The Jury EXPERT
from JANUARY/FEBRUARY 2013
Volume 25, Issue 1
The Art and Science of Litigation Advocacy
A publication of the American Society of Trial Consultants Foundation

Forensic Mental Health Evaluations:
Reliability, Validity, Quality, and Other Minor Details

by W. Neil Gowensmith, Daniel Murrie, and Marcus T. Boccaccini

Reliability is crucial to expert evidence. In cases involving mental health, the court usually relies on the opinions and testimony of forensic mental health expert witnesses (those experts who specialize in the intersection of mental health and the law). Even in adversarial proceedings, independent forensic experts appointed by the court are presumed objective and generally reliable. In other words, an opinion from one neutral expert should be similar to the opinion from another neutral expert when the two are considering the same case details.

But how reliable are these forensic experts? That is, how often do independent, court-appointed forensic experts agree with each other? Further, what factors might influence that reliability? Do some types of cases lead to more disagreement than others? Is agreement better for some questions (e.g., competence to stand trial) than others (e.g., insanity)?

To answer some of these questions, we reviewed nearly 350 real cases in which multiple forensic evaluators, in routine practice, evaluated the same defendants to answer questions of competency to stand trial, legal sanity (criminal responsibility), and readiness for release from a psychiatric hospital. Our goal was to examine how often we might expect forensic evaluators to agree on the most common psycho-legal questions the court asks of them. We calculated evaluator agreement across these cases, researched the eventual court dispositions, and explored factors that increased or decreased evaluator agreement. We present these findings later in this paper. First, we review how the evaluations in our study were ordered and conducted.

The Forensic Evaluation

We studied evaluations from Hawaii, where state statutes dictate a unique process that provides an excellent setting for examining reliability. In felony cases, the courts order three concurrent and independent evaluations of the defendant. One of these evaluations must be conducted by an employee of the state Department of Health. The other two evaluations are conducted by independent certified evaluators in the community. One of these independent evaluators must be a licensed psychiatrist, while the other may either be a licensed psychiatrist or a licensed psychologist. All evaluators are appointed by the court, not by the defense or prosecution.

In this way, evaluators in Hawaii are *independent*, so any disagreement we find is *not* likely to be attributable to "adversarial allegiance," the tendency for experts to form opinions that support the party who retained them (see Murrie et al, 2008). All of the evaluators in this study had been certified by the state Department of Health through a series of trainings on forensic evaluation. These conditions allowed for a unique, naturalistic study of the field reliability of forensic evaluations; because each case requires three independent and concurrent evaluations, we could easily compute agreement rates across each case and identify factors related to that reliability.

We reviewed opinions from the most common forensic evaluations: competency to stand trial, legal sanity, and readiness for "Conditional Release" (release from the state hospital subsequent to placement after a verdict of insanity).

Competency to Stand Trial

In lay terms, competency to stand trial (CST) refers to a defendant's ability to understand his or her court proceedings and work productively with his or her defense counsel. Like all states, Hawaii uses the *Duksy*criteria for competency (Dusky v United States, 1960). That is, the defendant must demonstrate a factual and rational understanding of the charges against him, and must be able to assist defense counsel (see Drope v Missouri, 1975).

How reliable are evaluations of a defendant's competency to stand trial? Previous results were mixed, with some showing reasonable agreement among clinicians and others showing poor agreement. Most previous research utilized artificial experimental conditions (such as hypothetical vignettes, or studies in which evaluators use the same instruments in the same hospital), which tended to reveal strong reliability but may not translate adequately to real-world forensic practice. Thus routine reliability "in the field," has been largely unknown.

We coded data from a total of 716 CST reports, taken from 241 cases (full details available in Gowensmith, Murrie & Boccaccini, 2012). Seven Department of Health psychologists, 15 independent psychologists, and 16 certified independent psychiatrists submitted the reports. In most cases, three different evaluators saw each defendant. Thus, evaluators could show unanimous agreement in one of two ways: all could agree that the defendant was competent to stand trial, or all could agree that the defendant was incompetent.

How often did all three evaluators agree with each other? In 71% of cases involving initial evaluations of competency to stand trial, all three evaluators unanimously agreed in their opinion about the defendant's competency. Most of those cases (59%) involved unanimous agreement that the defendant was competent, and fewer (12%) involved unanimous agreement that the defendant was *in*competent. For cases involving repeated evaluations of competency (i.e., re-evaluation after incompetent defendants received treatment to restore competence), agreement rates fell to 61.0%.

When it came to the actual court decisions about a defendant's competence, judges typically followed the "majority opinion" from evaluators. When judges ruled in the opposite direction of the majority of evaluators, they usually did so to find a defendant *in*competent to stand trial. This reflects the court's conservative stance towards competency; that is, they were apparently reluctant to find a defendant competent if there was any doubt among evaluators. Judges were also far more likely to rule against the majority recommendation of evaluators when evaluators presented a split decision on competency (i.e., two say competent, one says incompetent).

We explored several factors that we believed might influence evaluator agreement: the age, gender, and ethnicity of the defendant, the seriousness of the offense, the location of the evaluation, the referral court, the judge presiding over the case, the professional discipline or employer of the evaluators, and the defendant's proficiency with the English language. None of these factors significantly influenced agreement among evaluators. However, when evaluators agreed that a defendant was psychotic (that is, demonstrated severe symptoms such as hallucinations, delusions, or grossly disorganized behavior), they showed better agreement about competence. Fortunately, further analysis revealed that evaluators did not simply conflate a psychotic diagnosis with the finding of incompetence, a problem that has historically been common in competence evaluations (Skeem & Golding, 1998).

Legal Sanity / Criminal Responsibility

We also investigated rates of agreement regarding legal sanity (also known as criminal responsibility). Unlike competency to stand trial, which is a dynamic condition focused on a defendant's current functioning–which may change from moment to moment– legal sanity is a static, historical condition that requires retrospectively determining a defendant's functioning at the moment of his crime. The state of Hawaii uses a version of the two-pronged American Legal Institute standard for legal sanity, which considers both the M'Naughten standard (whether the defendant understood the criminal behavior was wrong) and the volitional capacity standard (whether the defendant could resist the impulse to commit the crime).

Very little previous research has been conducted on the field reliability of legal sanity evaluations. Indeed, *no* recent literature examines evaluator agreement in real cases involving legal sanity.

We coded 468 sanity evaluation reports across 161 cases (for details, see Gowensmith, Murrie & Boccaccini, in press). The proportion of psychologists (24) versus psychiatrists (12) was similar to the pattern we found in CST evaluations.

How often did evaluators agree with each other regarding a

defendant's legal sanity? We found unanimous agreement among evaluators in 55% of legal sanity cases. Evaluators unanimously agreed that the defendant was sane in 38% of cases, and they unanimously agreed the defendant was insane in 17% of cases. When evaluators disagreed, two of the three evaluators more often opined that the defendant was sane rather than insane.

When these sanity cases went to trial, judges were more likely to "overrule" the majority opinion of the evaluators in cases of legal sanity than in cases involving competency to stand trial. They typically did so to find defendants legally sane even when two or three evaluators opined them as *in*sane. In fact, in only one out of 91 cases did a judge find a defendant insane when the majority of evaluators believed the defendant to be sane.

Unlike competency to stand trial evaluations, several factors influenced rates of evaluator agreement in cases involving legal sanity. Evaluators were more likely to agree about sanity when they agreed the defendant warranted diagnosis of a psychotic disorder or when the defendant had been hospitalized in a psychiatric facility sometime in the six months prior to the evaluation. Evaluators were more likely to *dis*agree with each other when the defendant had been abusing substances (making it difficult to disentangle the effects of mental illness versus substance abuse) or when the defendant had committed a violent felony.

Readiness for Cconditional Release

Finally, we investigated agreement rates for evaluators assessing readiness for conditional release (CR). "Conditional release" in Hawaii refers to the community placement of a person previously acquitted by the insanity defense. Conditional release procedures are typically required in every jurisdiction that has an insanity defense. CR readiness evaluations typically involve some form of violence risk assessment, a broader category of evaluation that requires evaluators to measure and comment on an individual's likelihood to act violently.

Unlike competency to stand trial and legal sanity, there is little statutory guidance for the CR evaluation. The statute requires that evaluators form an opinion as to whether or not the insanity acquittee can "be safely managed in the community" once released from commitment status. However, the statutes give no additional guidance on this issue, making the legal question far less clear than competence or sanity.

We reviewed 175 real evaluation reports across 62 cases (McNichols, Gowensmith, Murrie & Boccaccini, 2011). Unanimous agreement rates were the lowest of all three evaluation types we studied. Evaluators agreed unanimously on a person's readiness for CR in only 53.2% of cases. Nearly 90% of these cases involved all three evaluators agreeing that the person was indeed ready for CR. When evaluators disagreed, the two evaluators in most of the split decisions were just about as likely to recommend against CR as they were to support the motion for CR. None of the additional factors that we examined in this study significantly influenced the agreement rates of evaluators on CR readiness evaluations.

Of all the psycho-legal questions that we studied, judges were most likely to "overrule" the majority recommendation of evaluators in cases involving readiness for CR. That is, judges appeared to err on the side of caution, by retaining a patient in the hospital, even when the majority of evaluators opined the patient was ready for release.

Did evaluator agreement relate to case outcome? Of the 62 patients who petitioned for conditional release, the court ultimately granted conditional release to 43 of them. We followed all 43 of these cases for up to three years post-hospital discharge and documented rates of rehospitalization. In cases in which evaluators unanimously agreed that the person was ready for CR, 34.5% were rehospitalized within three years. This approximates a base rate for rehospitalization within the Hawaii CR population, and is similar to other rates of rehospitalization in similar populations across the United States. In cases in which evaluators *dis*agreed, however, 71.4% of individuals granted CR were rehospitalized within three years. In other words, the patients about whom evaluators tended to disagree were indeed those patients who were more likely to "fail" on conditional release (or at least to require re-hospitalization).

Decision-making in Forensic Evaluations

We also explored the rationale behind the conditional release decision-making in the evaluators themselves. Previous work along these lines has been done for competency to stand trial evaluations; Skeem and Golding (1998) found substantial differences among competency reports, with many evaluators documenting little to no rationale for their decision on competency in their reports. Given the low rates of agreement in CR evaluations, and the lack of statutory guidance for CR readiness, we explored how evaluators make decisions on hospital discharge.

We gave 46 certified forensic evaluators a list of 21 potentially relevant factors to be considered in a CR evaluation. We asked them to rank these factors, and we then asked them to identify their understanding of the psycholegal question for CR readiness. Evaluators showed substantial agreement on the importance of "past violence" in determining readiness for conditional release. However, evaluators disagreed on the importance of all the other factors; no other factor was endorsed by more than half of the evaluators, but two-thirds were listed in individual evaluators' "top three" lists. Also, evaluators were nearly evenly split on how to interpret the statute ordering the evaluation. Forensic evaluators seem to have no clear agreement on what factors are important to consider in conditional release readiness applications, or even what the question means in the first place – likely causing the low reliability found across these evaluations. In other words, Hawaii's ambiguous legal

criterion for this particular type of evaluation apparently leaves evaluators interpreting and measuring the relevant issues in different ways.

What Do These Reliability Studies Mean for Attorneys and Trial Consultants?

First, we should expect to see some disagreement among forensic mental health experts, particularly in complex cases. Attorneys and consultants who routinely handle cases that require mental health testimony will inevitably encounter some in which reasonable experts seem to disagree.

Does this mean that expert mental health testimony is worthless? Not at all. The levels of agreement among evaluators in our studies were significantly better than chance. For example, using the base rates for sanity opinions found in our sample, the likelihood that three evaluators will agree on a dichotomous opinion of legal sanity *by chance alone* is 31%; our research showed that evaluators agreed at a rate of approximately 55%, which is well above chance. Agreement rates for competency to stand trial were substantially higher (71%), far exceeding chance levels. Thus, experts agreed in most cases, particularly when the legal question was more straightforward and well-defined (e.g., competence to stand trial).

Arriving at a unanimous decision on "straightforward" forensic evaluations—those that have clearly defined statutory criteria and sound psychometric assessments easily available to evaluators—is itself a tall order. Expecting unanimous agreement on evaluations that require retrospective decision-making (legal sanity) or interpreting fuzzy statutory criteria (conditional release) is simply unrealistic. In addition, the clinical data that evaluators must consider are rarely unambiguous. Complicating factors abound: defendants may misrepresent or malinger their symptoms, important records may be unavailable, and it is inevitably difficult to infer mental state in the past or present. Challenging and confusing cases will always exist; this is the rationale behind requesting a "second opinion" from a medical doctor – or behind checking that second weather report before holding your daughter's outdoor wedding in the backyard. Our findings of less-than-perfect agreement (even in non-adversarial contexts) suggest that it may be worthwhile and reasonable to seek a second opinion in complex cases.

Second, because disagreements among experts are not common, it is important to consider an expert's *procedure* not just the expert's final opinion. Although judges do tend to follow the evaluator's ultimate opinion, we suggest that the opinion itself is less important than the procedures and data that underlie that opinion. When litigation features disagreeing experts, consultants and attorneys should be ready to scrutinize—and help the court scrutinize—the procedures that an expert followed, and the data an expert considered, to reach a particular conclusion. Often, the reasons for disagreements become clear when evaluators are asked to detail the information they considered (or failed to consider) or the inferences they used to connect data and form an opinion. Because many forensic evaluations are genuinely complex and difficult, there are often decision points (e.g., Are additional collateral records necessary?) and inferences (e.g., how does this new data fit with the existing records?) in evaluations during which reasonable professionals might disagree. It is important to identify these decision points and ambiguous data for careful scrutiny. Ask forensic experts to "show their work," not just state their opinion.

Input from forensic mental health experts can be helpful—even essential—to answer certain legal questions. But, like any expert opinion on complex matters, opinions from mental health experts may vary, particularly on complex cases, and this requires educated consumers to carefully consider the data and procedure underlying forensic evaluations.

W. Neil Gowensmith, PhD is an Assistant Professor in the Master's of Forensic Psychology Program at the University of Denver's Graduate School of Professional Psychology. As a clinician, Dr. Gowensmith performs criminal forensic psychological evaluations and was previously the chief of statewide forensic services for the state of Hawaii. His research program focuses on issues related to forensic assessment (particularly field reliability, validity and quality) and the public forensic mental health system.

Marcus T. Boccaccini is an Associate Professor in the Psychology and Philosophy Department at Sam Houston State University. His recent consulting work has focused on strategies for explaining psychological test results to judges and jurors. His research program focuses broadly on the area of forensic assessment, with emphases in field reliability and validity.

Daniel Murrie, PhD serves as Director of Psychology at the University of Virginia's Institute of Law, Psychiatry and Public Policy (ILPPP), an Associate Professor in the School of Medicine, and an instructor in the School of Law. As a clinician, Dr. Murrie performs criminal and civil forensic psychological evaluations. As a researcher, Dr. Murrie studies topics related to forensic assessment, particularly bias and quality control. For details, see here or here.

References

Gowensmith, W., Murrie, D.C., & Boccaccini, M.T. (in press). How reliable are forensic evaluations of legal sanity? Law and Human Behavior*. doi: 10.1037/lhb0000001

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

Murrie, D.C., Boccaccini, M.T., Turner, D., Meeks, M., Woods, C. & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation?Psychology, Public Policy, and Law, 15,19–53. doi: 10.1037/a0014897

Skeem, J., & Golding, S. (1998). Community examiners' evaluations of competence to stand trial: Common problems and suggestions for improvement. Professional Psychology: Research and Practice, 29,357–367.

We asked two trial consultants to respond to this paper. On the following pages, Doug Green and Roy Aranda respond.

Doug Green responds:

Doug Green is the principal consultant with Douglas Green Associates, Inc. which is based in greater New Orleans, but has a national scope, working mostly in civil litigation. Doug has a Ph.D. in Psychology from Tulane University is once again serving on the board and is the President-Elect of the American Society of Trial Consultants.

Because my practice is focused almost exclusively on civil litigation, the principal implications of this research do not necessarily apply directly to my clients. But, underlying these findings is a core concept that I believe applies to expert testimony in any kind of case. While it is generally accepted that experts are indispensable in most kinds of civil litigation, in my experience jurors view experts in a much different way today than they did 20 years ago. This experience comes from conducting hundreds of mock jury studies and interviewing actual jurors after verdicts. The changing perception of experts has important implications for trial lawyers.

When I started working as a trial consultant in the 1980s, most of the work I did involved automotive, product liability cases. At issue in these cases was typically an allegation of design defect. Both sides hired experts in automobile design who would opine on the ultimate question in the case: does the design in question represent a defect. Along the way, the experts would discuss design standards and practices. One side or the other might conduct testing related to the case. And, the presentation of the expert witness at trial always began with an impressive presentation of his or her credentials. Ultimately, there was the opinion that the design was or was not defective. The same was true for injury causation and damages.

Back then, we counted a great deal on the credentials of the expert and his or her ability to persuade the jury that he or she was more experienced, more credentialed, and more of a "real expert" in the field. These factors were very important at the time and we focused mostly on getting the jury to trust the expert for his or her expertise and therefore accept the proffered opinion.

Things slowly started to change towards the end of the 1990s. At the time, I attributed the change to the collapse of Enron, and still do to some extent. Perhaps my bias was that I did a lot of work in Texas. But the Enron scandal exposed an ugly side of American business. At the core of the scandal was unbridled greed and arrogance, and the big losers were the average workers who went to the office every day and did their jobs for nothing more than their middle class wages. They stood to gain nothing by the risks that their employers took, but they paid a very heavy price.

At the same time, I saw a concerning escalation in the fees charged by expert witnesses. When I first started, expert fees were in the range of $150 to $250 per hour. In that range, jurors were impressed, but not shocked. But by the mid–1990s, some experts were charging as much as $500 to $650 per hour. At those rates, jurors started to take serious note of the money changing hands. Then, Enron came to light.

What the scandal stood for in the eyes of many people was that when there was enough money to be gained, some people would do, or say, almost anything. It also created tremendous skepticism about corporations and corporate governance. The role of government regulation in the scandal, or lack thereof, did not become apparent for some time. But, the perception of these events on the part of the average person, the average juror, became a dominant theme in how they perceived cases where individuals were pitted against corporations. Now, the $650 an hour expert was viewed with great skepticism. For that much money, many people believed, a person might say just about anything. The perception of the hired gun became very real. The idea of building trust in an expert became very difficult.

Nothing much has happened to change these attitudes in the intervening years. Around the same time, we saw the dot-com bubble bust and more recently we have seen the sub-prime mortgage crisis. There has also been a massive tort reform movement set in motion largely by the insurance industry, designed to question the motivation of anyone who files a lawsuit. Plaintiffs, after all, have a lot to gain and everyone knows about contingent fee lawyers.

So, what does all of this mean for the use of expert witnesses today? What strategies do we incorporate in my practice to deal with the increasing skepticism of anyone getting paid a lot of money to give opinions? Well, I turn back to the authors' recommendations, which is how I got started on this line of thought: "it is important to consider an expert's procedure not just the expert's final opinion. Ask forensic experts to 'show their work,' not just state their opinion."

As an initial proposition, the philosophy I use when working with experts is that their job is to educate the jury on the relevant field of study to the point where the jurors can examine the evidence and reach their own conclusions. The expert is, therefore, not someone who says, "trust me, I'm an expert," but rather, "let me teach you so you can become an expert."

If you start from this point of view, the qualifications of the expert you choose become clear. I get a lot of calls on this question and the client usually starts by telling me about the potential expert's qualifications. My response is usually, "but can he teach this to the jury?" The precise qualifications of experts, in my opinion, are less important than the individual's ability to communicate and to present difficult concepts to the jury in plain, simple terms. It is also tremendously helpful if the expert is likable and friendly. I find that lawyers tend to parse the qualifications of experts much more finely than do jurors. The gap between the knowledge and experiences of two potential experts will always be far less than the gap between either one and the jurors. When it comes to experts, one should worry more about the ability of a potential expert to communicate and relate to jurors and worry less about expert's specific credentials.

Finally, I believe that the impact of experts on jury decision making today has tremendously diminished compared to 20 years ago. I can't debate the conventional wisdom that experts are essential to most cases. They are often required as a matter of law. But what impact is the expert going to have on the jury verdict? My experience is that in most cases the impact is not much. Jurors today want to hear from fact witnesses. They want to know the story of what happened. If there is a design question in the case, they want to hear from someone actually involved in the design *at the time*. If the issue is patent infringement, they want to hear from the inventor of the patent and the designer of the accused product. The weakness of experts is that they were not involved at the time and are only involved now because they are getting paid – and usually a lot of money. From this point of view, jurors look at experts with great skepticism.

So, my advice to trial lawyers today is to choose experts carefully and use them wisely. Build your case around people who were there at the time – whether they are your witnesses or the other side's – and rely on experts as little as possible. Build the record you need to make your case and hold on to a verdict, but do not expect the jury to care much about the opinions of your experts.

Roy Aranda responds:

Roy Aranda, Psy.D., J.D. is a forensic psychologist with offices in N.Y. and Long Island. He has been involved in several high profile cases including traveling to Cuba and Puerto Rico and testifies frequently in criminal and civil cases throughout New York State.

Gowensmith, Murrie, and Boccaccini have taken their research about how often forensic experts agree with one another in the field up another notch. Drawing upon earlier research (Gowensmith, Murrie, & Boccaccini, 2012) that examined field reliability of competence to stand trial (CST), *Forensic Mental Health Evaluations: Reliability, Validity, Quality, and Other Minor Details* examines forensic evaluations in three contexts: CST; criminal responsibility; and conditional release from a state hospital.

Gowensmith, Murrie, and Boccaccini sought to answer several questions: 1) How often do forensic evaluators agree with another? 2) What factors might influence their reliability? 3) Do some types of cases lead to more disagreement than others? 4) Is agreement better in some contexts than others?

Gowensmith, Murrie, and Boccaccini reviewed nearly 350 cases in Hawaii of multiple forensic evaluators who evaluated the same defendants. Hawaii's unique process provided an excellent setting for several reasons. First, three evaluators are used. This adds a measure of validity that is lacking in settings that rely on a single examiner and when it is assumed that evaluators are interchangeable. Second, because precious little is known about reliability in the field, it provides a natural, real-world setting as opposed to a research setting that employs artificial experimental conditions. Third, the impact of adversarial or partisan allegiance is controlled because all evaluators are independent in as much as they are appointed by the court, not by the defense or prosecution.

Outcome:

CST: In 71% of cases there was unanimous agreement; 59% found that the defendant was competent, and 12% found that the defendant was not competent. Judges typically followed the majority opinion. When they did not they usually took a conservative stand finding that the defendant was not competent to stand trial. Judges also were more likely to rule against the majority when there was a split decision among the evaluators.

Gowensmith, Murrie, and Boccaccini examined the following factors:

- Age of the defendant
- Gender of the defendant
- Ethnicity of the defendant
- Seriousness of the offense
- Location of the evaluation

- Referral court
- Presiding judge
- Professional discipline
- Employer
- Defendant's English-speaking proficiency

Surprisingly, none of these factors significantly influenced agreement among the evaluators.

Analysis revealed that a psychotic diagnosis per se did not result in a finding of incompetence suggesting that functional abilities were looked at more closely.

Criminal responsibility: In 55% of cases there was unanimous agreement; 38% found that the defendant was sane, and 17% found that the defendant was insane. Judges were more likely to overrule the majority opinion of evaluators than in CST, and when they did, they found the defendant to be legally sane and thus subject to criminal prosecution.

Factors that led to increased agreement among the evaluators were 1) diagnosis of psychotic disorder, and 2) hospitalization in a psychiatric facility within six months prior to the evaluation. Factors that led to increased disagreement among the evaluators were 1) when the defendant had been abusing substances, and 2) when the defendant had committed a violent felony.

Conditional release: Unanimous agreement rates among evaluators were lowest of all three types of evaluations. In 53.2% of cases there was unanimous agreement; nearly 90% found that the defendant was ready for conditional release. Judges were most likely to overrule the majority opinion of evaluators in these cases keeping the patient hospitalized, apparently choosing to err on the side of caution.

Little statutory guidance in Hawaii makes the issue of conditional release – that involves whether or not the respondent can "be safely managed in the community" – much less clear than CST or criminal responsibility.

Significantly, of the 21 potentially relevant factors that Gowensmith, Murrie, and Boccaccini considered to be important in a conditional release evaluation, substantial agreement was found in only one: past violence. None of the other factors were endorsed by more than half of the evaluators. Moreover, evaluators were split on how to interpret the statute.

The study raises the following questions and implications:

Evaluators are not interchangeable. Expect to find disagreement among evaluators albeit levels of agreement were significantly better than chance. Agreement was greatest in CST.

Agreement is greater when the legal question is more straightforward and well defined.

Gowensmith, Murrie, and Boccaccini recommend that a second opinion be sought in cases that are complex.

How applicable are the findings to other jurisdictions? Evaluators comply with jurisdiction-specific requirements. What might reliability look like elsewhere?

How are judges' rulings influenced by the reports of evaluators? What other factors are weighed?

How do evaluators go about performing an evaluation? What factors do they consider to be important? What does their assessment consist of? What are the "best practices" or guidelines established by the profession?

Judges tend to follow the evaluators' opinions. The procedures employed by evaluators may shed more light. Thus, when there is disagreement, it is advisable to scrutinize the procedures, data employed, and to examine the evaluators' work product.

How can the procedure(s) in conditional release evaluations, the most problematic of three forensic contexts examined by Gowensmith, Murrie, and Boccaccini, be improved?

In their earlier study, Gowensmith, Murrie, and Boccaccini noted that a small percentage of evaluators (14.3%) used formal competency assessment measures. Why do they or don't they use measures? And when employed, which ones are used?

What weight can be attributed in different contexts to static factors (more applicable in criminal responsibility cases and providing a window into the past in conditional release) and dynamic factors (more applicable in CST and current functioning in conditional release)?

Do evaluators use risk assessment tools in conditional release evaluations? Why or why not? Which ones when used?

Would reliability be improved by use of context-specific instruments?

Do evaluator characteristics and factors identified by Gowensmith, Murrie, and Boccaccini need to be examined further? Would the conclusions extend to other jurisdictions and settings?

What can be done to improve the evaluation on the part of examiners in different forensic contexts?

Would mandated training and oversight improve reliability?

What policy implications can be drawn from this study in different jurisdictions and forensic contexts?

When evaluators are not court appointed, what is the likely impact of adversarial allegiance? How can this be controlled?

When mistakes are made, what are the consequences to the

defendant and to society in different jurisdictions and forensic contexts?

For attorneys who rely on the work of forensic experts, it behooves them to know their background, training, evaluation methodology, and experience and knowledge of the applicable law and statutes.

Forensic examiners need to remain up to date with the literature, evidence-based practice, and know the applicable law and statutes of the jurisdiction they work in. It also behooves them to routinely self-assess potential biases.

It may be that describing evaluation procedures and methodology more fully in forensic reports will add greater clarity to the judge to assist in making a ruling.

References
Gowensmith, W., Murrie, D.C., & Boccaccini, M.T. (2012). Field reliability of competency to stand trial evaluations: How often do evaluators agree, and what do judges decide when evaluators disagree? Law and Human Behavior, 36, 130–139.

After reading the reactions to their paper, the authors decided to issue a final comment.

Dr. Aranda raises several insightful questions about our research and the context for its findings. Although space precludes us from answering each of his questions, please allow us a brief moment to discuss some additional research that addresses his major themes.

First, Dr. Aranda wonders about how well this data generalizes to other jurisdictions and settings. The reliability values we found appear comparable to one of only a few other "real world" reliability studies (Skeem & Golding, 1998), though far more studies of this sort are needed. We are also researching additional settings and states to consider how our results generalize. Specifically, we are conducting additional research in multiple states on the decision-making of both the judges *and* the evaluators in forensic psychological assessments. What factors do mental health professionals prioritize in these types of cases? Do those comport with the factors that judges and attorneys view as most important? Does the state or setting matter? Some early trends are emerging, and we look forward to having more answers soon.

Second, Dr. Aranda poses questions about how to improve reliability and validity in forensic mental health evaluations. Of course there is no one easy answer. We have some evidence that the overall quality of forensic evaluations themselves has room for improvement (see Nguyen, Acklin, Fuger, Gowensmith, & Ignacio, 2011 for more information). We suspect that the largest improvements in reliability, validity and quality of forensic evaluations are likely to come from simply following the already-established standards in the field. We are working with several states to improve their evaluator certification processes and to ensure that best practices are infused into training and education for forensic evaluators. We must also work with the legal system as well to ensure that both legal and mental health audiences are well-informed about the most powerful factors to consider in various forensic cases, and the best ways to scrutinize forensic evaluations.

Finally, Dr. Aranda mentions the subject of adversarial allegiance. In contrast to our studies in Hawaii, where evaluators are appointed by the judge and presumed to be neutral experts, many jurisdictions let the defense and prosecution retain their own experts. Of course this raises questions about whether those experts can ever be impartial. This concept of "adversarial allegiance" continues to be a focus of our research, and we have found that opinions of mental health experts can differ depending on the side from which they were retained (please see Murrie et al, 2008; 2009 for more information).

We appreciate all of the reviewers' commentary and questions. As they suggest, a comprehensive understanding of forensic evaluations requires examining the evaluations, the evaluators, and the justice system in which they work. We have begun this process, and we have found some provocative results, but there is much work left to do.

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

Murrie, D.C., Boccaccini, M.T., Turner, D., Meeks, M., Woods, C. & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation?Psychology, Public Policy, and Law, 15,19–53. doi: 10.1037/a0014897

Nguyen, A. H., Acklin, M. A., Fuger, K., Gowensmith, W. N. & Ignacio, L. A. (2011). Freedom in paradise: Quality of conditional release reports submitted to the Hawaii judiciary. International Journal of Law and Psychiatry, 34, 341–348.

Skeem, J. L. & Golding, S. G. (1998). Community examiners' evaluations of competence to stand trial: Common problems and suggestions for improvement. Professional Psychology: Research and Practice, 29, 357–367.