Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations.

Article *in* Translational Issues in Psychological Science · June 2017

DOI: 10.1037/tps0000114

CITATIONS
40

READS
3,888

3 authors, including:

Lucy Guarnera
University of Virginia
11 PUBLICATIONS **370** CITATIONS

SEE PROFILE

Daniel C Murrie
University of Virginia
113 PUBLICATIONS **3,689** CITATIONS

SEE PROFILE

Translational Issues in Psychological Science
2017, Vol. 3, No. 2, 143–152

2332-2136/17/$12.00 http://dx.doi.org/10.1037/tps0000114

Why Do Forensic Experts Disagree? Sources of Unreliability and Bias in Forensic Psychology Evaluations

Lucy A. Guarnera
University of Virginia

Daniel C. Murrie
University of Virginia School of Medicine

Marcus T. Boccaccini
Sam Houston State University

Recently, the National Research Council, Committee on Identifying the Needs of the Forensic Science Community (2009) and President's Council of Advisors on Science and Technology (PCAST; 2016) identified significant concerns about unreliability and bias in the forensic sciences. Two broad categories of problems also appear applicable to forensic psychology: (1) unknown or insufficient field reliability of forensic procedures, and (2) experts' lack of independence from those requesting their services. We overview and integrate research documenting sources of disagreement and bias in forensic psychology evaluations, including limited training and certification for forensic evaluators, unstandardized methods, individual evaluator differences, and adversarial allegiance. Unreliable opinions can result in arbitrary or unjust legal outcomes for forensic examinees, as well as diminish confidence in psychological expertise within the legal system. We present recommendations for translating these research findings into policy and practice reforms intended to improve reliability and reduce bias in forensic psychology. We also recommend avenues for future research to continue to monitor progress and suggest new reforms.

What is the significance of this article for the general public?
This review summarizes and integrates research on sources of disagreement and bias in forensic psychology evaluations, including limited training and certification, unstandardized methods, individual evaluator differences, and allegiance to the retaining party. Disagreement can result in arbitrary or unjust legal outcomes for forensic examinees, as well as diminish confidence in psychological expertise. Thus, policy and practice changes are needed to improve the reliability of forensic opinions.

Keywords: forensic evaluation, forensic instrument, adversarial allegiance, human factors, bias

Imagine you are a criminal defendant or civil litigant undergoing a forensic evaluation by a psychologist, psychiatrist, or other clinician. The forensic evaluator has been tasked with answering a difficult psycholegal question about you and your case. For example, "Were you sane or insane at the time of the offense? How likely is it that you will be violent in the future? Are you psychologically stable enough to fulfill your job duties?" The forensic evaluator interviews you, reads records about your history, speaks to some sources close to you, and perhaps administers some psychological tests. The evaluator then forms a forensic opinion about your case—and the opinion is not in your favor. You might wonder whether most forensic clinicians would have reached this same opinion. Would a second (or third,

Lucy A. Guarnera, Department of Psychology, University of Virginia; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia School of Medicine; Marcus T. Boccaccini, Department of Psychology, Sam Houston State University.

Correspondence concerning this article should be addressed to Lucy A. Guarnera, Department of Psychology, University of Virginia, P.O. Box 400400, Charlottesville, VA 22904-4400. E-mail: lag8e@virginia.edu

or fourth) evaluator have come to a different, perhaps more favorable conclusion? In other words, how often do forensic psychologists disagree? And why does such disagreement occur?

Questions about reliability and bias in forensic psychology feel more pressing as authorities ask similar questions about long-trusted forensic science procedures. Recently, the National Research Council, Committee on Identifying the Needs of the Forensic Science Community (2009) and President's Council of Advisors on Science and Technology (PCAST; 2016) reviewed the state of forensic science, covering a wide range of disciplines including analyses of DNA, fingerprints, hair, tire treads, bite marks, and ballistics. Both governmental councils concluded that the error rates of many forensic techniques are unknown, and that forensic scientists are prone to a variety of contextual biases. Consistent with the National Research Council (NRC) and PCAST's concerns, research has documented subjectivity and bias even in the forensic science procedures that courts have considered most reliable, such as analyses of DNA (Dror & Hampikian, 2011) and fingerprints (Dror & Rosenthal, 2008).

While forensic evaluators strive for objectivity and seek to avoid conflicts of interest (American Psychological Association, 2013), a forensic opinion may be influenced by multiple sources of variability and bias that can be powerful enough to cause independent evaluators to form different opinions about the same defendant (see Figure 1). The purpose of this review is to summarize and integrate research documenting various sources of disagreement in forensic evaluations, as well as suggest promising avenues of future research. We also present recommendations for translating these research findings into policy and practice reforms intended to improve the reliability of forensic evaluations.

The NRC and PCAST reports identified two broad categories of problems in forensic science that appear applicable to forensic psychology: (1) unknown or insufficient field reliability of forensic procedures, and (2) experts' lack of independence from those requesting their services. We address both of these areas in turn.

Unknown or Insufficient Field Reliability of Forensic Opinions

The (un)Reliability of Forensic Psychology?

Interrater reliability is the degree of consensus among multiple independent raters.[1] Of particular interest within forensic psychology is field reliability—the interrater reliability among practitioners performing under routine practice conditions typical of real-world work (Wood, Nezworski, & Stejskal, 1996). In general, the field reliability of forensic opinions is either unknown or far from perfect. For example, a recent meta-analysis concluded that for evaluations of adjudicative competency—one of the most common forensic psychology procedures—pairs of independent evaluators assessing the same defendant disagreed in approximately 15%–30% of cases (Guarnera & Murrie, in press). This corresponds to rater agreement coefficients (i.e., Cohen's kappa) in the range of .30–.65, which indicates fair to moderate agreement according to most kappa interpretation schemes (e.g., Landis & Koch, 1977). Field reliability rates for other common forensic opinions are similar although generally somewhat lower; pairs of independent evaluators tend to disagree in approximately 25%–35% of sanity cases ($\kappa \approx .25$–.65; Guarnera & Murrie, in press) and almost half (45%) of conditional release cases ($\kappa = .19$; Acklin, Fuger, & Gowensmith, 2015). In a related issue, the interrater reliability of forensic assessment instruments scored under routine practice conditions in the field is often poorer than what has been documented in controlled validation studies and reported in test manuals (C. S. Miller, Kimonis, Otto, Kline, & Wasserman, 2012).

We discuss many possible reasons for these less-than-ideal field reliability rates, but one key foundational explanation is that forming a forensic opinion is an extraordinarily difficult task. For example, evaluations of legal sanity require clinicians to use limited and often contradictory information to draw conclusions about the mental state of a defendant at the time they committed the crime, which may have been months or even years ago. A survey of a variety of medical and psychological proce-

[1] See, generally, Gwet (2014) for a more in-depth definition and discussion of interrater reliability.

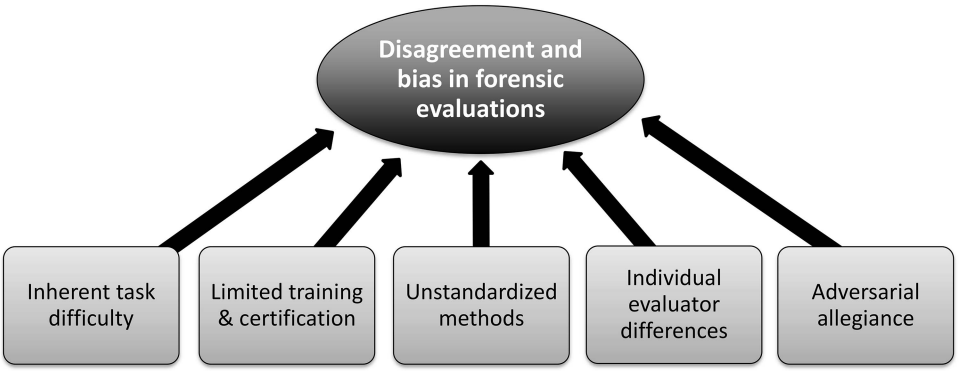

Figure 1. Sources of disagreement and bias in forensic evaluations.

dures confirms that complex decision tasks involving the integration of multiple sources of data, such as rating child behavior problems or classifying stroke severity, tend to settle at fair to moderate reliability rates (kappa or intraclass correlation [ICC] ≈ .30–.75; Meyer, Mihura, & Smith, 2005). This is in contrast to simple object counts (e.g., counting decayed or missing teeth) or physical measurements (e.g., measuring organ size on an ultrasound), where reliability tends to be higher, with rater agreement coefficients greater than .90 (Meyer et al., 2005). Along these lines, Mossman (2013) recently performed mathematical simulations of competency evaluations and concluded that fair to moderate reliability estimates were about as good as could reasonably be expected given the inherent difficulty of the task.

Limited Training and Certification for Forensic Evaluators

Besides the unreliability that may be intrinsic to a complex, ambiguous task such as forensic evaluation, research has identified multiple extrinsic sources of expert disagreement. One such source is limited training and certification for forensic evaluators. While specialized training programs and board certifications have become far more commonplace and rigorous since the early days of the field in the 1970s and 1980s, the training and certification of typical clinicians conducting forensic evaluations today remains variable and often poor (DeMatteo, Marczyk, Krauss, & Burl, 2009). For example, only about one third to one half of states have any state-level certification in forensic mental health assessment, and those that do may have weak standards (e.g., attend one brief, initial training session or have previous clinical experience; Gowensmith, Pinals, & Karas, 2015). Thus, many states continue to have the bulk of their forensic evaluations performed by "occasional experts," general clinicians without specialized forensic training (Grisso, 1987, p. 833).[2] Unsurprisingly, studies assessing the thoroughness, relevance, and accuracy of the reports forensic clinicians submit to the court routinely find them deficient (Fuger, Acklin, Nguyen, Ignacio, & Gowensmith, 2014). For example, Skeem and colleagues (1998) found that competency evaluators' reports in Utah failed to incorporate legally relevant aspects of competency and failed to adequately describe the reasoning underlying their final forensic opinion.

This training gap is important because empirical research suggests that evaluators with greater training produce more reliable forensic opinions. A compelling recent study conducted in Hawaii examined interrater reliability rates for three types of common forensic opinions (adjudicative competency, legal sanity, and violence risk assessment) both before and after the state adopted more stringent certification

[2] *Occasional experts* are likely more common in rural or other underresourced areas where forensic mental health assessments are needed, but no highly trained, board-certified forensic clinicians are available. Thus, the court's only option may be a general clinician without specialized training in forensic assessment.

standards in 2014 (Gowensmith, Sledd, & Sessarego, 2014). These new standards included a mandatory 3-day training, written test, submission of a mock report, peer review process, and continuing education. Postcertification, reliability rates improved for all three types of evaluations (competency: 13% increase, $p = .08$; sanity: 17% increase, $p = .04$; risk: 29% increase, $p = .001$). Gowensmith and colleagues' (2014) results provide the first direct evidence that more stringent state-level certification standards can improve the field reliability of forensic opinions.

Unstandardized Methods

One likely reason why training and certification increase interrater reliability is that they promote standardized evaluation methods among forensic clinicians. While there are now greater resources and consensus concerning appropriate practice than even a decade ago, forensic psychologists still vary widely in what they actually *do* during any particular forensic evaluation (Heilbrun & Brooks, 2010). For example, Neal and Grisso (2014) found that 74% of forensic clinicians in a large international sample reported using at least one structured assessment tool in their most recent assessments—meaning the remaining 26% used clinical judgment alone. The 434 clinicians in the sample reported using a surprising total of 286 different tools, many with unknown reliability or validity. Furthermore, the sources of information clinicians reported using (e.g., medical records, justice records, educational records, collateral interviews, psychological testing) varied widely even within a particular type of evaluation. This diversity of methods—including the variety and at times total lack of structured tools—is likely a major contributor to disagreement among forensic evaluators.

Even within the category of structured tools, research shows that forensic assessment instruments with explicit scoring rules based on objective criteria yield higher field reliability than instruments involving more holistic or subjective judgments. C. S. Miller and colleagues (2012) found that more structured risk assessment instruments such as the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012) and Minnesota Sex Offender Screening Tool—Revised (MnSOST-R; Epperson et al., 1998) showed higher field reliability (ICC_1 = .78 and .74, respectively) than less structured instruments like the Psychopathy Checklist—Revised (PCL-R; Hare, 2003), which showed an ICC_1 of .60 in the field. Even within the PCL-R, more objective items with explicit scoring rules (e.g., criminal versatility, juvenile delinquency, revocation of conditional release; ICC_{A1} = .75–.80) tend to show greater field reliability than more subjective items requiring impressionistic judgments (e.g., impulsivity, glibness, callousness; ICC_{A1} = .23–.36; Sturup et al., 2014).

Individual Evaluator Differences

In addition to evaluators' inconsistent training and methods, patterns of stable individual differences among evaluators—as opposed to mere inaccuracy or random variation—seem to contribute to divergent forensic opinions. For example, evaluators appear to vary widely in the base rates at which they find defendants incompetent or insane, even when all evaluators in the sample assessed defendants drawn essentially at random from the same population (Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008; Murrie & Warren, 2005). For example, in a sample of 59 clinicians conducting a total of 4,498 evaluations of legal sanity, seven clinicians found zero defendants insane while three clinicians found 50% of all defendants insane (Murrie & Warren, 2005). Similarly, some evaluators assign consistently higher or lower PCL-R scores than others, even when there are no obvious differences among examinees that might explain these scoring trends (Boccaccini, Turner, & Murrie, 2008). Stable patterns of differences suggest that evaluators may adopt idiosyncratic decision thresholds that consistently shift their forensic opinions or instrument scores in a particular direction, especially when faced with ambiguous cases (Mossman, 2013).

Two factors that may contribute to evaluators' different decision thresholds are evaluator personality and evaluator attitudes. Regarding personality, A. K. Miller and colleagues (2011) demonstrated that evaluators who described themselves as more agreeable on a personality questionnaire rated offenders as less psychopathic on PCL-R items assessing glibness, grandiosity, conning, and pathological lying (zero-order correlation = $-.51$, $p < .01$). The authors concluded that, since agreeable people tend to assume the best about others, evaluators higher

on agreeableness may have been less willing to assume that equivocal data from the case files indicated psychopathy.

Regarding attitudes, early studies found that evaluators' personal attitudes toward the insanity defense predicted whether they reached an insanity opinion in case vignettes (Homant & Kennedy, 1987). Vignette-based research and practitioner surveys have found that evaluators with pro-death-penalty attitudes are more likely to find hypothetical defendants competent for execution (Palker-Corell, 2007) or accept referrals for death penalty evaluations (Neal, 2016). Furthermore, evaluators themselves appear to acknowledge the potential influence of attitudes on their forensic work. In a recent qualitative study, many forensic evaluators identified preexisting personal, moral, or political values as influences on their forensic opinions (Neal & Brodsky, 2016).

Forensic Psychologists' Lack of Independence From the Retaining Party

Upon these concerns about unknown or less-than-ideal field reliability of forensic psychology procedures, we now add concerns about forensic experts' lack of independence from those requesting their services (NRC, 2009). As far back as the 1800s, legal experts have lamented the apparent frequency of scientific experts espousing the views of the side that hired them (perhaps for financial gain), leading one judge to comment, "[T]he vicious method of the Law, which permits and requires each of the opposing parties to summon the witnesses on the party's own account[,] . . . naturally makes the witness himself a partisan" (Wigmore, 1924). More modern surveys continue to identify partisan bias as judges' main concern about expert testimony, citing experts who appear to "abandon objectivity" and "become advocates" for the retaining party (Krafka, Dunn, Johnson, Cecil, & Miletich, 2002, p. 328).

Research on forensic psychologists working within adversarial settings appears to validate some of these concerns about adversarial allegiance, the tendency for experts to reach conclusions that support the party who retained them (Murrie et al., 2009). Some early studies suggested that clinicians drifted toward opinions favorable to the retaining party in a real-life civil litigation following a mining disaster (Zusman & Simon, 1983) and in case vignettes simulating sanity evaluations (Otto, 1989). More recently, using scores from structured risk instruments (e.g., PCL-R, Static-99R) as a convenient way to quantify differences in expert opinion, researchers examining archival data found large scoring differences according to side of retention—prosecution-retained evaluators produced higher risk scores that made the examinee look more dangerous, while defense-retained evaluators produced lower risk scores that made the examinee look more benign. For example, Murrie et al. (2009) found an average difference of 5.8 points on the PCL-R (score range: 0–40) between opposing sexually violent predator (SVP) evaluators in Texas, a difference twice the standard error of measurement reported in the test manual (Hare, 2003).[3]

Recent surveys also suggest that evaluators tend to interpret risk scores in a way that favors the side that retained them (Boccaccini, Chevalier, Murrie, & Varela, 2015; Chevalier, Boccaccini, Murrie, & Varela, 2015). For example, Chevalier et al. (2015) found that 94% of state-retained SVP evaluators reported using high-risk/need norms for the Static-99R (a way of interpreting scores that makes the examinee seem more risky, as compared to routine sample norms), but only 33% of respondent-retained evaluators reported using high-risk/need norms. Thus, two opposing evaluators who arrive at the same numerical score on a risk assessment instrument might still draw biased conclusions that favor the retaining side through differing norm selection.

These surveys and field studies of adversarial allegiance cannot rule out of the possibility of selection effects creating the observed scoring differences (Murrie & Boccaccini, 2015). Attorneys may preselect evaluators whom they know to be sympathetic to their point of view, or gather preliminary opinions from multiple evaluators and ultimately retain only the most favorable opinion. Furthermore, evaluators may self-select according to preexisting attitudes or preferences, choosing to accept or decline particular types of cases or cases from particular referral sources (Neal, 2016). To eliminate the possible influence of selection ef-

[3] *SVP* refers to sexually violent predator provisions, which allow for sexual offenders to be civilly committed after completing their criminal sentence. While SVP proceedings are technically civil, they still involve an adversarial arrangement, with different forensic psychologists testifying for the state and for the respondent (i.e., the individual being considered for commitment).

fects, Murrie and colleagues (2013) conducted an experiment where practicing forensic evaluators were randomly assigned to believe they were working for the prosecution or the defense on a real-world case consultation. Even with random assignment, evaluators still tended to score cases in the direction of allegiance. Unsurprisingly, allegiance effects were larger for the PCL-R (medium to large effect sizes; $d = 0.55–0.85$) than for the more structured and objective Static-99R (small effect sizes; $d = .20–.42$).[4] While the Murrie et al. (2013) experiment used sex offender case files scored with popular risk assessment instruments, other types of forensic evaluations and instruments likely show the same vulnerability to adversarial allegiance.

Future Directions for Research, Practice, and Policy

The research overviewed here points to the growing realization that some portion of every forensic opinion—perhaps a larger portion than we might now acknowledge—has more to do with the examiner than the examinee. This is a serious problem that risks arbitrary or unjust outcomes for those undergoing forensic evaluations, as well as diminishing the legal system's confidence in psychological expertise. Unreliable evaluations can also put the community at risk (e.g., assigning a low risk score to a truly high-risk individual likely to offend again). At the same time, some degree of unreliability and bias on complex human decision tasks is unavoidable in light of our "bounded rationality" (Gigerenzer & Goldstein, 1996). Given this tension, what next steps are possible to prevent forensic psychology from becoming the NRC or PCAST's next target?

Just as research has helped uncover these problems, further research can continue to define the scope of the problem and suggest solutions. As a much-needed first step, foundational research should establish field reliability rates for various types of forensic evaluations in order to assess the current situation and gauge progress toward improvement. Only a handful of field reliability studies exist for a few types of forensic evaluations (i.e., adjudicative competency, legal sanity, conditional release), and virtually nothing is known about the field reliability of other types of evaluations, particularly civil evaluations. If error rates of forensic psychology procedures were widely known, legal decision makers might be able to weight their confidence in psychological testimony according to the reliability of the procedure in question (Butler, 2013). In addition, by carefully cataloguing variables specific to the examiner, examinee, and evaluation context from which reliability figures are drawn, field reliability research can also shed light on factors associated with better or worse reliability, suggesting further avenues for improvement (Guarnera & Murrie, in press).

Given that increased standardization of forensic methods has the potential to ameliorate multiple sources of unreliability and bias described here, more investigation of forensic instruments, checklists, practice guidelines, and other methods of standardization is a second research priority (Ægisdóttir et al., 2006). Some of this research should continue to focus on creating standardized tools for forensic evaluations and populations for which none are currently available, particularly civil evaluations such as guardianship, child protection, fitness for duty, and civil torts like emotional injury (Heilbrun & Brooks, 2010). Future research can also continue to seek improvements to the currently modest predictive accuracy of risk assessment instruments (Fazel, Singh, Doll, & Grann, 2012). However, given the current gap between the availability of forensic instruments and their limited use by forensic evaluators in the field, perhaps more pressing is research on the implementation of forensic instruments in routine practice. More qualitative (e.g., Pinals, Tillbrook, & Mumley, 2006) and quantitative (e.g., Neal & Grisso, 2014) investigations of how instruments are administered in routine practice, why instruments are or are not used, and what practical obstacles evaluators encounter are needed. Without greater understanding of how instruments are (or are not) implemented in practice—particularly in rural or other underresourced areas—continuing to develop new tools may not translate to their increased use in the field.

Third, a clear recommendation for improving evaluator reliability is that states without stan-

[4] These effect sizes held true for three out of four cases included in the study. One case, involving an individual with exceptionally low risk, did not show evidence of adversarial allegiance. All evaluators rated this individual as similarly low risk, regardless of side of retention.

dards for the training and certification of forensic experts should adopt them, and states with weak standards (e.g., mere workshop attendance) should strengthen them. What is less clear, however, is what kinds and doses of training can improve reliability with the greatest efficiency. Drawing from extensive research in industrial and organizational psychology, credentialing requirements that mimic the type of work evaluators do as part of their job (e.g., mock reports, peer review, apprenticing) may foster professional competency better than requirements dissimilar to job duties (e.g., written tests; Phillips, 1998). Given that both evaluators and certifying bodies have limited time and resources, research into the most potent ingredients of successful forensic credentialing is a third research priority.

Even while this important research remains to be done, practicing forensic evaluators still have many options to reduce the impact of unreliability and bias in their own work. While many clinicians cite introspection (i.e., looking inward in order to identify one's own biases) as a primary method to counteract personal ideology, idiosyncratic responses to examinees, and other individual differences (Neal & Brodsky, 2016), research suggests that introspection is ineffective and may even be counterproductive (Pronin, Lin, & Ross, 2002). Thus, more disciplined changes to personal practice are needed. For example, when conducting evaluations for which well-validated structured tools exist, evaluators could commit to using such tools as a personal standard of practice. This would entail justifying to themselves (or preferably colleagues) why they did or did not use an available tool for a particular case. Practicing forensic evaluators could also use simple debiasing methods to counteract confirmation bias, such as the "consider-the-opposite" technique in which evaluators ask themselves, "What are some reasons my initial judgment might be wrong?" (Mussweiler, Strack, & Pfeiffer, 2000). To increase personal accountability, evaluators could keep organized records of their own forensic opinions and instrument scores, or even help organize larger databases for evaluators within their own institution or locality (Lerner & Tetlock, 1999). Using these personal data sets, evaluators might look for mean differences in their own instrument scores when retained by the prosecution versus the defense, or compare their own base rates of incompetency and insanity findings to those of their colleagues.

Ambitious evaluators could even experiment with blinding themselves to the source of referral in order to counteract adversarial allegiance (Robertson & Kesselheim, 2016). For example, evaluators could try using a case manager, an individual who communicates with attorneys and controls the inflow and outflow of information, in order to prevent irrelevant biasing information (such as the identity of the retaining party) from reaching the evaluator (Dror, 2013). Evaluators may soon be able to market (to attorneys or the court) their willingness to serve as blinded experts, since research suggests that mock jurors view the testimony of blinded experts as more credible (Robertson & Yokum, 2012).

Although individual evaluators can make many voluntary changes today in order to reduce the impact of unreliability and bias on their forensic opinions, other reforms require wider-ranging structural transformation. For example, state-level legislative action is needed to mandate more than one independent forensic opinion. Requiring more than one independent opinion is a powerful way to combat unreliability and bias by reducing the impact of any one evaluator's error (Larrick, 2004). For example, by statute, Hawaii mandates three independent, nonadversarial forensic opinions for all felony defendants being evaluated for adjudicative competency and legal sanity (Hawaii Revised Statutes, 2003, sections 704–404 and 704–406). Only nine other states require more than one competency evaluator, and 14 states allow (but do not require) more than one evaluator (Gowensmith et al., 2015). For more states to join these ranks, state legislators would need to prioritize funding for multiple forensic evaluations per defendant, likely a substantial outlay. Similarly, more stringent state-level certification standards would require considerable financial investment in the infrastructure necessary to organize trainings, vet certification materials, maintain records, and enforce compliance.

Even slower to change than state legislation and infrastructure might be existing legal norms, such as judges' current willingness to admit nonblinded, partisan experts. While authoritative calls to action like the NRC and

PCAST reports may have some influence, most legal change only happens by the accretion of legal precedent, which is a slow and unpredictable process. Thus, radical changes regarding the roles and expectations of forensic experts—such as "hot tubbing," a system pioneered in Australia where opposing experts are questioned simultaneously and can also question each other (Edmund, 2009)—seem unlikely to take root any time soon in the American legal system. Regardless, we hope the growing awareness of problems of unreliability and bias in the forensic sciences—in the wake of the NRC and PCAST reports—can spur on legal reforms, as well as create urgency to prioritize some of these larger structural and funding changes within forensic psychology.

References

Acklin, M. W., Fuger, K., & Gowensmith, W. N. (2015). Examiner agreement and judicial consensus in forensic mental health evaluations. *Journal of Forensic Psychology Practice, 15,* 318–343. http://dx.doi.org/10.1080/15228932.2015.1051447

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34,* 341–382. http://dx.doi.org/10.1177/0011000005285875

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68,* 7–19. http://dx.doi.org/10.1037/a0029889

Boccaccini, M. T., Chevalier, C. S., Murrie, D. C., & Varela, J. G. (2015). Psychopathy Checklist–Revised use and reporting practices in sexually violent predator evaluations. *Sexual Abuse*. Advanced online publication. http://dx.doi.org/10.1177/1079063215612443

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283. http://dx.doi.org/10.1037/a0014523

Butler, H. A. (2013). *Debiasing juror perceptions of the infallibility of forensic identification evidence: The utility of educational and perspective-taking debiasing methods* (Unpublished doctoral dissertation). Claremont Graduate University, Claremont, CA.

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior, 39,* 209–218. http://dx.doi.org/10.1037/lhb0000114

DeMatteo, D., Marczyk, G., Krauss, D. A., & Burl, J. (2009). Educational and training models in forensic psychology. *Training and Education in Professional Psychology, 3,* 184–191. http://dx.doi.org/10.1037/a0014582

Dror, I. E. (2013). Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management, 4,* 105–113. http://dx.doi.org/10.1080/19409044.2014.901437

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice, 51,* 204–208. http://dx.doi.org/10.1016/j.scijus.2011.08.004

Dror, I., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences, 53,* 900–903. http://dx.doi.org/10.1111/j.1556-4029.2008.00762.x

Edmund, G. (2009). Merton and the hot tub: Scientific conventions and expert evidence in Australian civil procedure. *Law and Contemporary Problems, 72,* 159–189. http://www.jstor.org/stable/40647170

Epperson, D. L., Kaul, J. D., Goldman, R., Hout, S. J., Hesselton, D., & Alexander, W. (1998). *Minnesota Sex Offender Screening Tool—Revised (MnSOST-R)*. St. Paul, MN: Minnesota Department of Corrections.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal, 345,* e4692. http://dx.doi.org/10.1136/bmj.e4692

Fuger, K. D., Acklin, M. W., Nguyen, A. H., Ignacio, L. A., & Gowensmith, W. N. (2014). Quality of criminal responsibility reports submitted to the Hawaii judiciary. *International Journal of Law and Psychiatry, 37,* 272–280. http://dx.doi.org/10.1016/j.ijlp.2013.11.020

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gowensmith, W. N., Pinals, D. A., & Karas, A. C. (2015). States' standards for training and certifying evaluators of competency to stand trial. *Journal of Forensic Psychology Practice, 15,* 295–317. http://dx.doi.org/10.1080/15228932.2015.1046798

Gowensmith, W. N., Sledd, M., & Sessarego, S. (2014). *The impact of stringent certification standards on forensic evaluator reliability*. Paper pre-

sented at the annual meeting of the American Psychological Association, Washington, DC.

Grisso, T. (1987). The economic and scientific future of forensic psychological assessment. *American Psychologist, 42,* 831–839. http://dx.doi.org/10.1037/0003-066X.42.9.831

Guarnera, L. G., & Murrie, D. C. (in press). Field reliability of competency and sanity opinions: A systematic review and meta-analysis. *Psychological Assessment*.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics.

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Hawaii Revised Statutes, Vol. 14, 704–404 (2003). http://dx.doi.org/10.1007/BF01044699

Heilbrun, K., & Brooks, S. (2010). Forensic psychology and forensic science: A proposed agenda for the next decade. *Psychology, Public Policy, and Law, 16,* 219–253. http://dx.doi.org/10.1037/a0019138

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, 24,* 64–101.

Homant, R. J., & Kennedy, D. B. (1987). Subjective factors in clinicians' judgments of insanity: Comparison of a hypothetical case and an actual case. *Professional Psychology: Research and Practice, 18,* 439–446. http://dx.doi.org/10.1037/0735-7028.18.5.439

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,* 309–332. http://dx.doi.org/10.1037/1076-8971.8.3.309

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. http://dx.doi.org/10.2307/2529310

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–338). Oxford, UK: Blackwell. http://dx.doi.org/10.1002/9780470752937.ch16

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255–275. http://dx.doi.org/10.1037/0033-2909.125.2.255

Meyer, G. J., Mihura, J. L., & Smith, B. L. (2005). The interclinician reliability of Rorschach interpretation in four data sets. *Journal of Personality Assessment, 84,* 296–314. http://dx.doi.org/10.1207/s15327752jpa8403_09

Miller, A. K., Rufino, K. A., Boccaccini, M. T., Jackson, R. L., & Murrie, D. C. (2011). On individual differences in person perception: Raters' personality traits relate to their Psychopathy Checklist—Revised scoring tendencies. *Assessment, 18,* 253–260. http://dx.doi.org/10.1177/1073191111402460

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24,* 944–953. http://dx.doi.org/10.1037/a0028411

Mossman, D. (2013). When forensic examiners disagree: Bias, or just inaccuracy? *Psychology, Public Policy, and Law, 19,* 40–55. http://dx.doi.org/10.1037/a0029242

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among forensic experts. *Annual Review of Law and Social Science, 11,* 37–55. http://dx.doi.org/10.1146/annurev-lawsocsci-120814-121714

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24,* 1889–1897. http://dx.doi.org/10.1177/0956797613481812

Murrie, D. C., Boccaccini, M. T., Turner, D. B., 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. http://dx.doi.org/10.1037/a0014897

Murrie, D. C., Boccaccini, M. T., Zapf, P. A., Warren, J. I., & Henderson, C. E. (2008). Clinician variation in findings of competence to stand trial. *Psychology, Public Policy, and Law, 14,* 177–193. http://dx.doi.org/10.1037/a0013578

Murrie, D. C., & Warren, J. I. (2005). Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice, 36,* 519–524. http://dx.doi.org/10.1037/0735-7028.36.5.519

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26,* 1142–1150. http://dx.doi.org/10.1177/01461672002611010

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: National Academies Press. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

Neal, T. M. (2016). Are forensic experts already biased before adversarial legal parties hire them? *PLoS ONE, 11,* e0154434. http://dx.doi.org/10.1371/journal.pone.0154434

Neal, T., & Brodsky, S. L. (2016). Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. *Psychology, Public Policy, and Law, 22,* 58–76. http://dx.doi.org/10.1037/law0000077

Neal, T., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior, 41,* 1406–1421. http://dx.doi.org/10.1177/0093854814548449

Otto, R. K. (1989). Bias and expert testimony of mental health professionals in adversarial proceedings: A preliminary investigation. *Behavioral Sciences & the Law, 7,* 267–273. http://dx.doi.org/10.1002/bsl.2370070210

Palker-Corell, A. M. (2007). *Mental health professionals' decision-making in competence for execution evaluations* (Unpublished doctoral dissertation). Sam Houston State University, Huntsville, TX.

Phillips, J. M. (1998). Effects of realistic job previews on multiple organizational outcomes: A meta-analysis. *Academy of Management Journal, 41,* 673–690. http://dx.doi.org/10.2307/256964

Pinals, D. A., Tillbrook, C. E., & Mumley, D. L. (2006). Practical application of the MacArthur competence assessment tool-criminal adjudication (MacCAT-CA) in a public sector forensic setting. *Journal of the American Academy of Psychiatry and the Law, 34,* 179–188.

President's Council of Advisors on Science and Technology (PCAST). (2016). *Report to the President: Forensic science in the criminal courts: Ensuring scientific validity of feature-comparison methods*. Washington, DC: Executive Office of the President of the United States. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perception of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369–381. http://dx.doi.org/10.1177/0146167202286008

Robertson, C. T., & Kesselheim, A. S. (2016). *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. San Diego, CA: Elsevier.

Robertson, C. T., & Yokum, D. V. (2012). The effect of blinded experts on juror verdicts. *Journal of Empirical Legal Studies, 9,* 765–794. http://dx.doi.org/10.1111/j.1740-1461.2012.01273.x

Skeem, J. L., Golding, S. L., Cohn, N. B., & Berge, G. (1998). Logic and reliability of evaluations of competence to stand trial. *Law and Human Behavior, 22,* 519–547. http://dx.doi.org/10.1023/A:1025787429972

Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist—Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38,* 315–324. http://dx.doi.org/10.1037/lhb0000063

Wigmore, J. H. (1924). To abolish partisanship of expert witnesses, as illustrated in the Loeb-Leopold case. *Journal of the American Institute of Criminal Law and Criminology, 15,* 341–343.

Wood, J. M., Nezworski, T., & Stejskal, W. J. (1996). The comprehensive system for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00658.x

Zusman, J., & Simon, J. (1983). Differences in repeated psychiatric examinations of litigants to a lawsuit. *American Journal of Psychiatry, 140,* 1300–1304. http://dx.doi.org/10.1176/ajp.140.10.1300

Received April 2, 2016
Revision received March 7, 2017
Accepted March 24, 2017 ■