



NEW JERSEY
ASSOCIATION
for JUSTICE

EDUCATIONAL FOUNDATION, INC.

Formerly the ATLA NJ Educational Foundation, Inc.

Mark S. Ali, Esq.

Criminal and Municipal Law

Meadowlands Seminar 2013

Wednesday - Friday

November 13 - 15, 2013

Sheraton Meadowlands

East Rutherford, NJ

Are forensic experts biased by the side that retained them?

Daniel C. Murrie

University of Virginia

Marcus T. Boccaccini

Sam Houston State University

Lucy A. Guarnera

University of Virginia

Katrina Rufino

Sam Houston State University

Manuscript in Press: *Psychological Science*

Please do not cite or distribute without author permission

Author Note

Daniel C. Murrie and Lucy Guarnera, Institute of Law, Psychiatry, and Public Policy, University of Virginia. Marcus T. Boccaccini and Katrina Rufino, Psychology Department, Sam Houston State University. Correspondence concerning this article should be addressed to Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy. UVA Box 800660, Charlottesville, VA, 22908-0660. Email: Murrie@Virginia.edu

This research was supported by the National Science Foundation, Law & Social Science Program, Award SES 0961082

ARE FORENSIC EXPERTS BIASED?

Abstract

How objective are forensic experts when they are retained by opposing sides in adversarial legal proceedings? Despite longstanding concerns from the legal system, we know little about whether experts can provide opinions unbiased by the side that retained them. In this experiment, we paid 108 forensic psychologists and psychiatrists to review the same offender case files, but deceived some to believe they were consulting for the defense, and some to believe they were consulting for the prosecution. Experts scored each offender on two commonly-used, well-researched risk assessment instruments. Experts who believed they were working for the prosecution tended to assign higher risk scores to offenders, whereas those who believed they were working for the defense tended to assign lower risk scores to the same offenders, with effect sizes up to $d = .85$. Results provide strong evidence of an allegiance effect among some forensic experts in adversarial legal proceedings.

ARE FORENSIC EXPERTS BIASED?

Are forensic experts biased by the side that retained them?

Recently, the National Research Council (NRC, 2009) warned that the accuracy and reliability of many popular forensic science techniques are unknown, that error rates are rarely acknowledged, and that forensic scientists are prone to bias because they “lack independence” from those requesting their services. Emerging research has clearly documented subjectivity and bias even in the forensic science procedures that courts have tended to consider most reliable, such as analyses of DNA (Dror & Hampikian, 2011) and fingerprints (Dror & Cole, 2010). Thus the NRC urged further research on the cognitive and contextual biases that influence forensic experts.

The NRC report did not specifically address mental health experts or forensic psychological evaluations. But psychological evaluations—like other forensic science procedures—are often admitted as evidence or presented via expert testimony in adversarial legal proceedings. Indeed, evaluations by mental health experts influence decisions as grave as death sentences (*Barefoot v. Estelle*, 1983) and indefinite civil confinement (*Kansas v. Hendricks*, 1997). Therefore, recent concerns regarding forensic science raise questions about whether forensic psychological evaluations might suffer similar problems of unreliability and bias.

So how reliable are forensic psychologists and psychiatrists when they are retained as experts in adversarial legal proceedings? For more than a century, courts and legal scholars have lamented apparent bias among medical experts (Bernstein, 2008; Hand, 1901; Mnookin, 2008; Wigmore, 1923). Likewise, practicing judges and attorneys complain that experts sacrifice objectivity for advocacy (e.g., Krafska et al., 2002). But little psychological research has investigated what we call “adversarial allegiance” (Murrie et al., 2009), the presumed tendency for experts to reach conclusions that support the party who retained them. Psychology’s delay in

ARE FORENSIC EXPERTS BIASED?

investigating adversarial allegiance is disappointing, because psychologists are uniquely suited to explore reliability and bias in decision-making.

Field Studies of Risk Instruments Suggest, but Do Not Prove, Adversarial Allegiance

Recently, we investigated adversarial allegiance by examining sex offender civil commitment proceedings, also known as Sexually Violent Predator (SVP) trials. SVP trials provide an ideal context for studying the possibility of adversarial allegiance, because court decisions depend largely on weighing testimony from opposing experts. Twenty states and the federal system have SVP laws, which allow them to identify sexual offenders whom they consider likely to re-offend and confine them indefinitely after their incarceration (*Kansas v. Hendricks*, 1997). SVP proceedings routinely involve forensic psychologists and psychiatrists retained by opposing sides who conduct risk assessments of the same offender and consider the same data, often using the same instruments. So we could study adversarial allegiance in SVP proceedings by comparing the scores that defense-retained and prosecution-retained evaluators assigned on popular risk assessment instruments (Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009).

Scores on risk instruments are an ideal metric to measure expert opinions, because (a) experts routinely administer these instruments to inform legal proceedings, and (b) dozens of studies document strong rater agreement when clinicians score these instruments in research and practice contexts that are *not* adversarial. For example, Hare's (2003) Psychopathy Checklist-Revised (PCL-R), an instrument that relies on clinical interview and review of records, is widely used in forensic assessments of risk for violence or sexual violence (Skeem et al, 2011). The PCL-R manual reports strong rater agreement statistics (intraclass correlation [ICC] = .87; Hare, 2003). Indeed, most (92%) pairs of scores from trained raters who score the same offender differ

ARE FORENSIC EXPERTS BIASED?

by fewer than two points (Gacono & Hutton, 2004), even though PCL-R scores can range from 0-40.

However, in a small sample of SVP proceedings that featured PCL-R scores from defense-retained and prosecution-retained evaluators, the ICC for opposing evaluators was .42, indicating that less than half of the variance in PCL-R scores could be attributed to the offenders' true standing on the PCL-R (Murrie et al., 2009). Moreover, the average PCL-R score from prosecution experts was 24, whereas the average score from defense experts was only 18 (Cohen's $d = .78$). The PCL-R may be especially vulnerable to this allegiance effect because it requires clinicians to make inferences about an offender's personality and emotions (e.g., lack of guilt or remorse, superficial charm). The adversarial allegiance effect was smaller ($d = .34$) for the Static-99 (Hanson & Thornton, 2000), a highly structured measure scored from file information about criminal history that requires less subjective judgment.

These field studies (Murrie et al., 2008; 2009) strongly suggest adversarial allegiance, in that prosecution-retained evaluators assigned higher scores and defense-retained evaluators assigned lower scores to the same offenders. But we cannot draw firm conclusions from these field studies alone, because they investigated scores from experts selected by attorneys. Conceivably, attorneys could have chosen specific experts because they perceived the experts *already* held attitudes or scoring tendencies conducive to their case. Or perhaps attorneys consulted many experts, but arranged testimony only from those whose opinions were most supportive of their case. For example, a defense attorney might retain several evaluators to examine a client, but request testimony only from the evaluator who assigned the lowest risk scores. Thus, the apparent allegiance in field studies might reflect *selection* effects, whether in

ARE FORENSIC EXPERTS BIASED?

terms of which expert an attorney selected to perform an evaluation or which findings an attorney selected to present at trial.

Understanding Adversarial Allegiance Requires a True Experiment

Field studies raise an important question that can only be answered with a true experiment. Is apparent allegiance due simply to attorneys choosing evaluators who have pre-existing attitudes that favor their side, or attorneys calling only experts with the most favorable findings to testify in court (selection effects)? Or do evaluators, once retained and promised payment by one side, tend to form opinions that favor that side (allegiance effects)? If an experiment using random assignment failed to find allegiance effects, it would suggest that the apparent allegiance in the field is due primarily to one or both of these selection effects. But if an experiment using random assignment *did* find allegiance effects, it would suggest that being retained and paid by one side in an adversarial system may compromise objectivity among experts.

To answer this question, we recruited over 100 experienced forensic psychologists and psychiatrists, provided two days of in-person training on risk instruments from established experts, had them meet with an attorney, and then paid them to score risk instruments for up to four offenders. We deceived participants to believe they were performing a large-scale, paid forensic consultation. But unbeknownst to participants, they all received exactly the same four offender files, and each participant was randomly assigned to believe that they were working for the prosecution *or* the defense.

Method

Participants

ARE FORENSIC EXPERTS BIASED?

We sent recruitment correspondence to a broad group of practicing forensic evaluators offering “gold standard” training (and continuing education credits) on the two most commonly used measures in sex offender risk assessments: the PCL-R and Static-99R. This training was offered at no cost to participants who could commit to returning a few weeks later to spend one day scoring offenders at a pay rate typical of forensic consultation (\$400). We received over 100 applications from practicing, doctoral-level forensic clinicians (PhD, PsyD, or MD).

Of the 118 clinicians who participated in the risk measure training, 108 returned to score files for the experiment.¹ Five who scored cases did not pass a manipulation check (i.e., they could not identify which side had retained them) and four expressed some suspicion that the cover story of scoring cases for a forensic consultation was a sham (see *Debriefing*). So we report results for the 99 participants (49 = defense, 50 = prosecution) who accepted the study manipulation and believed they were scoring cases for one side of an adversarial process.

Participants (60% female) came from 15 states. Most reported doctoral degrees in psychology (PhD/PsyD, 88%). Others reported a medical degree (7%) or another type of doctoral degree (5%). Most (84%) reported experience conducting forensic evaluations, and most (75%) reported experience conducting sex offender risk assessments. About half (51%) had used the PCL-R in practice, and about half (49%) had used the Static-99R in practice.

Training

The participants attended a single two-day training. The first 1.5 days (14 hours) involved training on the PCL-R, conducted by a well-known expert who co-authored one Psychopathy Checklist measure and provided many formal PCL-R workshops. The final half-day of training (4 hours) focused on the Static-99R. Our goal was not to train participants to a

¹ Of the 10 clinicians who failed to return for file scoring, most explained they were absent because they had been called to court to provide testimony as part of their professional practice.

ARE FORENSIC EXPERTS BIASED?

predetermined level of reliability (common practice in validity studies) because evaluators in the field are never required to demonstrate a specified level of reliability before accepting cases. Rather, we provided training to ensure that all participants had, at a minimum, completed the type of high-quality workshop that is offered to professionals in the field. Many evaluators cite workshop training as evidence of their qualifications to score risk measures for SVP cases (Rufino, Boccaccini, Hawes, & Murrie, 2012), although it is possible some evaluators administer these measures with less formal training. Regarding deception at the training stage, participants were informed only that the training and subsequent scoring was funded by an “out-of-state agency” that wanted to ensure that all participants had rigorous training before they scored offender files.

Deception and Experimental Manipulation: Scoring Cases for Prosecution or Defense

Participants returned about three weeks later to score offender files. They were randomly assigned² to either a prosecution-allegiance or defense-allegiance group and deceived to believe that they were a part of a formal, large-scale forensic consultation paid for by either a public defender service or a specialized prosecution unit that prosecutes SVP cases. Immediately after arrival, participants met for 10-15 minutes with a confederate (a former SVP attorney) who posed as an attorney for either the public defender service or the specialized prosecution unit. The same attorney played both roles, but followed a slightly different script (see online supplement) depending on whether the participant had been randomly assigned to the defense or prosecution.

² To reduce the possibility of researcher expectations influencing study results, three of the four researchers were blinded to participant study conditions (inevitably one research assistant, who managed the random assignment, was aware).

ARE FORENSIC EXPERTS BIASED?

The attorney addressed the defense-allegiance participants with statements that are typical of many defense attorneys (e.g., “We try to help the court understand that the data show not every sex offender really poses a high risk of reoffending”). Likewise, he addressed participants in the prosecution-allegiance condition with statements that are typical of prosecutors. Across both conditions, he asked experts to score the relevant risk instruments. He also hinted at the possibility of future opportunities for paid consultation (see Online Supplement).

Participants were led to believe that, as a group, they were reviewing and scoring cases from a large cohort. But in truth, each participant was scoring the same four case files, which we selected to span the range from low risk to high risk. Each set of case materials was authentic (i.e., from an actual SVP case). The files included de-identified, but real, court, criminal and correctional records. Specifically, these included real police investigation and arrest documents; victim and witness statements; plea, judgment, and sentencing documents from court; pre-sentence investigation reports; criminal history summary documents; prison intake and case summary documents; prison placement documents; and prison disciplinary records. Prison records also included some material from routine psychological assessments performed by the prison’s sex offender treatment program, i.e., results from the Personality Assessment Inventory (Morey, 1991) and a clinical interview (similar in content to a PCL-R interview) conducted by treatment staff. Again, all of these records were real, but de-identified material unique to each of the four cases. Finally, each file also included a realistic transcript of a *fabricated* PCL-R interview that we wrote to correspond to each set of records. The fabricated PCL-R interview transcripts were cosmetically altered to appear as if they were part of the original records.

ARE FORENSIC EXPERTS BIASED?

The four offender files were selected to be representative of SVP cases generally. One sex offender had adult victims, whereas three had child victims. All had been convicted of multiple sexual offenses. After the participants reviewed a case file,³ they scored the PCL-R and Static-99R.

Measures

Psychopathy Checklist-Revised. Hare's (2003) PCL-R is a 20-item measure of interpersonal, emotional, and behavioral traits, which clinicians score on the basis of an offender's records and a clinical interview. PCL-R items are rated on a scale from 0 to 2, with higher scores reflecting a higher level of the psychopathic trait, and summed to yield a Total score that can range from 0 to 40. Although forensic evaluators usually emphasize PCL-R *Total* scores in reports or testimony, PCL-R items are divided into two factors: Factor 1 consists of an *Interpersonal* facet and an *Affective* facet, and Factor 2 consists of an *Impulsive Lifestyle* facet and an *Antisocial Behavior* facet.

The PCL-R is the most widely-used and well-researched measure of psychopathy, a personality construct characterized by a self-serving interpersonal style, shallow emotions, an unstable lifestyle, and antisocial behavior. Although not originally developed for risk assessment, ample research suggests PCL-R scores correspond with violence and recidivism. For example, meta-analyses find that PCL-R Total scores tend to be moderately associated with antisocial behavior (Leistico, Salekin, DeCoster, & Rogers, 2008), including sexual violence

³ The order of administration was randomized for three of four cases. Pilot testing suggested that most participants would be able to score three files in one day, but that some might be unable to complete four. Therefore, we provided the first three offender files to participants in a randomized order, to ensure we would have a similar, sufficient *n* for each of these three cases. A fourth case was provided to all participants last, anticipating that time constraints may preclude many participants from completing it.

ARE FORENSIC EXPERTS BIASED?

(Hawes, Boccacini, & Murrie, in press). Thus, the measure has become widely used in assessments of violence risk or sexual violence risk, and courts routinely admit expert testimony regarding PCL-R scores (DeMatteo & Edens, 2006).

The PCL-R manual (Hare, 2003) reports strong agreement among independent raters for PCL-R Total scores (intraclass correlation [ICC] = .87), at least outside of adversarial legal proceedings. But the manual also reveals that inter-rater agreement tends to be stronger for Factor 2 items that relate to antisocial behavior (e.g., “Criminal Versatility,” “Juvenile Delinquency”) and weaker for Factor 1 items such as “Failure to Accept Responsibility,” and “Glibness/Superficial Charm,” which may require more clinical inference.

Static-99R. The Static-99R is an actuarial risk assessment instrument designed to predict sexual recidivism among sex offenders (Helmus, Thornton, Hanson, & Babchishin, 2012). Composed of 10 items that address an offender’s age, prior living arrangements, and several aspects of his offense history, the Static-99R is scored based on file review. According to Static-99.org, the Static-99 is “the most widely used sex offender risk assessment instrument in the world, and is extensively used in the United States, Canada, the United Kingdom, Australia, and many European nations.” It is widely accepted in legal proceedings, given its strong empirical relation to important outcomes, and strong evidence of validity and reliability. For example, Static-99 scores are among the best known predictors of sexual recidivism, with a meta-analytic effect of $d = .67$ across more than 60 studies (Hanson & Morton-Bourgon, 2009). A recent review of rater-agreement coefficients found a median rater agreement value of .90 (Hanson & Morton-Bourgon, 2009), suggesting that the Static-99 and Static-99R meet or exceed commonly accepted standards for reliability in psychological measures. Compared with PCL-R items,

ARE FORENSIC EXPERTS BIASED?

Static-99R items (e.g., “Age at release,” “Any male victims”) appear fairly straightforward and require less clinical inference to score.

Clinician Attitudes. One potential explanation for any allegiance effects we might observe would be pre-existing differences in clinician attitudes (i.e., if participants assigned to score for the prosecution tended to have a harsher perspective on sexual offenders than participants assigned to score for the defense). So, although we randomly assigned participants to prosecution and defense conditions, we nevertheless had participants complete two additional measures that allowed us to check whether participants in each condition were similar regarding attitudes about sexual offenders.

Participants completed a five-item questionnaire at the *end* of the scoring day, to avoid revealing that their attitudes and scoring patterns were the focus of study. The questionnaire asked them to rate the extent to which restrictive policies for sex offenders (e.g., SVP laws) are necessary and reasonable. For example, “Laws that allow states to civilly commit potentially dangerous sex offenders who have completed their sentences are reasonable strategies to protect people in the community” (1 = strongly disagree, 5 = strongly agree). Internal consistency for this attitudes measure was .79. We also asked participants (at the end of PCL-R training) to report their best estimate of the typical PCL-R Total score among offenders who have committed sexually violent crimes against (a) adults and (b) children.

Debriefing

After participants completed the presumed forensic consultation, we performed a manipulation check, in which a member of the research team met privately with each participant. The researcher asked about the participant’s understanding of study goals, and then asked explicitly whether they were suspicious about any additional or hidden study goals. The four

ARE FORENSIC EXPERTS BIASED?

participants who conveyed *any* degree of suspicion (ranging from vague suspicion to more specific guesses about alternate study goals) were excluded from subsequent data analysis, as were the five who could not identify which side retained them. The researcher then described the experimental manipulation and the true study goals. Although all participants had the option of withdrawing their data from the study, none did so. All received the payment (\$400) and CEU credits originally promised.

Results

Overall, the risk measure scores assigned by prosecution and defense experts showed a clear pattern of adversarial allegiance. As expected, allegiance effects were stronger for the PCL-R, a measure that requires more subjective clinical judgment, than the Static-99R, a measure that requires less clinical judgment (see Table 1). For the PCL-R Total score, independent-samples *t*-tests indicated that prosecution evaluators assigned significantly higher scores than defense evaluators for Case 1 ($t [94] = 4.15, p < .001$), Case 2 ($t [94] = 3.73, p < .001$), and Case 3 ($t [97] = 2.71, p = .008$), but not Case 4 ($t [62] = -0.33, p = .97$). Cohen's *d* effect sizes for the three cases with significant effects ranged from .55 to .85, similar in magnitude to effects ($d = .63$ to $.83$) documented in a sample of actual SVP proceedings (Murrie et al., 2009). The one case without a PCL-R allegiance effect was one we selected to be unusually low in psychopathy;⁴ this case received unusually low scores both from prosecution-retained ($M = 7.81$) and defense-retained ($M = 7.84$) evaluators.

⁴ We included this unusual case for exploratory purposes because we hypothesized that there may be some “floor effect” to allegiance. That is, we wondered whether some offenders might be so low in psychopathy that evaluators (regardless of side that retained them) would score the offender similarly. This seemed to be the case. However, because this exploratory case was the last file provided to participants (the order of administration was randomized for the first three cases), and fewer participants

ARE FORENSIC EXPERTS BIASED?

Adversarial allegiance effects were evident for both Factor 1 (Interpersonal/Affective) and Factor 2 (Social Deviance) scores from the PCL-R, as detailed in Table 2. In terms of absolute value, Factor 1 effects were larger than Factor 2 effects in two of the three cases with Total score allegiance effects, which is consistent with findings that Factor 1 items tend to require more subjective judgment to score (Rufino, Boccaccini, & Guy, 2011). For case 3, however, there was a significant effect for Factor 2 scores ($d = .73, p < .001$), but not Factor 1 scores ($d = .24, p = .24$). Examination of the Factor 1 facets indicated that there was some evidence for an allegiance effect for Facet 2 (Affective traits) scores [$t(97) = 1.94, p = .06, d = .39, 95\% \text{ CI} [-.01, .79]$], but not Facet 1 (Interpersonal traits) scores [$t(97) = 0.08, p = .94, d = .01, 95\% \text{ CI} [-.38, .41]$].

For the Static-99R, a more structured measure, prosecution-retained evaluators again tended to assign higher scores than defense evaluators in each of the four cases (see Table 1). But the difference was large enough to reach statistical significance for only Case 1 ($d = .42$). The effect sizes across these four cases ($d = .14, .20, .24, .42$) are similar to, although somewhat smaller than, the effect sizes ($d = .29$ to $.37$) reported across 27 actual SVP cases (Murrie et al., 2009).

Differences among Pairs of Prosecution and Defense Evaluators

In court, judges and juries would never consider risk instrument scores that have been averaged across many experts. Rather, they usually hear expert testimony about risk scores from two experts: one called by each opposing side. Moreover, because *all* test scores are influenced to some extent by random measurement error, it is unrealistic to expect two experts to assign exactly the same score in every case. Small score differences may be trivial, even if they are in

completed it (see Table 1), it is conceivable that some of the difference in results was attributable to these other factors.

ARE FORENSIC EXPERTS BIASED?

the direction of allegiance. The mean scores in Table 1 do not provide any information about how often, if ever, we might expect large, non-trivial differences in risk scores from pairs of opposing experts.

Therefore, we conducted a series of follow-up analyses to examine how likely it was that a randomly selected prosecution expert and a randomly selected defense expert would assign scores that were so different that they could not be explained by expected random measurement error. We considered the difference between a pair of scores to be meaningfully different if it was more than twice the standard error of measurement (SEM) for the risk instrument. The SEM is the amount that experts' scores could be expected to differ due to random measurement error when evaluating the same offender. Based on the normal curve, we would expect only about 32% of difference scores to be larger than the SEM, and only about 4% to be more than twice as large as the SEM (i.e., ≥ 2 SEM units). In the absence of adversarial allegiance, we would expect prosecution evaluators to assign scores that are more than twice the SEM higher than defense evaluators in about 2% of cases, and vice versa.

For each of the four cases, we calculated a difference score for each possible pairing of prosecution and defense experts. This process yielded approximately 2,400 difference scores for each measure, for each case. We then calculated the percentage of difference scores that were more than twice the SEM in the direction of allegiance (prosecution > defense) and the percentage that were more than twice the SEM in the opposite direction (see Table 2). The SEM for the PCL-R is about 3.0 points, and the SEM for the Static-99R is about 1.0 point.

The findings in Table 2 show two clear effects. First, more than 20% of the score pairings for each case led to score difference that were more than twice the SEM, although only about 4% of score pairings in research contexts lead to score differences this large. For several

ARE FORENSIC EXPERTS BIASED?

offender files, nearly 40% of the difference scores were larger than two SEMs apart. Second, most large (i.e., ≥ 2 SEM) differences were in the direction of adversarial allegiance, with the prosecution expert assigning higher scores and the defense expert assigning lower scores. This pattern was especially clear for the PCL-R. For the three cases with clear PCL-R allegiance effects, 28% or more of all possible score pairings led to a score difference of more than two SEM in the direction of allegiance. Again, score differences of ≥ 2 SEM in one direction (e.g., prosecution > defense) should occur in only about 2% of cases, according to rater-agreement values from non-adversarial research contexts. Between 4% and 9% of PCL-R score pairings in the three cases with clear allegiance effects led to large differences in the opposite direction, which is also more than the 2% expected based on non-adversarial research, but these differences clearly were not as common as large differences in the direction of allegiance (>28%).

Potential Explanations for Allegiance Effects

One possible alternate explanation for our findings is that, despite random assignment, evaluators assigned to score for the prosecution maintained harsher attitudes toward sex offenders or had different types of clinical experience than those assigned to score for the defense. But we found *no* evidence for this alternate explanation. Prosecution and defense evaluators did not differ in their ratings on our on our five-item measure of support for restrictive sex offender policies [$t(97) = 0.07, p = .95, d = .02$], their estimate of the typical PCL-R Total score among sex offenders with adult victims [$t(93) = 0.51, p = .62, d = .10$], or their estimate of the typical PCL-R Total score assigned to sex offenders with child victims [$t(93) = .25, p = .80, d = .05$]. Likewise, prosecution and defense evaluators did not differ in having used the Static-99R in practice (52% vs. 45%; $\chi^2(1, N = 99) = 0.50, p = .48, OR = 1.33$). Those assigned to score for the prosecution were somewhat more likely (62%) to have used the PCL-R in practice than those

ARE FORENSIC EXPERTS BIASED?

assigned to score for the defense [41%; $\chi^2(1, N = 99) = 4.45, p = .04, OR = 2.36$], but this is a difference that would actually *reduce* the likelihood of observing an allegiance effect.

Those with less prior clinical experience using these measures and those with higher scores on the attitude measures did tend to assign higher scores in some cases, but these effects were similar in size and direction for state and defense evaluators (Guarnera, Murrie, Boccaccini, & Rufino, 2012). We could find only one instance in which an attitude or experience measure might help explain an allegiance effect. Recall that the strongest Static-99R allegiance effect occurred in Case 1 ($d = .42$). A two-way analysis of variance revealed a statistically significant interaction between condition assignment and prior use of the Static-99R in practice, $F(1, 91) = 4.38, p = .04$. Specifically, there was a clear allegiance effect for those who had *not* used the Static-99R in practice ($d = .71, 95\% CI [.12, 1.29]$), but no evidence of an effect for those who had used the Static-99R in practice ($d = .00, 95\% CI [-.12, .12]$). However, there was no evidence of a similar interaction for Static-99R scores from other cases, or for PCL-R scores from any case. In short, we could find *no* variables that seemed to explain the allegiance effects we observed overall.

Discussion

Results from this study underscore recent concerns about forensic sciences (NRC, 2009)—and raise concerns specific to forensic psychology—by demonstrating that some experts who score ostensibly objective assessment instruments assign scores that are biased toward the side that retained them. In the field, some apparent adversarial allegiance may result partly from selection effects (i.e., savvy attorneys select experts who are predisposed to the attorney's perspective, or present input only from experts who favored their perspective), but our results

ARE FORENSIC EXPERTS BIASED?

suggest that *even without selection effects*, the pull of adversarial proceedings tends to influence opinions by paid forensic experts.

Of course, there was considerable variability in scores even from experts assigned to the same side, and certainly not every expert produced scores consistent with adversarial allegiance. But the systematic score differences among opposing experts could not be explained by chance, standard error of measurement, or pre-existing differences among the experimental groups.

This evidence of allegiance was particularly striking because our experimental manipulation was less powerful than experts are likely to encounter in most real cases. For example, our participating experts spent only 15 minutes with the retaining attorney, whereas experts in the field may have extensive contact with retaining attorneys over weeks or months. Our experts formed opinions based on files only, which were identical across opposing experts. But experts in the field may *elicit* different information by seeking different collateral sources or interviewing offenders in different ways. Therefore, the pull toward allegiance in this study was relatively weak compared with the pull typical of most cases in the field. So the large group differences provide compelling evidence for adversarial allegiance.

Our study could not identify the mechanisms responsible for the allegiance effect. We do not know whether the effect was more attributable to the initial conversation with an attorney, a sense of team loyalty, monetary payment, or the promise of future work. We do not know the role of confirmation bias, anchoring, or other potentially important cognitive mechanisms. Of course, the role of each mechanism may have varied by participant, and not all participants demonstrated an allegiance effect. Future research must attempt to disentangle the roles of these mechanisms, and identify evaluator characteristics that are associated with allegiance.

ARE FORENSIC EXPERTS BIASED?

Although this study addressed only one kind of evaluation (i.e., assessment of risk for sexual re-offense), there is little reason to believe that this is the only kind of forensic psychological evaluation or forensic science procedure vulnerable to allegiance effects. Indeed, the evidence of allegiance effects on structured, ostensibly objective instruments that usually reveal strong rater agreement leaves us even more concerned about the possibility of allegiance effects on expert procedures that are less structured or less guided by scoring rules. Many forensic science procedures rely heavily on subjective judgment (e.g., matching bite marks, hair fibers, or tire treads; NRC, 2009), as do many opinions psychologists offer in court (e.g., assigning diagnoses or assessing emotional injury). Our findings underscore the need for research addressing the cognitive and procedural biases that may facilitate adversarial allegiance, and potential interventions to reduce allegiance. Indeed, our findings suggest there may be opportunities to improve forensic psychological practice, broader forensic science practice, and even legal policy and procedures in ways that might better promote scientific objectivity and reduce adversarial allegiance.

ARE FORENSIC EXPERTS BIASED?

References

- Barefoot v. Estelle*, 463 U. S. 880 (1983).
- Bernstein, D. E. (2008). Expert witnesses, adversarial bias, and the (partial) failure of the Daubert Revolution. *Iowa Law Review*, *93*, 101-137.
- DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist-Revised in court: A case law survey of U.S. courts (1991-2004). *Psychology, Public Policy, and Law*, *12*, 215-241. doi: 10.1037/1076-8971.12.2.214
- Dror, I. E. & Cole, S. A. (2010). The visit in “blind” justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*, 161-167. doi: 10.3758/PBR.17.2.161
- Dror, I. E. & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science and Justice*, *51*, 204-208. doi:10.1016/j.scijus.2011.08.004
- Gacono, C., & Hutton, H. (1994). Suggestions for the clinical and forensic use of the Hare Psychopathy Checklist–Revised (PCL-R). *International Journal of Law and Psychiatry*, *17*, 303-317.
- Guarnera, L. A., Murrie, D. C., Boccaccini, M. T., & Rufino, K. (2012). Do attitudes affect psychopathy scores evaluators assign to sexual offenders in Sexually Violent Predator proceedings? Paper presented at the meeting of the American Psychology-Law Society, San Juan, Puerto Rico.
- Hand, L. (1901). Historical and practical considerations regarding expert testimony. *Harvard Law Review*, *15*, 40-58.
- Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis. *Psychological Assessment*, *21*, 1-21. doi: 10.1037/a0014421
- Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119-136. doi: 10.1023/A:1005482921333

ARE FORENSIC EXPERTS BIASED?

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised: Second Edition*. Toronto, Ontario, Canada: Multi-Health Systems.

Hawes, S. W., Boccaccini, M. T., & Murrie, D. C. (in press). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist-Revised. *Psychological Assessment*.

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: Journal of Research and Treatment, 24*, 64-101.

doi:10.1177/1079063211409951

Kansas v. Hendricks, 521 U. S. 346 (1997).

Krafka, C., Dunn, M. A., Johnson, M. T., Cecil, J. S., & Miletich, D. (2002). Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. *Psychology, Public Policy, and Law, 8*, 309–32. doi: 10.1037/1076-8971.8.3.309

Leistico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28-45. doi: 10.1007/s10979-007-9096-6

Mnookin, J. (2008). Expert evidence, partisanship, and epistemic confidence. *Brooklyn Law Review, 73*, 587-611.

Morey, L.C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior, 32*, 352-362. doi: 10.1007/s10979-007-9097-5

ARE FORENSIC EXPERTS BIASED?

- Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C. & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*, 19-53. doi: 10.1037/a0014897
- Murrie, D. C., Boccaccini, M. T., Caperton, J., & Rufino, K. (2012). Field validity of the Psychopathy Checklist-Revised in sex offender risk assessment. *Psychological Assessment, 24*, 524-529. doi: 10.1037/a0026015
- National Research Council, Committee on Identifying the Needs of the Forensic Science Community (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC. The National Academies Press.
- Rufino, K. A., Boccaccini, M. T., & Guy, L., S. (2011). Scoring subjectivity and item performance on measures used to assess violence risk: The PCL-R and HCR-20 as exemplars. *Assessment, 18*, 453-463. doi: 10.1037/t01167-000
- Rufino, K. A., Boccaccini, M. T., Hawes, S., & Murrie, D. C. (2012). When experts disagreed, who was correct? A comparison of PCL-R scores from independent raters and opposing forensic experts. *Law and Human Behavior, 36*, 527-531. doi: 10.1037/h0093988
- Skeem, J., Polaschek, D., Patrick, C., Lilienfeld, S. (2011). Psychopathic personality: Bridging the gap between scientific evidence and public policy. *Psychological Science in the Public Interest, 12*, 95-162. doi: 10.1177/1529100611426706
- Wigmore, J. (1923). *A treatise on the Anglo-American system of evidence in trials at common law: Including the statutes and judicial decisions of all jurisdictions of the United States and Canada*. Little Brown; Boston, MA.

ARE FORENSIC EXPERTS BIASED?

Table 1

Risk measure scores from evaluators randomly assigned to score cases for the prosecution and evaluators randomly assigned to score cases for the defense

Case	Prosecution		Defense		Cohen's <i>d</i>	95% CI
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Psychopathy Checklist-Revised (PCL-R) Total Score						
1	16.64	3.50	13.41	4.10	.85***	[.43, 1.26]
2	26.53	4.32	23.22	4.37	.76***	[.35, 1.17]
3	26.40	4.69	24.00	4.14	.55**	[.14, .94]
4	7.81	4.09	7.84	3.36	-.01	[-.32, .31]
PCL-R Factor 1 (Interpersonal/Affective)						
1	11.22	2.60	8.95	3.20	.78***	[.36, 1.18]
2	8.34	2.72	6.51	2.95	.65**	[.23, 1.05]
3	11.91	2.80	11.27	2.52	.24	[-.15, .63]
4	4.74	3.30	4.60	2.66	.05	[-.44, .54]
PCL-R Factor 2 (Social Deviance)						
1	3.86	1.68	3.13	1.60	.44*	[.04, .85]
2	15.61	2.26	14.45	2.19	.52**	[.11, .93]
3	12.26	2.36	10.65	2.00	.73***	[.33, 1.14]
4	2.58	1.45	2.98	1.79	-.25	[-.74, .25]
Static-99R						
1	4.46	0.85	4.06	1.05	.42*	[.01, .82]
2	5.56	1.35	5.27	1.05	.24	[-.16, .64]
3	5.62	1.81	5.29	1.57	.20	[-.20, .59]
4	1.85	1.21	1.69	1.11	.14	[-.35, .64]

Note. Statistical significance determined using independent samples t-tests (two-tailed). *N* = 96 (Case 1), 96 (Case 2), 99 (Case 3), 64 (Case 4). **p* ≤ .05. ***p* ≤ .01. ****p* ≤ .001.

ARE FORENSIC EXPERTS BIASED?

Table 2

Percentage of opposing evaluator pairs that led to score differences greater than twice the standard error of measurement.

Case	Prosecution > Defense	Defense > Prosecution
	<hr/>	
	PCL-R	
1	29%	4%
2	33%	7%
3	28%	9%
4	13%	12%
	Static-99R	
1	18%	7%
2	20%	12%
3	28%	21%
4	20%	18%