations of the expert psychological evidence. Thus, it appears that judges may not use peer review as a heuristic for judging the quality of psychological science; this is true whether or not the judges making the decisions have received scientific training. Perhaps publication in a peer-reviewed journal is not sufficient to persuade judges that the experiment has been accepted by the scientific community. Rather, judges may require additional testimony from another psychological expert who attests to the acceptance of this research by the scientific community before they are willing to admit psychological research into evidence.

We received partial support for our hypothesis that internal validity would interact with scientific training to influence judges' admissibility decisions. This interaction was not obtained for the dichotomous admissibility decision or judges' evaluations of study quality. However, the results from the legal admissibility measure suggest that a judge's background in science interacted with the study's internal validity to influence his or her willingness to admit psychological science into evidence. Specifically, judges with training in the scientific method gave higher admissibility ratings to the valid study than did untrained judges. Trained judges gave lower admissibility ratings to the study with a confound. Training did not help sensitize judges to the problems associated with

missing control groups or nonblind experimenters, nor did it influence whether they used the internal validity flaws as justification for their decisions.

On the final page of our survey, we asked the judges to provide any further comments that they wished to make. One judge wrote "Judges and attorneys do not have a basic understanding of the scientific method and scientific evidence." The findings from our research suggest that this statement may be true for many judges, even those who have received some scientific training. As we noted previously, the judges were no more likely to admit the valid, award-winning study than they were to admit a flawed study. Moreover, analysis of the judges' justifications for their admissibility decisions demonstrated that very few judges reported detecting the flaws present in the research proffered by the expert. Indeed, only 12% of the judges who read flawed research noted the flaw as a justification for their admissibility decision. Although one could hypothesize that judges are more likely to provide legal justifications for their decisions than they are to provide scientific rationales, a majority of our judges cited validity issues to support their admissibility decision. However, when they mentioned the study's validity, they did so by making vague statements (e.g., "the study is not valid") rather than specifically mentioning the internal validity threats present in the study. They voiced these vague concerns about the study's validity irrespective of the study's methodological quality or the level of their scientific training.

These findings suggest that the scientific training that these judges have received may not be sufficient to produce critical consumers of psychological science. It is possible that a background in biology or chemistry may sensitize judges to certain scientific flaws but not to the types of internal validity threats that can plague psychological research, such as missing control groups and nonblind experimenters. Perhaps training in other disciplines (e.g., psychology, sociology) would better enable judges to recognize these methodological flaws (e.g., Lehman & Nisbett, 1990). Unfortunately, very few of the judges in our sample had received this kind of training; therefore we were unable to test this hypothesis. Moreover, the dichotomous nature of our assessment of scientific training may have masked significant differences that would have been revealed if we had used a continuous measure of scientific training. Because it was necessary to curtail the length of the survey to encourage judges to respond, we were constrained in the number of questions we could ask about scientific training. Our measure of scientific training did have some strengths in that it tapped both intensive training in the scientific method that occurred years ago (i.e., undergraduate and graduate majors in the sciences) and briefer training that occurred more recently (i.e., continuing legal education). Further research is needed to determine whether other operationalizations of scientific training produce significant differences in judicial admissibility decisions.

Even if continued research suggests that the scientific training judges receive is insufficient to help them recognize flawed psychological research, it is possible that even untrained judges can be made sensitive to internal validity threats through the course of pretrial arguments about the admissibility of expert evidence. When opposing counsels object to testimony that attorneys propose to enter into evidence, they file a motion in limine that questions the admissibility of the expert testimony. This pretrial motion and the accompanying oral arguments provide a vehicle for attorneys to educate judges about any flaws that might be present

in the scientific research. To be effective in sensitizing judges to internal validity threats, attorneys need to be able to identify the threats themselves or hire an expert to help them evaluate the scientific validity of the proposed testimony.

We are currently conducting research to examine whether attorneys are sensitive to flaws in scientific research and under what conditions they are willing to consult their own expert for advice. In the meantime, the current research provides some, but very little, insight into whether this safeguard might be effective. Judges opined that attorneys would be no more able to identify flawed methodology than would judges. If they are correct, then attorneys may also benefit from additional scientific training. After all, judges were once attorneys. If they have difficulty identifying internal validity threats, it is possible that attorneys have similar difficulties.

Judges' Justifications for Their Admissibility Decisions

Perhaps not surprisingly, the judges were most likely to provide legal justifications for their admissibility decisions. Those judges who chose not to admit the testimony were more likely to cite the *Frye* rule or a general acceptance criterion as the reason for their decision than were judges who chose to admit the testimony. Judges who admitted the expert testimony were more likely to argue that the expert was highly qualified and that the expert's testimony would be helpful to the jury than were judges who did not admit the testimony. It is also interesting to note the proportion of judges who cited criteria found in Florida's evidentiary rules (e.g., helpfulness, relevance, qualifications, common understanding, and probative vs. prejudicial nature), given that the Florida Supreme Court has specifically ruled that the *Frye* general acceptance test and not the *Daubert* test is to be used in Florida courts (*Jordan v. State*, 1997). Apparently, a significant proportion of our judges were unaware of this decision and continue to use criteria other than general acceptance when judging the admissibility of scientific evidence.

It is interesting to note that a significant proportion of the judges mentioned concerns about the scientific validity of the expert's research when justifying their admissibility decisions. However, judges were most likely to express general concerns about the study's validity and expressed these concerns irrespective of whether the study was flawed or valid. If the judges did express specific concerns about the study's methodology, they were most likely to mention concerns about external or ecological validity. Perhaps it is not surprising that judges often objected to the laboratory setting of the expert's research, stating that it did not resemble the plaintiff's workplace. Judges were less likely to mention concerns with the study's construct validity (e.g., how the expert measured sexual harassment) and internal validity. Indeed, of all the judges who read descriptions of a fundamentally flawed experiment, only 12% mentioned any of the methodological flaws present in the research. These findings provide further evidence that judges may be ill-equipped to evaluate the quality of scientific methods.

Beliefs About Decision-Maker Ability and Legal Safeguards

As we had predicted, the judges reported that judges and attorneys were better able to identify invalid scientific research than

were jurors. Shuman, Whitaker, and Champagne (1994) also reported that judges believe they are more capable of scrutinizing the quality of expert testimony than are jurors. Is it true that judges and attorneys are better able to evaluate scientific evidence than jurors? Researchers have investigated whether a law school education improves methodological and statistical reasoning skills (Lehman et al., 1988). This research suggests that completion of law school does not train people to reason about statistical and methodological issues. Moreover, the present research suggests that judges may have difficulty in differentiating high quality research from junk science. Thus, they seem slightly overconfident in their competence in evaluating scientific evidence.

However, the judges' distrust of juror capabilities may be appropriate. Basic psychological research has shown that many laypeople have weak methodological and statistical reasoning skills (Lehman & Nisbett, 1990). Although no one has examined whether jurors are able to detect internal validity defects in expert evidence, they have previously demonstrated that jurors are insensitive to variations in construct validity (Kovera et al., 1999). We currently are conducting research to determine whether this insensitivity extends to issues of internal validity.

Although the judges generally did not exhibit much faith in jurors' abilities to reason about scientific evidence, we received support for our hypothesis that those judges who had greater confidence in jurors' abilities were also more likely to admit the expert evidence. In contrast, judges' ratings of the abilities of judges and attorneys to identify flawed research were not related to their admissibility decisions. This pattern of results suggests that judges recognize that once evidence is admitted at trial, jurors are responsible for judging the reliability of the evidence and assigning it appropriate weight in their decision. However, this pattern also suggests that judges may not be mindful that attorneys who are able to identify flawed research are better equipped to help jurors identify flaws in research by drawing attention to them during cross-examination or closing arguments. If judges were mindful of this, then they should be more likely to admit the expert evidence if they believe that attorneys have the ability to identify flawed research.

Judges also expressed the opinion that cross-examination was a more effective mode for educating jurors about scientific validity than were opposing experts. However, they expressed the belief that both cross-examination and opposing experts were relatively effective methods for sensitizing jurors to scientific validity. As predicted, they were more likely to admit the expert evidence if they believed that these safeguards are effective methods of exposing flawed research to jurors. Although researchers have not investigated whether cross-examination or opposing experts are better able to sensitize jurors to internal validity flaws, trial simulation studies have examined the effect of cross-examination on juror evaluations of expert evidence. In one study, Kovera and her colleagues (1994) demonstrated that the strength of the cross-examination of an expert did not influence jurors' trial judgments or their evaluations of the expert and her testimony. In another study, they showed that a cross-examination that highlighted methodological features of the expert's research did not sensitize jurors to the quality of the scientific evidence (Kovera et al., 1999). Thus, it appears that judges may have misplaced faith in the ability of cross-examination to educate jurors about scientific validity. It is possible, however, that a different type of cross-examination (e.g.,

a cross-examination that provides a more in-depth explanation of methodology) may be more effective.

Compared to the research on cross-examining experts, less is known about the effects of opposing experts on juror decision making. One study suggests that opposing experts may not sensitize jurors to important aspects of the evidence they are evaluating. In a trial simulation, Cutler and Penrod (1995) varied the conditions under which a witness viewed a crime. They also varied whether expert testimony was presented at trial. Jurors either heard a defense expert discuss the factors that influence the accuracy of eyewitness identifications, the defense expert and a prosecution expert who discussed the limitations of the research presented by the defense expert, or no expert. Cutler and Penrod found that a single expert sensitized jurors to the factors associated with the expert evidence. The addition of the opposing expert caused jurors to be skeptical of all eyewitness identifications, irrespective of the quality of the witnessing conditions. Although this study did not examine the effects of opposing experts on juror evaluation of the expert testimony itself, it suggests that the opposing expert safeguard may be less effective than judges think. More research is needed to assess whether these results generalize to juror evaluations of expert evidence.

Conclusions

Many of the judges in our sample expressed their negative views about psychology and the experts who testify about the psychological research. One particularly eloquent judge wrote, "Too often psychologists seem to lay claim to some mysterious or alchemistic powers, viewing their fragile assessment of such matters a virtual certainty. Even with a much broader and sophisticated database, theirs is at best, in most cases, no more than educated conjecture." This judge read the valid, award-winning version of the expert evidence. The judge's negative opinions about psychology seemed to be shared by others. Recall that the judges infrequently admitted even the valid expert evidence (i.e., only 17% of the judges admitted the valid evidence). Thus, although judges may have admitted the flawed research relatively infrequently (i.e., 17% of the judges admitted invalid evidence), their reluctance to admit the evidence may be because of their general bias against psychology rather than their ability to identify flawed evidence. Further research is needed to determine if flawed research from other disciplines is admitted more frequently.

Another judge who refused to admit the valid version of the expert evidence recommended that we read Margaret Hagen's (1997) scathing critique of psychological expert testimony, in which she argues for the exclusion of expert testimony based on clinical psychology. In her book, Hagen called for expert testimony based on clinical opinion to be barred from evidence; however, she argued that expert testimony based on high quality experimental psychological research could help jurors make better decisions. It seems to us that the judge who cited her book missed the nuances of Hagen's argument (which is understandable, given the vitriolic nature of her attack on psychological evidence). It is unclear how many of our judges also misinterpreted Hagen's argument. However, it is clear that most judges in our study voted to keep out any type of psychological evidence, even the types of award-winning evidence based on sound experimental procedures of which Hagen would approve.

Our findings suggest that a significant proportion of judges may admit into evidence findings from flawed psychological experiments, not just the flawed clinical evidence about which Hagen is concerned. It is possible that they admit flawed research from other scientific disciplines as well. Thus, it is likely that some invalid scientific research will confront jurors in some trials. Given that jurors may also have difficulty differentiating valid research from junk science (Kovera et al., 1999), it is important to identify methods of training judges and jurors to evaluate scientific evidence as if they were scientists. Pretrial hearings may be a vehicle for educating judges about scientific evidence. It is also possible that an extended cross-examination, which teaches jurors to reason about methodology and statistics, will assist jurors in their task of weighing the scientific evidence. However, both of these mechanisms rely on attorneys to educate the trier of fact. Therefore, it is important for researchers to evaluate the ability of attorneys to formulate oral arguments or cross-examination questions that will enhance the ability of judges and jurors to evaluate scientific evidence.

References

- American Psychological Association. (1991). In the Supreme Court of the United States, *Price Waterhouse v. Ann B. Hopkins*, Amicus curiae brief for the American Psychological Association. *American Psychologist*, 46, 1061–1070.
- Axson, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of Personality and Social Psychology*, 53, 30–40.
- Berger, M. A. (1994). Procedural paradigms for applying the *Daubert* test. *Minnesota Law Review*, 78, 1345–1386.
- Black, B., Ayala, F. J., & Saffran-Brinks, C. (1994). Science and the law in the wake of *Daubert*: A new search for scientific knowledge. *Texas Law Review*, 72, 715–802.
- 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.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Cooper, J., Bennett, E. A., & Sukel, H. L. (1996). Complex scientific testimony: How do jurors make decisions? *Law and Human Behavior*, 20, 379–394.
- Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law*. New York: Cambridge University Press.
- Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 113 S.Ct. 2786 (1993).
- Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
- Faigman, D. L. (1995). The evidentiary status of social science under *Daubert*: Is it “scientific,” “technical,” or “other” knowledge? *Psychology, Public Policy, and Law*, 1, 960–979.
- Federal Judicial Center. (1995). *Reference manual on scientific evidence*. New York: Thomson Legal Publishing.
- Federal rules of evidence*, 28 U.S.C. (West 1975).
- Fisher, D. E. (1994). *Daubert v. Merrell Dow Pharmaceuticals*: The Supreme Court gives federal judges the keys to the gate of admissibility of expert scientific testimony. *South Dakota Law Review*, 39, 141–158.
- 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.
- Frye v. United States*, 54 App.D.C. 46, 293 F. 1013 (1923).
- Giner-Sorolla, R., & Chaiken, S. (1997). Selective use of heuristic and systematic processing under defense motivation. *Personality and Social Psychology Bulletin*, 23, 84–97.
- Hagen, M. A. (1997). *Whores of the court: The fraud of psychiatric testimony and the rape of American justice*. New York: Harper Collins.
- Huffman v. Pepsi-Cola Bottling Co. of Minneapolis–St. Paul*, et al., Hennepin County Trial Court, File No. MC 92-10995, Decided June 20, 1994.
- Jepson, C., Krantz, D. H., & Nisbett, R. E. (1983). Inductive reasoning: Competence or skill? *Behavioral and Brain Sciences*, 6, 494–501.
- Jordan v. State*, 694 So. 2d 708; 1997 Fla. LEXIS 554.
- Kovera, M. B., Levy, R. J., Borgida, E., & Penrod, S. D. (1994). Expert testimony in child sexual abuse cases: Effects of expert evidence type and cross-examination. *Law and Human Behavior*, 18, 653–674.
- Kovera, M. B., McAuliff, B. D., & Hebert, K. S. (1999). Reasoning about scientific evidence: The effects of juror gender and evidence quality on juror decisions in a hostile work environment case. *Journal of Applied Psychology*, 84, 362–375.
- 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, 431–443.
- Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, 26, 952–960.
- Leippe, M. R., & Elkin, R. A. (1987). When motives clash: Issue involvement and response involvement as determinants of persuasion. *Journal of Personality and Social Psychology*, 52, 269–278.
- 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.
- McAuliff, B. D., & Kovera, M. B. (1999a). [Juror evaluations of flawed expert evidence.] Unpublished raw data.
- McAuliff, B. D., & Kovera, M. B. (1999b, July). Juror sensitivity to methodological flaws in expert evidence. In S. D. Penrod (Chair), *The use of scientific evidence: Empirical, legal, and comparative perspectives*. Symposium conducted at the meeting of the European Association for Psychology and Law, Dublin, Ireland.
- Nisbett, R. E. (1993). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.
- Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science*, 238, 625–631.
- 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.
- Ratneshwar, S., & Chaiken, S. (1991). Comprehension’s role in persuasion: The case of its moderating effects on the persuasive impact of source cues. *Journal of Consumer Research*, 18, 52–62.
- Redding, R. E., & Reppucci, N. D. (1999). Effects of lawyers’ socio-political attitudes on their judgments of social science in legal decision making. *Law and Human Behavior*, 23, 31–54.
- Robinson v. Jacksonville Shipyards, Inc.*, 760 F.Supp. 1486 (M.D.Fla. 1991).
- Rudman, L. A., & Borgida, E. (1995). The afterglow of construct accessibility: The behavioral consequences of priming men to view women as

- sexual objects. *Journal of Experimental Social Psychology*, 31, 493–517.
- Sallant, P., & Dillman, D. A. (1994). *How to conduct your own survey*. New York: Wiley.
- Shuman, D. W., Whitaker, E., & Champagne, A. (1994). An empirical examination of the use of expert witnesses in the courts—Part II: A three city study. *Jurimetrics: Journal of Law, Science and Technology*, 34, 193–208.
- Simard, L. S., & Young, W. G. (1994). *Daubert's gatekeeper*: The role of the district judge in admitting expert testimony. *Tulane Law Review*, 68, 1457–1475.
- Walker, L., & Monahan, J. (1996). *Daubert and the Reference Manual*: an essay on the future of science in law. *Virginia Law Review*, 82, 837–857.

Received October 28, 1998

Revision received September 29, 1999

Accepted September 29, 1999 ■

Instructions to Authors

Journal of Applied Psychology

Articles submitted for publication in the *Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, articles must be clearly written in concise and unambiguous language and must be logically organized. The goal of APA primary journals is to publish useful information that is accurate and clear.

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (4th ed.). Articles not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract containing a maximum of 960 characters and spaces (which is approximately 120 words) typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are copyedited for bias-free language (see chap. 2 of the *Publication Manual*). Original color figures can be printed in color provided the author agrees to pay half of the associated production costs.

The journal will publish both regular articles, or Feature Articles, and Research Reports. Authors can refer to recent issues of the journal for approximate length of Feature Articles. (Total manuscript pages divided by 3 provides an estimate of total printed pages.) Longer articles will be considered for publication, but the contribution they make must justify the number of journal pages needed to present the research. Research Reports feature shorter manuscripts that make a distinct but relatively narrow contribution, such as important replications or studies that discuss specific applications of psychology. Authors may request Research Report status at the time of submission, or the editor may suggest that a regular-length submission be pared down to Research Report length. Research Reports are limited to no more than 17 pages of text proper; these limits do not include the title page, abstract, references, tables, or figures. Different printers, fonts, spacing, margins, and so forth can substantially alter the amount of text that can be fit on a page. In determining the length limits of Research Reports, authors should count 25–26 lines of text (60 characters per line) as the equivalent of one page.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 6.24). As this journal is a primary journal that publishes original material only, APA

policy prohibits as well publication of any manuscript that has already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data upon which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 6.25). APA expects authors submitting to this journal to adhere to these standards. Specifically, authors of manuscripts submitted to APA journals are expected to have their data available throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242 (or see "Ethical Principles," December 1992, *American Psychologist*, Vol. 47, pp. 1597–1611).

The journal will accept submissions in masked (blind) review format only. Each copy of a manuscript should include a separate title page with author names and affiliations, and these should not appear anywhere else on the manuscript. Furthermore, author identification notes should be typed on the title page (see *Manual*). Authors should make every reasonable effort to see that the manuscript itself contains no clues to their identities. Manuscripts not in masked format will not be reviewed.

Authors must submit five (5) copies of the manuscript. The copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. In addition to addresses and phone numbers, authors should supply electronic mail addresses and fax numbers, if available, for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to Kevin R. Murphy, Editor, Department of Psychology, Pennsylvania State University, 458 Bruce V. Moore Building, University Park, Pennsylvania 16802-3104.
