Psychology, Public Policy, and Law
2018, Vol. 24, No. 1, 11–23

© 2017 American Psychological Association
1076-8971/18/$12.00 http://dx.doi.org/10.1037/law0000140

A Hierarchy of Expert Performance Applied to Forensic Psychological Assessments

Itiel E. Dror
University College London

Daniel C. Murrie
University of Virginia

Experts in forensic psychology must make skilled observations and conclusions, minimally compromised by bias, in order to try and provide reliable and accurate conclusions to the courts. But the field has little data revealing how well forensic psychologists actually perform these tasks, in part because there has been no clear framework for systematic research of their expertise. Therefore, we consider forensic psychological assessments in light of Dror's (2016) Hierarchy of Expert Performance (HEP). HEP addresses reliability and biasability, both within and between experts, at the levels of observations and conclusions. Applying this framework to forensic psychological assessments reveals a few domains in which there are some meaningful data, particularly addressing reliability between experts in certain types of forensic assessments. But applying HEP reveals more domains in which we lack data addressing fundamental aspects of expert performance, such as reliability at the level of observations, and reliability and biasability *within* experts. Understanding these strengths and gaps in forensic assessment research should guide testimony of forensic psychologists, policies around forensic assessment, and further research in forensic assessment.

Keywords: bias, reliability, adversarial allegiance, contextual effects, forensic assessment

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Expert performance can be characterized and measured in a variety of ways. When it comes to decision making, a basic measurement is accuracy. In other words, are expert decisions correct, and do they reflect the ground truth? But there are many domains—including many areas in forensic science and forensic psychology—wherein the "real" answer is never known. For example, who actually committed the crime may be unknown, or the defendant's actual mental state at the time of the crime may be unknown. Because ground truth is unknown in real cases, we cannot always directly measure the accuracy of forensic experts. Given this problem, controlled research is needed to help ascertain accuracy less directly, by determining the components involved in forensic decision making. Then we can experimentally isolate and measure them, with a view that such understandings will reveal strengths and weaknesses, and then inform strategies and policies to improve the work of forensic experts.

A Hierarchy of Expert Performance

Two basic properties of decision making are biasability and reliability (Dror, 2016; Dror & Rosenthal, 2008). *Biasability*, within the forensic science and legal community, refers to the potential effects of irrelevant contextual information and other biases that may impact the decision. For example, would a criminal defendant's race influence forensic evaluators' decisions (as it seems to among other decision-makers in the justice system; Mitchell, Haw, Pfeifer, & Meissner, 2005; Smalarz, Madon, Tang, Guyll, & Buck, 2016)? *Reliability* refers to the consistency, reproducibility, or repeatability of decisions, regardless of bias. For example, would different forensic evaluators reach the same conclusions about a defendant's legal sanity when they review and examine the same collateral records and recorded interview of the defendant? Reliability and biasability are distinct concepts, but both contribute to variability in decision making.

Without considering and teasing apart the different elements underpinning expert performance, such as reliability and biasability, it is hard to properly quantify expert performance, particularly because there are no parameters to research. Organizing different elements in expert performance enables us to understand the different aspects of expert performance and how they relate to one another. Reliability and biasability are often lumped together, and therefore variability in decisions may not be correctly attributed to the biasing effects or to the reliability per se. Furthermore, teasing apart the different components of decision making allows us to identify gaps in the literature and to prescribe further research, as well as to direct and focus policies on specific problems, where needed. For example, policies need not address basic reliability (e.g., by requiring multiple assessments) if biasability is the major contributor to

This article was published Online First September 21, 2017.

Itiel E. Dror, University College London; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia.

We thank John Monahan for his helpful comments on a draft of this article. Some content of this article was presented at the annual conference of the American Psychology-Law Society (Seattle, Washington), the Risk and Recovery Conference (Hamilton, Ontario), and the British Psychological Society annual forensic conference (Bristol, United Kingdom), all in 2017. Each author made equal contributions to this article, so authorship order was determined by coin flip.

Correspondence concerning this article should be addressed to Daniel C. Murrie, Institute of Law, Psychiatry, & Public Policy, University of Virginia Box 800660, Charlottesville, VA 22908-0660. E-mail: murrie@virginia.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the variance; conversely, policies need not address biasability (e.g., by requiring a neutral, court-appointed expert) if simple unreliability is the underlying problem.

A second distinction, in addition to reliability and biasability, is to quantify experts' performance relative to other experts (*between experts*, or interexpert performance) versus experts' performance relative to themselves (*within experts*, or intraexpert performance). For example, a between-expert study might examine whether several different fingerprint examiners identify the same features in the same fingerprint mark, whereas a within-expert performance study might examine whether the same fingerprint examiner will identify the same features in the same fingerprint when it is presented multiple times at different occasions (e.g., Dror, et al., 2011).

Within-expert performance is perhaps the most basic and essential level of expertise. When experts vary in their performance among one another (between-experts), this variability can be attributed to individual differences (e.g., different philosophies and ideologies, different training and experience, different subjective decision thresholds, different eyesight, different risk tolerance, and a variety of factors that make experts different from one another). However, if an individual expert cannot be consistent with himself—that is, if he or she cannot draw the same observations and conclusions from the same data—this unreliability cannot be attributed to individual differences. Thus within, intraexpert, performance measurements are a more basic metric, and foundational to expert performance.

A third distinction in studying expert decision making is the distinction between observations and conclusions. Conclusions depend on assessment and interpretation of observations. Therefore, to understand decisions, one must be able to distinguish performance in interpreting observations (i.e., drawing conclusions) versus performance in actually *making* the original observations (Dror, 2016). Lumping these together obscures the initial observational performance and may be misleading because observations underpin the resulting conclusions.

In the medical domain the distinction between observation and conclusion is made clear and explicit in the SBAR (Situation, Background, Assessment, and Recommendation) protocol (e.g., Thomas, Bertram, & Johnson, 2009; Wacogne & Diwakar, 2010). The Situation and Background focuses on observations (such as, patient heart rate is 140, patient has a history of heart attacks, etc.), whereas the Assessment and Recommendation focuses on the conclusions based on the observations (such as, patient is having a heart attack, patient should be provided with oxygen, etc.). Similarly, the "SOAP" method (Subjective, Objective, Assessment, Plan) guides clinicians to work from observations to conclusions (Weed, 1970). "Subjective" and "Objective" refer to observations about the patient's presentation, whereas "Assessment" and "Plan" are conclusions based on the initial observations.

Research from the forensic sciences illustrates the crucial distinction between observations and conclusions. For example, imagine two fingerprint experts reach different decisions: one expert concludes with high confidence that the two prints 'match' (i.e., come from the same source) whereas the other expert concludes with high confidence that they do not match. It may be that both examiners observed the same data in the fingerprints (the minutia; see Figure 1), but nevertheless reached opposing conclusions because they use different similarity thresholds for concluding a 'match' (e.g., one examiner requires a minimum of 12 matching points, whereas the other requires at least 13). Alternatively, it may be that the two examiners used identical thresholds to reach a conclusion of a 'match' (e.g., 12), but reached different decisions because they observed different data in the fingerprints (e.g., one examiner observed the 12 minutia needed to call a 'match,' whereas the other only observed 9) (Dror et al., 2011). Understanding this distinction is crucial for designing interventions to increase reliability; for example, do the experts need better training in observing fingerprint minutia, better training in drawing conclusions, or both?

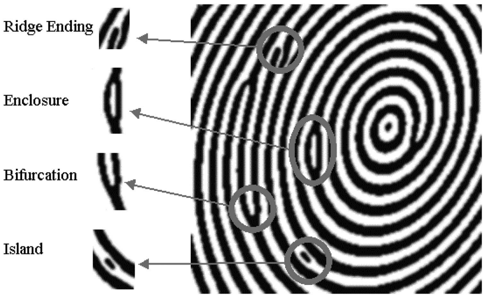

Figure 1. Different characteristics (minutia) that may be present in a fingermark.

This between-expert (interexpert) example illustrates why it is important to tease apart observations from conclusions, but the distinction can equally apply to within-expert (intraexpert) performance. If the same expert makes a different decision when the same fingerprints are presented multiple times at different occasions, this can happen because the expert used different decision rules at the different times, or it can be because the expert observed different data at the different times. Here, when the variance is within-expert, the interventions will be different than those for between-expert variance (hence the importance of distinguishing between- and within-expert variance); for example, has change in eyesight underpinned observing different minutia (and if so, shall there be a policy for yearly eyesight testing of fingerprint examiners)? Or, do changing work environment and pressures need addressing because they underpin variance in drawing conclusions? Of course, it can be that both observation and conclusion differences contribute to the differences in decisions, and the distinction is not always clear. Nevertheless, it is important to tease them apart as much as possible.

Using these three dimensions of: a) reliability and biasability; b) within experts and between experts; and c) observations and conclusions, Dror (2016) suggested an 8-level Hierarchy of Expert Performance (HEP). As illustrated in Figure 2, at the bottom of the HEP is the most basic measurement of expert performance: reliability within expert observation; that is, how consistent an expert is with herself in what she observes. The question and quantification at Level 1 is the extent to which an expert will observe the same things when presented with the same data. For example, fingerprint experts, presented with the same print at Time 1 and Time 2, will often observe different features (see Dror et al., 2011, for details). Level 2 of HEP remains in the observation stage and still focuses on reliability, but at this level the measurement is differences in performance between experts, rather than within

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

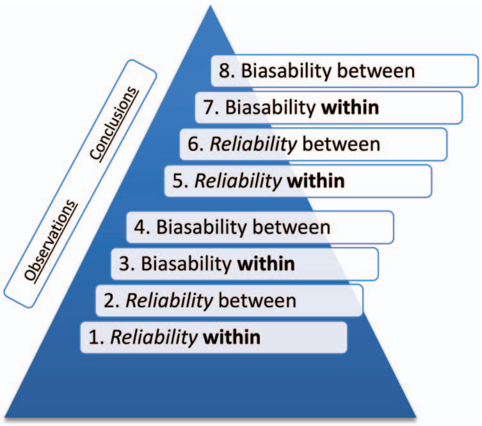

Figure 2. Dror's (2016) HEP: Hierarchy of Expert Performance. See the online article for the color version of this figure.

experts. Levels 3 and 4 examine biasability in observations—Level 3 within experts and Level 4 between experts. These levels address the impact of irrelevant contextual information on observations of data. For example, will irrelevant context (such as the nature of the crime or the suspect) influence what experts observe in the fingerprint evidence (e.g., Earwaker, Morgan, Harris, & Hall, 2015)?

While Levels 1–4 focus on observations, Levels 5–8 of HEP focus on the conclusions. Following the structure of Levels 1–4, Levels 5–8 address quantification of reliability within experts (Level 5), reliability between experts (Level 6), biasability within experts (Level 7), and biasability between experts (Level 8).

Expert Performance in Forensic Psychology and Psychiatry: Identifying Research Needs and Policy Implications

Like other forensic science practitioners, forensic psychologists and psychiatrists are trained experts, often assigned to review complex case materials, interview criminal defendants or civil litigants, and provide expert opinions to assist the court. Also like other forensic science experts, their services usually culminate in formal written reports or expert testimony in which they present conclusions to assist a judge or jury in reaching a verdict. To the extent that their conclusions are reliable and resistant to bias, the justice system can have greater faith in them. But to the extent that these conclusions are unreliable or biased, they are at best unhelpful, and at worst misleading, to the justice system's goals of administering justice with accuracy and equity.

Despite their similarities to other forensic scientists, forensic psychologists and psychiatrists have escaped much of the public and government scrutiny that other forensic science domains have received. The National Research Council (NRC, 2009) and the 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, tool marks, bite marks, and firearms. Both government councils concluded that the reliability of many forensic techniques is unknown and that forensic scientists are prone to a variety of contextual biases.

Wide-scale calls for reform prompted vigorous national efforts, such the National Institute of Standards and Technology's (NIST) efforts to develop best practices and standards in the forensic sciences, and the formation of a National Commission on Forensic Science to develop policies. Both entities include specialized Human Factors groups to address human decision making and bias, and both have produced a wide range of policies to increase reliability and reduce bias in the forensic sciences (e.g., NCFS, 2015). NIST has even provided substantial funding to improve the scientific foundations of forensic science (National Institute of Standards and Technology, 2016). *None* of the recent wide-scale reviews or reform efforts have addressed forensic psychology or psychiatry. But regardless of whether authorities expand their scrutiny (and funding) of forensic sciences to include forensic psychology, their efforts evoke similar questions about forensic psychology and psychiatry (Guarnera, Murrie, & Boccaccini, 2017; Heilbrun & Brooks, 2010; Murrie, Boccaccini, Guarnera, & Rufino, 2013). Likewise, all relevant ethical and professional standards (e.g., AERA, APA, & NCME, 2014; APA, 2002, 2013) suggest the field has a duty to examine and optimize the reliability and objectivity of our methods, regardless of external scrutiny.

Just as Dror's (2016) hierarchy of expert performance (HEP) provided both theoretical and practical benefits to understanding expert performance in the forensic sciences, applying this HEP to forensic mental health evaluations allows us to better understand the expertise and performance of forensic psychologists and psychiatrists. It should help identify areas of strength and weakness, as well as areas in which we simply lack adequate data about evaluators' expert decision making. Following Dror's (2016) HEP in forensic science, we will work from the top of the hierarchy (Level 8) to the bottom (Level 1), beginning with the considerations that appear most visible (e.g., biasability between expert conclusions) and working down to the most basic questions of reliability in observations within experts, which have been least observed or researched. A primary goal in this paper is to identify where empirical data is lacking—that is, where we know little about the performance of forensic psychologists providing expert services to the justice system—and prescribe research that will address these gaps and ultimately improve expert decision making.

In performing this literature review, we used several strategies. We performed standard searches of online research databases using many variations of terms related to bias, reliability, and forensic psychology. We reviewed authoritative texts (e.g., Melton, Petrila, Poythress, & Slobogin, 2007; Packer, 2009; Zapf & Roesch, 2009) and meta-analyses (e.g., Guarnera & Murrie, 2017) addressing forensic evaluation. Upon locating appropriate studies, we reviewed reference lists and ran "cited by" searches to seek additional relevant studies. Finally, we distributed drafts of this manuscript to authorities in forensic psychology—particularly authorities in the subdisciplines with which we were least familiar—asking that they notify us of any potentially relevant research we may have missed.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Dror's (2016) Hierarchy of Expert Performance Applied to Forensic Psychology

8. Biasability Between Experts' Conclusions

Are different experts, considering identical data, biased by irrelevant contextual information? In forensic science, irrelevant contextual information may include whether a suspect confessed to the crime, whether other lines of evidence suggest he is the culprit, and so forth. Researchers have identified the biasing effect of contextually irrelevant information in several ways. For example, Dror and Hampikian (2011) examined whether irrelevant contextual information that implicated a suspect in a sexual assault biased the conclusions of DNA experts. The examiners who were exposed to the biasing information concluded that he could not be excluded from being a contributor to the DNA mixture, whereas most examiners (16 out of 17) who were not exposed to the biasing information did not reach the same conclusion.

In forensic psychology and psychiatry, experts likely encounter a variety of irrelevant contextual information (e.g., case details that are irrelevant and go beyond the specific referral question for forensic evaluation and beyond the expertise of the evaluator). The field has certainly not considered or identified task-irrelevant information in the ways that other forensic sciences have (see, e.g., NCFS, 2015).[1] However, one piece of contextual information that is almost always irrelevant is the "side," or party, requesting an expert opinion.[2] Forensic mental health professionals are to strive for accuracy and neutrality, impartial to the side that retained them (APA, 2013).[3] But judges, legal scholars, and even laypersons have long lamented that forensic experts appear biased by the side that retained them (e.g., Foster, 1897; Hand, 1901; Wigmore, 1923). Likewise, a series of field studies strongly suggest that at least some forensic mental health experts may be vulnerable to *adversarial allegiance,* a bias toward reaching conclusions that favor the party retaining their services (Murrie & Boccaccini, 2015; Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009).

Compelling evidence of biasability between expert conclusions comes from an experimental study in which researchers recruited over 100 practicing, doctoral-level forensic psychologists and psychiatrists and led them to believe they were performing a formal, large-scale forensic consultation (Murrie et al., 2013). These forensic experts were—unbeknownst to them—randomly assigned to either believe that they were paid by the public defender service or the special prosecution unit. These participants met with an attorney who posed as leading either the public defender service or the specialized prosecution unit, and requested that they score particular risk instruments based on extensive offender records (a type of consultation not uncommon in forensic practice). Each participant scored the same four case files, and each file was authentic (i.e., from an actual case), including extensive records (e.g., police, court, correctional, mental health) typical of those evaluators use to score risk instruments in forensic evaluations. Thus, participating forensic experts were able to score the same two commonly used risk instruments that served as the metrics for bias in earlier field studies (Murrie et al., 2008, 2009): that is, the PCL-R (Hare, 2003) and the Static-99R (Helmus et al., 2012).

Overall, the risk measure scores assigned by the experts showed a clear pattern of differences (i.e., experts who thought they were working for the prosecution assigned significantly higher scores, and those who thought they were working for the defense assigned lower scores), revealing biasability between experts as a function of the side they believed retained them. Allegiance effects were stronger for the PCL-R—a measure that requires more subjective clinical judgment—than for the Static-99R, a more structured measure that permits less judgment.

These experimental results provide strong evidence that even scores on ostensibly objective forensic instruments can be compromised by bias (Murrie et al., 2013). To be clear, there was considerable variability in scores even among experts assigned to the same side, a form of poor reliability between experts (distinct from biasability), and not all experts demonstrated allegiance effects. But finding such evidence of allegiance in a context where all other possible explanations have been experimentally controlled suggests that adversarial allegiance is a significant biasing influence among some experts. Put simply, the "side" retaining an expert is one piece of biasing, but irrelevant, contextual information that can cause experts to reach different conclusions.

Beyond this one type of irrelevant biasing information, the field has little data on other types of task-irrelevant contextual information that may bias experts. But many questions about contextual bias are important for practice and policy. For example, an evaluator's role and workplace are theoretically irrelevant to a defendant's true adjudicative competence. But might a community-based evaluator reach a different decision about the competence of a volatile and disruptive defendant than the hospital-based evaluator who would also be tasked with providing treatment or competence restoration services if the defendant is found incompetent and hospitalized? One prominent judge lecturing to psychologists mused, "I wonder how many clinical evaluations that inform decisions about [civil] commitment are influenced by such extraneous considerations as the amount of bed space in the receiving institution" (Bazelon, 1982).

Other potentially biasing information could be fairly easily researched between evaluators (or even within evaluators; see

[1] We acknowledge that determining what is task-irrelevant information is not always clear in forensic science and can be even less clear in forensic psychology. For example, crime details are highly relevant when evaluating legal sanity (a defendant's mental state at the time of the crime), but usually much less relevant when evaluating a defendant's trial competence. A full discussion of this important issue is beyond the scope of this paper, though we note that some contextual information should almost always be irrelevant (e.g., the side retaining the evaluation, the fee for the evaluation, the census pressures in a state hospital, etc.), whereas some (e.g., defendant's sexual behaviors and interests, defendant's family relationships, etc.) may be relevant only to some referral questions and not others.

[2] To be clear, there are a few contexts in which a forensic evaluator's role may properly differ depending on the side retaining that evaluator. For example, Virginia statute directs the mitigation evaluator in a capital sentencing evaluation to actively *assist the defense* in presenting mitigation evidence (Virginia Code § 19.2-264.3:1). But in most situations, the evaluation task, scope, or conclusions should not differ based on the party requesting services.

[3] In the U.S., the Federal Rule of Evidence 702 requirement that "the expert has reliably applied the principles and methods to the facts of the case" suggests an expectation of objectivity. More than 40 states have adopted this federal rule in their state evidence codes. In the United Kingdom, The Criminal Procedures Rule 33.2(1)(a) clearly and explicitly states that the expert's duty to the court must be "objective and unbiased;" bias toward the retaining party is a clear violation.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

below) using popular research strategies such as vignettes or case descriptions. For example, would a psychologist reach the same conclusions about a defendant's trial competence if that defendant was described as amiable and nonviolent versus psychopathic and pedophilic (paraphilias and personality disorders are usually irrelevant to trial competence)? Could the race, ethnicity, religion, or sexual orientation of the defendant impact evaluator conclusions if all other case details were identical (see Smalarz et al., 2016 for a similar study within forensic science evidence)? Studies of between-expert biasability may be among the easiest to explore with common social science research methods.

7. Biasability Within Experts' Conclusions

Would the *same* expert reach the same (or different) conclusions when an identical case is presented within a different, irrelevant, biasing context? Whereas Level 8 in the HEP (see Figure 2) deals with differences *between* experts, Level 7 deals with intraexpert, *within*-expert biasability in conclusions. Research in the forensic sciences provides examples of biasability within the same expert's conclusions. For example, the same fingerprint experts did not always reach the same conclusions when the same fingerprints were presented to them on different occasions within different irrelevant contexts (Dror & Rosenthal, 2008). The level of variability in their conclusions depended on a number of factors: the strength of the biasing irrelevant contextual information (which can be relatively weak, such as "the detective believes the person is guilty," or can be relatively strong, such as "the suspect was in custody in jail when the crime was committed"), and the difficulty of the decision (when decisions are more complex and difficult, there is more leeway for bias to influence the conclusions).

In forensic psychology, we know of no comparable research. Within-expert studies are generally harder to conduct than between-expert studies and probably even more difficult to conduct in forensic psychology than in forensic science. Whereas a forensic science expert may not so easily recognize that she is examining the same fingerprints or gunshot residue sample she examined one year earlier, a forensic psychologist would more likely recognize she is examining the same defendant for the same referral question. Nevertheless, these types of studies are critical if we are to explore, understand, and minimize potential biases. Researchers could use methods that do not require repeated exposure to an actual defendant (e.g., presenting vignettes, case summaries, case referrals, or results of psychological tests). For example, would a forensic psychiatrist respond the same way to an attorney's case description if that attorney mentioned a minimal pay rate typical for evaluations of indigent defendants versus a lucrative hourly rate with the promise of more lucrative referrals? Does a psychologist who works in a psychiatric hospital make the same decision about a defendant's restoration to competence when the hospital census is high and hospital beds are scarce as compared to times when there are fewer space constraints? Though such studies may be more challenging to conduct in forensic psychology than in forensic science, the underlying questions are ripe for study and important to explore.

6. Reliability Between Experts' Conclusions

Even without biasing contextual information, will experts examining the same information reach the same conclusions? In forensic science, studies of the basic reliability among experts show that even DNA and fingerprint experts will reach a spectrum of different (and conflicting) conclusions when they examine the same evidence, even absent any irrelevant biasing information about the suspect's likely guilt (e.g., Coble, 2015; Dror & Hampikian, 2011; Dror & Rosenthal, 2008).

Poor reliability between experts threatens the goals of accurate, equitable justice. Stated bluntly, poor between-expert reliability means that whether a suspect goes to prison or not may depend on the 'luck of the draw' as to which expert examines the evidence. Furthermore, since forensic evidence is rarely contested in court, when a fingerprint expert testifying that the fingerprint from the crime scene matches the suspect, it is highly incriminating and often results in a conviction. However, if a different fingerprint expert would have examined the same fingerprints—purely by chance lab assignment procedures—that expert may have not concluded that the prints matched. Hence the dire consequence of a dangerous mix: forensic science evidence is very powerful in court, but different forensic science experts may reach different conclusions. Indeed, many wrongful convictions have relied on confident, uncontested expert conclusions about forensic evidence (Garrett & Neufeld, 2009).

In forensic psychology and psychiatry, several studies provide data about the reliability among expert conclusions (the between, interexpert performance). These data may take the form of reliability estimates for clinicians administering the same instrument (e.g., Boccaccini et al., 2012; Otto et al., 1998; Rogers, Jackson, Sewell, Tillbrook, & Martin, 2003) or clinicians performing the same forensic assessment (e.g., Gowensmith, Sessarego, et al., 2017). Also relevant are studies of whether clinicians assign the same diagnosis to the same individual or case description. For example, in medical research on psychiatric diagnosis, agreement has been poor ($\kappa < .50$) among clinicians who use unstandardized procedures to assign diagnoses (Aboraya, Rankin, France, El-Missiry, & John, 2006; Spitzer & Fleiss, 1974), and the minimal research on diagnosis in forensic evaluations finds similarly poor agreement (Gowensmith, Murrie, Boccaccini, & McNichols, 2017). Recently, scholars have described an important distinction between reliability under the optimal conditions in formal research studies versus "field reliability" among real-world practicing clinicians performing their routine duties within the conditions typical of their work (Edens & Boccaccini, 2017; Wood, Nezworski, & Stejskal, 1996).

A recent review of the field reliability of the most common forensic evaluations—adjudicative competence and legal sanity—identified 59 studies that purported to address the reliability of competence or sanity opinions, but only 8 (for sanity) and 9 (for competence) actually addressed the reliability among practicing forensic evaluators performing real evaluations (Guarnera & Murrie, 2017). These reported pairwise percent-agreement rates ranged from 57% to 100%, and kappa values ranged from .28 (poor) to 1.0 (perfect).

The studies that best shed light on the routine field-reliability among forensic evaluators are from Hawaii, which historically required (Hawaii Revised Statutes, 2014) three independent evaluations for all felony defendants referred for competence or sanity evaluations. Because the evaluators are relatively independent (court-appointed, not retained by the prosecution or defense) they are less vulnerable to the biasing effect of adversarial allegiance.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Thus, Hawaii provides a "natural experiment" for studying field reliability, without the obvious confounds that bedevil other field studies. Regarding adjudicative competence, a review of 216 Hawaii felony defendants referred for evaluation revealed that the three independent evaluators reached different conclusions in 29% of the cases (Gowensmith, Murrie, & Boccaccini, 2012). Regarding legal sanity, a review of 165 defendants revealed that three independent experts reached different conclusions in 45% of the cases (Gowensmith et al., 2013). Finally, regarding evaluations of whether or not a patient who had been hospitalized as not guilty by reason of insanity (NGRI) was ready for conditional release—a legal question less well-defined in statute than competence or sanity—three independent experts reached different conclusions in 47% of the cases (Gowensmith et al., 2017). Hence, field reliability research suggests that expert reliability tends to be modest, even with common evaluations and with court-appointed forensic experts.

Although reliability among experts is usually measured by examining whether experts reach the same conclusion in the same case, it is also possible to explore reliability by examining other more detailed and sensitive measures within a case, or to examine *patterns* of findings across cases from the same referral stream. One example of more detailed or sensitive measures within a case might be confidence levels (e.g., Douglas & Ogloff, 2003). Even if experts reach the same conclusion, they may have different levels of confidence in their judgments, a distinction that is theoretically and practically important. Theoretically, confidence level serves as a more sensitive measure to understand the decision processes (much like researchers use response time as a more sensitive measure to understand differences even when participants give the same response). Practically, confidence is important because a judge or jury may weigh expert opinions based on the testimony and how strongly experts present their conclusions, which depend on the experts' own confidence in their conclusions. Two experts arriving at the same conclusion may nevertheless convey it quite differently because they differ in the confidence they have in the conclusion. In short, reliability can be examined with more sensitive measures beyond overall conclusion.

Another focus in examining reliability (or biasability) might be examining *how* evaluators communicate findings. For example, evaluators presenting the results of certain violence risk assessments may choose between: sharing numerical estimates in frequency or probability formats (Slovic, Monahan, & MacGregor, 2000), describing potential outcomes in "vivid" or "pallid" terms (Monahan et al., 2002), or describing risk factors in "packed" or "unpacked" format (Scurich, Monahan, & John, 2012). Research reveals that each of these decisions about risk communication substantially influences how decision-makers interpret the risk message, even when that message is substantively identical. Thus risk communication strategies may be another focus of reliability (or biasability) research.

Reliability patterns *across* cases is another possible measurement. If evaluators are generally reliable with one another, evaluators working in the same context with the same referral stream should generally display similar patterns of findings across cases. For example, they might find similar percentages of defendants incompetent or insane, and they might, overall, assign similar mean scores on the same instrument, at least *if* all examinees they evaluate are selected randomly from the same population and there is a sufficient number of cases.

In a study that best illustrates this kind of reliability research, Boccaccini, Turner, and Murrie (2008) examined 20 forensic mental health experts who had contracted with the state of Texas to perform screening evaluations of offenders—including administering and scoring Hare's (2003) Psychopathy Checklist-Revised—as part of specialized "sexually violent predator" laws. All experts examined offenders from the same correctional system, referred through the same office, which made efforts to ensure that offenders were assigned to experts in an essentially random manner. In other words, all experts worked for the same entity, examined offenders from the same "referral stream," and should not have seen systematically different samples of offenders. Nevertheless, results revealed that in the 321 cases they examined, the experts differed drastically in the average PCL-R score they assigned across the cases they saw. Indeed, 34% of the variance in PCL-R total scores was attributable to differences between experts, rather than differences in the offenders they evaluated. For example, some evaluators assigned average scores around or above 30 (indicating highly psychopathic personality) whereas some assigned average scores as low as 8 or 18 (scores on the PCL-R can range from 0 to 40, and research reveals average scores of 22–23 in correctional settings). Similar research suggests that experts differ in the proportion of criminal defendants they conclude are incompetent to stand trial (Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008) or not guilty by reason of insanity (Murrie & Warren, 2005), though in most field studies, it is difficult to exclude all confounds (beyond the expert) that might contribute to this variability.

Overall, most research addressing expert performance in forensic psychology tends to fall within Level 6 in the HEP, *reliability between expert conclusions* (see Figure 2). But even in this Level 6 where there are the most studies (e.g., competence and sanity evaluations, scoring forensic assessment instruments), data are sparse and lack some details necessary to answer important practical questions (Guarnera & Murrie, 2017). For example, the forensic psychology literature has virtually no Level 6 data addressing reliability of forensic psychological evaluations in civil litigation or death penalty cases, where the stakes are highest. There is also lack of reliability data for commonplace civil evaluations like those addressing civil commitment or parenting capacity—situations in which courts may restrict rights (or tolerate risk of great harm) based on the opinion of an evaluator. Better reliability data may reveal training needs and may help inform policies such as those that allow or assign multiple evaluations (as in Hawaii) or allow for second-opinion evaluations (as in other jurisdictions).

With further insights on the nature and factors that affect reliability, specific policies and measures can be developed to increase reliability in forensic psychology. In the forensic sciences, more extensive reliability research has produced policies and tools that help quantify and calibrate expert performance, for example, Fingerprint Analyses Consistency Tester (FACT; Dror et al., 2011). We acknowledge that the challenges to develop such policies and tools in forensic psychology may be greater than those in forensic science; for example, some behavioral evidence may be harder to quantify than physical evidence. However, the aspects of forensic psychology that make it more difficult to develop such tools are

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the same aspects that are likely to contribute to variability in expert performance, and hence the need for interventions to reduce them.

5. Reliability Within Experts' Conclusions

Level 5 in the HEP examines the reliability of conclusions *within,* rather than *between* experts. This within-expert level is a more basic measure of reliability, in that we would expect an expert to reach the same conclusion if considering the same data repeatedly (even if not reaching the same conclusions as other experts). Forensic science researchers examining this type of reliability found, for example, that even the same fingerprint expert examining the same pair of prints will not reach the same conclusions 10% of the time (Ulery, Hicklin, Buscaglia, & Roberts, 2012).

In forensic psychology and psychiatry, we know of *no* analogous research or any data that examines this aspect of expert performance. As we saw in Level 7 of the HEP, *within*-expert studies are much more challenging and difficult to conduct in general, and even more so in the mental health fields. As Kraemer, Kupfer, Clarke, Narrow, and Regier (2012) observed when reviewing diagnostic reliability in psychiatry, "Intra-rater reliability requires that the same rater be asked to "blindly" review the same patient material two or more times . . . Intrarater reliability is almost never assessed for psychiatric diagnosis because it is difficult to ensure blinding of two diagnoses by the same clinician viewing, for example, the same diagnostic interview" (p. 14). Forensic mental health experts rarely examine exactly the same case data (the way a forensic scientist might reexamine the same evidence), and even if they examine the same defendant, the defendant may have changed. For example, an expert might conclude that a defendant is not competent to stand trial but then evaluate her again after she receives competence restoration treatment and conclude that she is competent. Likewise, an expert might evaluate an individual and conclude that he presents little risk of violence, but evaluate him again later, after changes in dynamic risk factors such as psychosis and substance abuse, and conclude that he is at higher risk for violence (see generally, Douglas & Skeem, 2005). In neither scenario would we consider the experts' differing opinions to be unreliable; the changing conclusions may be appropriate because the case has changed. However, the critical question is: all things being equal, would the same examiner reach the same conclusions from the same data? That is a most basic form of reliability underlying forensic conclusions, and there has been no research examining this fundamental aspect of forensic psychology expertise.

Within-expert reliability is so fundamental for expertise that it should become a priority for future research. The challenge for researchers studying the reliability within expert conclusions would be to present precisely the same case data to an expert at different points in time (without the expert recalling having seen it before). Although it may be nearly impossible to do such studies with live defendants, it may be quite feasible with case files or test results, much like sharing fingerprints or other forensic evidence with forensic science experts.

1–4. The Observational Levels

Whereas levels 5 through 8 of the HEP (above) address expert conclusions, levels 1 through 4 address the underlying, more fundamental observations on which conclusions are based (see Figure 2). Failing to study observations is a significant oversight because observations underpin conclusions. Apparent unreliability or biasability at the level of conclusions may actually lie deeper, at the observational level, which would require different intervention. Of course, if research revealed only perfect reliability and minimal biasability at the level of conclusions, there may be little need to conduct similar research at the level of observations. But, as reviewed above, the limited data available suggests that forensic psychological assessments are less than perfectly reliable, and at least sometimes vulnerable to bias. Therefore, it remains critical to disentangle observations and conclusions in order to study them both.

In forensic science, the distinction between observations and conclusions is usually fairly clear. For example, a fingerprint expert will observe the minutia (distinct and well-defined individual characteristics; see Figure 1) in the friction ridge of the fingerprints. The expert then forms a conclusion as to whether they "match" based on whether the observed minutia in the fingerprints are "similar enough."[4]

Recognizing this distinction between observations and conclusions, the forensic science literature has revealed problems at the observation levels 1–4 within and between experts, and with regard to reliability and biasability. For example, fingerprint examiners observe different data in the evidence depending on irrelevant contextual information (Earwaker et al., 2015), and they observe different data in the same evidence, between and within themselves, even without irrelevant contextual information (see Dror et al., 2011). These types of studies measure different aspects of expert performance, quantify them, and reveal the phenomena that contribute to differences at the observational level. These phenomena may arise from human factors (e.g., training and methods, exposure to irrelevant information) and/or from the data itself (i.e., problems of unreliability and biasability are most apparent when the data is of low quality). But studies that explore these phenomena can inform training and best practices to minimize such differences (e.g., the Fingerprint Analyses Consistency Tester, which helps to quantify and calibrate expert observations; Dror et al., 2011).

In forensic psychology, the distinction between observations and conclusions can be more complicated than in forensic science. For example, assigning a clinical diagnosis or reaching an opinion about a defendant's adjudicative competence or legal sanity are certainly *conclusions*, based on observations from a clinical interview, record review, and other sources. But each final conclusion (e.g., is the defendant competent or incompetent?) rests on numerous intermediate conclusions (e.g., does the defendant factually understand the legal process and his charges? does he rationally understand these? can he assist counsel?), *which depend on ob-*

[4] This—whether two prints are "similar enough"—is a subjective determination, because most jurisdictions have no definition or criteria as to what constitutes "similar enough" (Dror & Cole, 2010). Having such subjective criteria may contribute to problems of biasability and unreliability. But similarly subjective criteria are common in forensic psychology. Forensic psychologists may observe a defendant's deficits or impairments, but the law provides no precise guidance as to when the deficits are "deficient enough" to conclude that the defendant is incompetent to stand trial.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

servations (e.g., what information in records is relevant to his capacities? what symptoms did he display during interview?).

Likewise, assigning an overall score on a complex assessment instrument—such as Hare's (2003) Psychopathy Checklist-Revised (PCL-R)—is a conclusion, reached only after countless observations necessary to score each of the instrument's 20 items. But even scoring many of the individual PCL-R items can be viewed as a *conclusion*, based upon many observations and inferences during interview and record review to decide whether a particular criterion—such as superficial charm, pathological lying, shallow emotions, or lack of remorse—should be scored as 0 (*absent*), 1 (*partially present*), or 2 (*clearly present*).

On the other hand, some items on forensic assessment instruments are much more similar to observations than conclusions. For example, the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012) includes items such as "age at release" and "any male victims [in the examinee's sexual offense history]" that require relatively straightforward observations from criminal records. Thus, it sometimes seems clear what is an observation and what is a conclusion; however, other times it is difficult to clearly delineate observations from conclusions.

Forensic psychology research has almost solely focused on expert conclusions and has for the most part neglected researching expert observations. Indeed, our review identified *no* studies of forensic mental health evaluation that specifically addressed Levels 1 to 4 of observations. The only observation-level data that the field appears to offer are certain item-level data from test instrument manuals or test reliability studies. Again, some items on risk measures like the Static-99R (Helmus et al., 2012) are essentially coded observations from criminal records; these tend to show high reliability (Phenix & Epperson, 2015; Phenix, Helmus, & Hanson, 2015). Likewise, item-level data in Hare's (2003) PCL-R manual reveals stronger reliability values for those few items that are more like *observations* from the criminal record (e.g., juvenile delinquency, revocation of conditional release; $\text{ICC}_{\text{A1}} = .75–.80$) than those items that are *conclusions* about behavior (e.g., impulsivity, glibness, callousness; $\text{ICC}_{\text{A1}} = .23–.36$; see Hare, 2003; Sturup et al., 2014). Generally, on forensic assessment instruments that require clinician inference, interrater reliability and predictive validity are both stronger for less subjective items that are more like observations, and weaker for those more subjective items that are more like conclusions (Rufino, Boccaccini, & Guy, 2011) Of course, even ostensibly simple observations may depend on how the data are collected and examined, and these "data collection" procedures (e.g., how a forensic evaluator asks a defendant questions, or which collateral sources a forensic evaluator seeks and prioritizes) may be vulnerable to biases. Furthermore, many of data sources that evaluators use to score these instruments (e.g., police, defendant, victim, or witness statements; prior evaluation reports) may themselves have been influenced by various biases.

Beyond item-level data from forensic assessment instruments, we know of no reliability data for the many observations that inform forensic psychological evaluations, particularly those requiring clinical expertise (e.g., observing symptoms of psychosis, observing indicators of past mental state in collateral records, observing evidence of malingering), nor any data on the biasing effects that irrelevant contextual information may have on such observations. A broad review of interrater reliability estimates across a variety of medical and psychological procedures suggests that what we would consider *observations*—that is, "circumscribed judgment tasks requiring relatively few bits of information—such as test scoring, object counts, or physical measurements (e.g., counting decayed teeth, measuring organ size on ultrasound)"—tend to show stronger reliability than what we could call *conclusions*—"complex tasks requiring the synthesis of multiple, higher inferences (e.g., job performance ratings, stroke classification by neurologists)"; see Meyer, Mihura, & Smith (2005, p. 310). Nevertheless, even reliability for simple observations or judgments was imperfect, suggesting that we should not take observation-level reliability for granted in forensic psychology.

Discussion

The forensic sciences have long focused solely on the objects of their inquiries (DNA, fingerprints, firearms, handwriting, etc.) while neglecting—if not totally ignoring—the critical role that the human experts play in forensic decision making. However, the past decade has seen a dramatic shift in the forensic sciences, now recognizing, researching, developing policies, and mandating changes to reduce variability in expert decision making.

Forensic psychology has also tended to focus on the object of their inquiries—human behavior vis-à-vis legal standards—with far less focus on the critical role that the actual forensic psychologists—as human expert examiners—play in forensic assessments. To be fair, forensic psychology has historically explored certain aspects of reliability (e.g., Poythress & Stock, 1980), particularly in the context of psychological assessment instruments, long before the recent reforms in the forensic sciences. Nevertheless, for a field so rooted in the study of human behavior, cognition, and psychology, there has been surprisingly little attention to the role of human experts and human decision making in forensic psychological assessment. Put bluntly, the field tends to value reliability and objectivity, but tends to consider these more as qualities to be studied and maximized in instruments, with less attention to studying and maximizing these among the human experts rendering forensic opinions.

The revolution in forensic science has brought about scrutiny and examination of what factors influence forensic science experts. Research in this area has demonstrated that variability in forensic science decision making arises from two distinct factors: basic *reliability* (i.e., repeatability, reproducibility, consistency in decision making) and *biasability* (being inappropriately influenced by task-irrelevant information) (Dror & Rosenthal, 2008). Research in this area has further disentangled different components in expert forensic science decision making between the *conclusions* (e.g., whether the fingerprints match; Ulery et al., 2012) versus the *observations* on which the conclusions are based (e.g., what characteristics are observed in the fingerprint; Dror et al., 2011). Finally, the performance and variability of forensic science experts has been examined and quantified *between-experts* (variability among experts, the intervariability performance; Dror & Hampikian, 2011) and *within-experts* (variability within a single expert, the intravariability performance; Dror & Rosenthal, 2008). Combining these elements yields an eight-level framework for expert decision making—the Hierarchy of Expert Performance (HEP; Dror, 2016, see Figure 2), which has helped organize and frame the existing research, shown gaps where further research is

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

needed, and identified specific problematic areas that require improved policies and practices.

Applying HEP to forensic psychology reveals a few areas of relative strength, but more areas in which basic research is sorely lacking. Regarding relative strengths, forensic psychology has offered some research regarding *reliability between experts*' *conclusions*. This research comprises many studies detailing reliability in scoring specific instruments (e.g., Otto et al., 1998; Rogers et al., 2003) and a few studies documenting the field reliability of common criminal forensic evaluations such as those addressing competence and sanity (Guarnera & Murrie, 2017). But the field lacks adequate data for many other types of forensic evaluations, as well as other types of expert conclusions (e.g., those regarding diagnosis or psychological injury) that are central to many forensic evaluations.

Forensic psychology also lacks data at the level of *observations* (in contrast to conclusions). Decades ago, the influential jurist David Bazelon (1982) noticed this weakness and warned psychologists,

> Behavioral scientists who appear in the public arena all too often focus on little more than making conclusory pronouncements. Either they omit any real discussion of underlying observations and methods of inference, or they drown such discussion in a sea of jargon . . .
>
> What the public needs most from any expert, including the psychologist, is a wealth of intermediate observations and conceptual insights that are adequately explained. (p. 116)

Though Bazelon's primary concern was that individual experts disclose the methods, limits, and values underlying their work, his critique remains apt for research has well. We have some data to shed light on evaluator conclusions, but almost none to shed light on the "intermediate observations" on which conclusions should rest.

Consider a practical example: is the reliability in sanity conclusions modest because evaluators disagree in how they make a final inference regarding mental state at the time of offense? Or because evaluators disagree even earlier in the process, by reviewing different sources of information and observing different data in those sources? No available research sheds light on such critical questions, though it is plausible to imagine studies that could do so. For example, studies can remove the observational components from the forensic evaluation: in such studies, the same observations will be provided to the examiners (rather than the data and information which they use to make the observations), hence ensuring that they all start with the same observations; any differences can then be attributed to their inferences rather than to differences in their observations. In contrast to examining such differences in inferences, one can study the observations per se by providing examiners with identical records or videos of interviews, and comparing the evaluators' observations of the data. Furthermore, one can research the interactions between the observations and conclusions, for example, studying whether changes in observations drive changes in conclusions (as they should) versus situations in which expectations about conclusions (e.g., prompted by research designs that provide irrelevant information) influence what data is observed (a biasing effect of working backward, or circular reasoning, which the LSU approach was designed to minimize in forensic science, Dror et al., 2015).

This type of research has the potential to inform tools and resources for evaluators (such as checklists to guide procedures; see Gawande, 2010) or even policies governing evaluations (e.g., mandating that evaluators receive and consider certain uniform information from records). Efforts to study and enhance observation-level reliability could fit practically into many forensic training programs, whether for early stage trainees like graduate students or for practicing professionals participating in continuing education. Indeed, more training emphasis on reliability at the level of observations will likely improve reliability at the level of conclusions. In short, data at the level of observations is conspicuously absent from forensic psychology research, but addressing this gap in the research may be relatively simple.

Whether at the level of observations or conclusions, the field seems to offer no data on reliability *within* experts. This fundamental, foundational form of reliability (see Kraemer et al., 2012) must be examined and quantified rather than presumed. Again, we acknowledge that such studies are challenging, but they are not impossible. Researchers might use case materials (e.g., psychological testing results, collateral records) rather than actual defendants, and incorporate reliability research into training or continuing education programs (see, e.g., Blais, Forth, & Hare, 2017 for an example of incorporating reliability research—albeit *between* experts—into training). Should studies reveal poor within-expert reliability (as have some studies in the forensic sciences), results may help inform more rigorous procedures (checklists, protocols, etc.) for forensic evaluations and early training.

The final domain in which the HEP reveals clear gaps in forensic psychology research is biasability. The available research on bias in forensic evaluation has addressed adversarial allegiance (Murrie & Boccaccini, 2015), a clear threat in adversarial justice systems such as those in the U.S., U.K., and many other countries.[5] But even this research body is small, and limited only to certain types of evaluations (particularly sex offender risk assessment).

Although adversarial allegiance may have received more research attention than other forms of bias, this is certainly not the only threat to objective evaluations. Biases related to race, sex, sexual orientation, age, disability, and religion have, to our knowledge, *never* been explored among forensic evaluators. This research gap is striking, considering that these potential biases are such a popular foci of other types of psychological research, and even forensic psychology research addressing jury-decision making (e.g., Sommers & Norton, 2008). Other potential biases—for example, basic base-rate expectation biases, or biases related to crime details or criminal stereotypes—are currently understudied. Indeed, the field is increasingly attuned to many ways in which forensic evaluators may be vulnerable to bias (see Neal & Grisso, 2014; Zapf & Dror, 2017 for reviews), but empirical research on these biases lags behind.

[5] We do not claim that alternative (nonadversarial) justice systems, such as the inquisitorial system, are better overall. Though less vulnerable to adversarial allegiance, they may be more vulnerable to other biases or they may sacrifice strengths inherent in the adversarial system. Comparative research may reveal relative strengths and weaknesses of each, or interventions that the adversarial legal system can learn from other systems (such as expert "hot-tubbing;" Edmond, 2009).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Policy Implications

Our goal in this review has been primarily to provide a conceptual and practical framework for understanding and studying the performance of forensic evaluators. Without such a framework, it is easy to overlook the gaps and limitations in the available literature, which then leaves us more likely to pursue policies and practice that overlook (or even exacerbate) underlying problems. Indeed, our review of forensic psychology using the HEP reveals that there are many gaps in our knowledge base that prevent us from prescribing specific remediation in many areas.

Thus, the first priority should be performing research (some of which we have recommended throughout this review) that helps us better understand problems of unreliability and biasability within and between experts, at the levels of observations and conclusions. That said, some of this recommended research could *also* serve as pilot testing for certain policy or practice interventions. Likewise, certain policy or intervention studies might identify the nature of underlying problems, much like intervention studies with certain pharmaceuticals or medical procedures can shed light on the mechanisms underlying a disease. Therefore, we provide general suggestions for potential studies of interventions or policies that may help us better understand underlying unreliability and bias.

First, studies of existing policy arrangements are valuable. For example, the few jurisdictions that assign more than one evaluator per case (see Gowensmith, Pinals, & Karas, 2015) allow for studies of field reliability under various conditions. Take for example the studies of Hawaii's three-evaluator system (Gowensmith et al., 2012, 2013; 2017), which revealed that real-world reliability may be modest even in arrangements that have minimized the biasability attributable to adversarial allegiance. To take another example, some hospitals and at least one state (i.e., Virginia) have developed oversight policies that monitor conclusions from each of their evaluators on every assigned evaluation; these may allow for naturalistic reliability-type studies examining *patterns* of findings across evaluators.

Second, addressing bias in particular, forensic psychology may benefit from considering the interventions and policies recently emerging from the forensic sciences. After documenting the powerful influence of irrelevant contextual information (Kassin, Dror, & Kukucka, 2013), and identifying specific mechanisms such as *bias cascade* versus *bias snowball* (Dror, Morgan, Rando, & Nakhaeizadeh, 2017), the field is working to distinguish task-relevant from task-irrelevant information (National Commission on Forensic Science, 2015), developing policies and procedures to shield analysts from the latter. Indeed, these efforts have resulted in well-developed strategies for laboratories and agencies to process cases and evidence in ways that minimize analyst exposure to potentially biasing information; these include the use of case managers (Dror, 2013) and Linear Sequential Unmasking (LSU; Dror et al., 2015).

Such research addressing bias and best practices has been adopted in forensic science policy in the United States (NCFS, 2015) and in the United Kingdom (Forensic Science Regulator, 2015). Again, we acknowledge that distinguishing task-relevant from task-irrelevant information is more challenging in forensic psychology, but it is no less important. Furthermore, the policies do not only address issues of what is task irrelevant, but also 'when' task relevant information should be provided, that is, the best sequence and order that relevant information should be provided to examiners to minimize bias (such as circular and backward reasoning, see, e.g., the LSU policy). Trying to apply such de-biasing policies and procedures from the forensic sciences to forensic psychology is an important and promising first step. Some of the policies may be easily applied, others may require modifications and adaptations, whereas others may not be applicable. In any case, such an attempt is worthwhile and will help in creating policies and research strategy to better understand biasability, and ways to reduce bias.

Regarding the adversarial allegiance bias, the most commonly recommended policy is to explore court-appointed experts. Though intuitively appealing, court-appointed experts bring new challenges and dilemmas (Mnookin, 2008; Murrie & Boccaccini, 2015) that no research has yet addressed. However, researchers could explore a few domains, particularly child custody litigation, in which informal policies have begun to shift toward court-appointed, or jointly appointed experts over adversarial, opposing experts. Researchers might explore other appealing, but untested legal reform proposals such as "blinding" experts to the side retaining their services (Robertson & Kesselheim, 2016; Robertson & Yokum, 2012).[6] Again, we do not necessarily recommend any of these as broad policy reforms because there are not yet sufficient data to support them, but there is clearly value in exploring some of these in order to shed light on unreliability and biasability, as well as their potential solutions.

Conclusion

Forensic psychology offers well-developed procedures for the expert assessment of criminal defendants and civil litigants. But the field has offered much less data exploring the decision making of the forensic experts themselves. Recent reforms in the forensic sciences underscore the need to carefully study forensic experts, and Dror's (2016) HEP conceptualizes and defines the aspects involved in expert decision making, thus helping to frame the existing research and identify gaps. Forensic psychology can learn from these insights and use HEP to benefit and enhance forensic psychology decision making.

[6] Of course, not all potential interventions would reduce allegiance effects. Neal and Grisso (2014) proposed a "thought experiment", in which adversarial experts simply present the most compelling case they can for the side that retained them, with no pretense of objectivity.

References

Aboraya, A., Rankin, E., France, C., El-Missiry, A., & John, C. (2006). The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry, 3,* 41–50.

American Educational Research Association, American Psychological Association, and National Council on Measurement Education. (2014). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA Publications.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073. http://dx.doi.org/10.1037//0003-066X.57.12.1060

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Bazelon, D. L. (1982). Veils, values, and social responsibility. *American Psychologist, 37,* 115–121. http://dx.doi.org/10.1037/0003-066X.37.2.115

Blais, J., Forth, A. E., & Hare, R. D. (2017). Examining the interrater reliability of the Hare Psychopathy Checklist-Revised across a large sample of trained raters. *Psychological Assessment, 29,* 762–775. http://dx.doi.org/10.1037/pas0000455

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A., & Jeglic, E. (2012). Implications of Static-99 field reliability findings for score use and interpretation. *Criminal Justice and Behavior, 39,* 42–58. http://dx.doi.org/10.1177/0093854811427131

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

Coble, M. (2015). *Interpretation errors detected in a NIST interlaboratory study on DNA mixture interpretation in the U.S. (MIX13)*. Presentation at the International Symposium on Forensic Science Error Management: Detection, Measurement, and Mitigation, Washington, DC.

Douglas, K. S., & Ogloff, J. R. P. (2003). The impact of confidence on the accuracy of structured professional and actuarial violence risk judgments in a sample of forensic psychiatric patients. *Law and Human Behavior, 27,* 573–587. http://dx.doi.org/10.1023/B:LAHU.0000004887.50905.f7

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11,* 347–383. http://dx.doi.org/10.1037/1076-8971.11.3.347

Dror, I. E. (2013). Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management: An International Journal, 4*(3–4), 105–113. http://dx.doi.org/10.1080/19409044.2014.901437

Dror, I. E. (2016). A hierarchy of expert performance. *Journal of Applied Research in Memory & Cognition, 5,* 121–127. http://dx.doi.org/10.1016/j.jarmac.2016.03.001

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. *Forensic Science International, 208*(1–3), 10–17. http://dx.doi.org/10.1016/j.forsciint.2010.10.013

Dror, I. E., & Cole, S. A. (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review, 17,* 161–167. http://dx.doi.org/10.3758/PBR.17.2.161

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

Dror, I. E., Morgan, R. M., Rando, C., & Nakhaeizadeh, S. (2017). The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making. *Journal of Forensic Sciences, 62,* 832–833. http://dx.doi.org/10.1111/1556-4029.13496

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

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Sciences, 60,* 1111–1112. http://dx.doi.org/10.1111/1556-4029.12805

Earwaker, H., Morgan, R. M., Harris, A. J. L., & Hall, L. J. (2015). Fingermark submission decision-making within a UK fingerprint laboratory: Do experts get the marks that they need? *Science & Justice: Journal of the Forensic Science Society, 55,* 239–247. http://dx.doi.org/10.1016/j.scijus.2015.01.007

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment "out of the lab" and into "the real world": Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29,* 599–610. http://dx.doi.org/10.1037/pas0000475

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

Forensic Science Regulator. (2015). *Cognitive bias effects relevant to forensic science examinations: Guidance*. Birmingham: The Forensic Science Regulator. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/510147/217_FSR-G-217_Cognitive_bias_appendix.pdf

Foster, W. L. (1897). Expert testimony, prevalent complaints and proposed remedies. *Harvard Law Review, 11,* 169–186. http://dx.doi.org/10.2307/1321970

Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97.

Gawande, A. (2010). Checklists for success inside the OR and beyond: An interview with Atul Gawanda, MD, FACS. Interview by Tony Peregrin. *Bulletin of the American College of Surgeons, 95,* 24–27.

Gowensmith, W. N., Murrie, D. C., & Boccaccini, M. T. (2012). Field reliability of competence to stand trial opinions: How often do evaluators agree, and what do judges decide when evaluators disagree? *Law and Human Behavior, 36,* 130–139. http://dx.doi.org/10.1037/h0093958

Gowensmith, W. N., Murrie, D. C., & Boccaccini, M. T. (2013). How reliable are forensic evaluations of legal sanity? *Law and Human Behavior, 37,* 98–106. http://dx.doi.org/10.1037/lhb0000001

Gowensmith, W. N., Murrie, D. C., Boccaccini, M. T., & McNichols, B. J. (2017). Field reliability influences field validity: Risk assessments of individuals found not guilty by reason of insanity. *Psychological Assessment, 29,* 786–794. http://dx.doi.org/10.1037/pas0000376

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., Sessarego, S. N., McKee, M. K., Horkott, S., MacLean, N., & McCallum, K. E. (2017). Diagnostic field reliability in forensic mental health evaluations. *Psychological Assessment, 29,* 692–700. http://dx.doi.org/10.1037/pas0000425

Guarnera, L., & Murrie, D. C. (2017). Field reliability of adjudicative competence and legal sanity opinions: A systematic review and meta-analysis. *Psychological Assessment, 29,* 795–818. http://dx.doi.org/10.1037/pas0000388

Guarnera, L., Murrie, D. C., & Boccaccini, M. T. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science, 3,* 143–152.

Hand, L. (1901). Historical and practical considerations regarding expert testimony. *Harvard Law Review, 15,* 40–58. http://dx.doi.org/10.2307/1322532

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

Hawaii Revised Statutes, Vol. 14, §704–111 (2014).

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. http://dx.doi.org/10.1177/1079063211409951

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Research in Memory & Cognition, 2, 42–52. http://dx.doi.org/10.1016/j.jarmac.2013.01.001

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). *DSM–5*: How reliable is reliable enough? *The American Journal of Psychiatry, 169,* 13–15. http://dx.doi.org/10.1176/appi.ajp.2011.11010050

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (2007). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers*. New York, NY: Guilford Press.

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

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior, 29,* 621–637. http://dx.doi.org/10.1007/s10979-005-8122-9

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

Monahan, J., Heilbrun, K., Silver, E., Nabors, E., Bone, J., & Slovic, P. (2002). Communicating violence risk: Frequency formats, vivid outcomes, and forensic settings. *The International Journal of Forensic Mental Health, 1,* 121–126. http://dx.doi.org/10.1080/14999013.2002.10471167

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among expert witnesses. *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., 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. http://dx.doi.org/10.1007/s10979-007-9097-5

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

National Commission on Forensic Science. (2015). *Ensuring that forensic analysis is based upon task relevant information*. National Institute of Standards and Technology. Retrieved from https://www.justice.gov/ncfs/file/818196/download

National Institute of Standards and Technology. (2016). *New NIST Center of Excellence to Improve Statistical Analysis of Forensic Evidence*. Retrieved May 30, 2017 from https://www.nist.gov/news-events/news/2015/05/new-nist-center-excellence-improve-statistical-analysis-forensic-evidence

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

Neal, T. M. S., & Grisso, T. (2014). The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy, and Law, 20,* 200–211. http://dx.doi.org/10.1037/a0035824

Otto, R. K., Poythress, N. G., Nicholson, R. A., Edens, J. F., Monahan, J., Bonnie, R. J., . . . Eisenberg, M. (1998). Psychometric properties of the MacArthur Competence Assessment Tool–Criminal Adjudication. *Psychological Assessment, 10,* 435–443. http://dx.doi.org/10.1037/1040-3590.10.4.435

Packer, I. K. (2009). *Evaluation of criminal responsibility*. New York, NY: New York: Oxford University Press. http://dx.doi.org/10.1093/med:psych/9780195324853.001.0001

Phenix, A., & Epperson, D. L. (2015). Overview of the development, reliability, validity, scoring, and uses of the Static-99, Static-99R, Static-2002, and Static-2002R. In A. Phenix & H. M. Hoberman (Eds.), *Sexual offending: Predisposing conditions, assessments, and management* (pp. 437–455). New York, NY: Springer.

Phenix, A., Helmus, H., & Hanson, R. K. (2015). Static-99R and Static-2002R evaluators' workbook [Unpublished manual]. Retrieved from www.static99.org

Poythress, N. G., & Stock, H. V. (1980). Competency to stand trial: A historical review and some new data. *The Journal of Psychiatry & Law, 8,* 131–146.

President's Council of Advisors on Science and Technology. (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.theiai.org/president/201609_PCAST_Forensic_Science_Report_FINAL.pdf

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

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

Rogers, R., Jackson, R. L., Sewell, K. W., Tillbrook, C. E., & Martin, M. A. (2003). Assessing dimensions of competency to stand trial: Construct validation of the ECST-R. *Assessment, 10,* 344–351. http://dx.doi.org/10.1177/1073191103259007

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. http://dx.doi.org/10.1177/1073191110378482

Scurich, N., Monahan, J., & John, R. S. (2012). Innumeracy and unpacking: Bridging the nomothetic/idiographic divide in violence risk assessment. *Law and Human Behavior, 36,* 548–554. http://dx.doi.org/10.1037/h0093994

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24,* 271–296. http://dx.doi.org/10.1023/A:1005595519944

Smalarz, L., Madon, S., Yang, Y., Guyll, M., & Buck, S. (2016). The perfect match: Do criminal stereotypes bias forensic evidence analysis? *Law and Human Behavior, 40,* 420–429. http://dx.doi.org/10.1037/lhb0000190

Sommers, S. R., & Norton, M. I. (2008). Race and jury selection: Psychological perspectives on the peremptory challenge debate. *American Psychologist, 63,* 527–539. http://dx.doi.org/10.1037/0003-066X.63.6.527

Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *The British Journal of Psychiatry, 125,* 341–347. http://dx.doi.org/10.1192/bjp.125.4.341

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

Thomas, C. M., Bertram, E., & Johnson, D. (2009). The SBAR communication technique: Teaching nursing students professional communication skills. *Nurse Educator, 34,* 176–180. http://dx.doi.org/10.1097/NNE.0b013e3181aaba54

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE, 7*(3), e32800. http://dx.doi.org/10.1371/journal.pone.0032800

Wacogne, I., & Diwakar, V. (2010). Handover and note-keeping: The SBAR approach. *Clinical Risk, 16,* 173–175. http://dx.doi.org/10.1258/cr.2010.010043

Weed, L. L. (1970). *Medical records, medical evaluation, and patient care: The problem-oriented medical record as a basic tool*. Cleveland, OH: Press of Case Western Reserve University.

Wigmore, J. H. (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*. Boston, MA: Little, Brown.

Wood, J. M., Nezworski, M. 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

Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *International Journal of Forensic Mental Health*.

Zapf, P. A., & Roesch, R. (2009). *Evaluation of competency to stand trial*. New York, NY: Oxford University Press.

Received March 20, 2017
Revision received June 22, 2017
Accepted June 23, 2017 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.