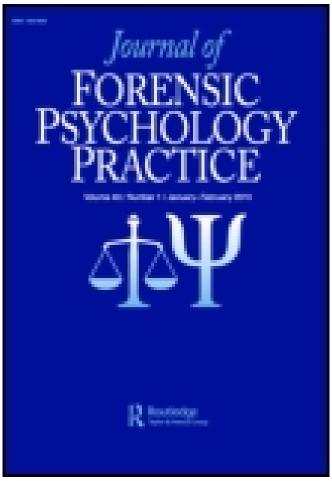

Click for updates

Journal of Forensic Psychology Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/wfpp20

Examiner Agreement and Judicial Consensus in Forensic Mental Health Evaluations

Marvin W. Acklin PhD[a], Kristen Fuger PhD[b] & William Gowensmith PhD[c]

[a] Department of Psychiatry, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, Hawaii

[b] Independent Practice, Honolulu, Hawaii

[c] Graduate School of Professional Psychology, University of Denver, Denver, Colorado

Published online: 04 Aug 2015.

To cite this article: Marvin W. Acklin PhD, Kristen Fuger PhD & William Gowensmith PhD (2015) Examiner Agreement and Judicial Consensus in Forensic Mental Health Evaluations, Journal of Forensic Psychology Practice, 15:4, 318-343, DOI: 10.1080/15228932.2015.1051447

To link to this article: http://dx.doi.org/10.1080/15228932.2015.1051447

Journal of Forensic Psychology Practice, 15:318–343, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1522-8932 print/1522-9092 online
DOI: 10.1080/15228932.2015.1051447

Routledge
Taylor & Francis Group

Examiner Agreement and Judicial Consensus in Forensic Mental Health Evaluations

MARVIN W. ACKLIN, PhD
Department of Psychiatry, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, Hawaii

KRISTEN FUGER, PhD
Independent Practice, Honolulu, Hawaii

WILLIAM GOWENSMITH, PhD
Graduate School of Professional Psychology, University of Denver, Denver, Colorado

*The reliability of forensic methods continues to be controversial. Hawaii is unique in utilizing a three-panel system for evaluating criminal defendants for competency to stand trial (CST), not guilty by reason of insanity (NGRI), and postacquittal conditional release (CR). The study examined independent forensic reports with judicial determinations to assess examiner agreement and judicial consensus. Examinees (*N = *450) were defendants charged with felony offenses. Three groups of examiners conducted independent forensic mental health evaluations: community-based psychiatrists, community-based psychologists, and psychologists employed by the Courts & Corrections branch of the Hawaii Department of Health. Five classes of reliability estimators were examined in a noncrossed data measurement design. The study examined field reliability of CST, NGRI, and CR as operationalized psycholegal constructs. Overall, findings revealed wide variability in examiner consensus and agreement between examiners and judges, depending on type of examination. Factors associated with examiner disagreement are discussed. Findings are similar to field reliability for other types of complex decision making. Procedural standardization, application of structured professional methods, use for forensic assessment instruments, and de-bias assessment are recommended to improve the quality of forensic mental health opinions.*

Address correspondence to Marvin W. Acklin, PhD, Department of Psychiatry, John A. Burns School of Medicine, University of Hawaii at Manoa, 850 W. Hind Drive, Suite 203, Honolulu, HI 96821. E-mail: acklin@hawaii.edu

KEYWORDS forensic psychology, judicial decision-making, forensic interrater reliability

The reliability of methods in all fields of forensic science has received scrutiny (e.g., Miller, Kimonis, Otto, Kline, & Wasserman, 2012; National Research Council, 2009; Neal & Grisso, 2014). "Partisan allegiance" (Murrie, Boccaccini, Guarnera, & Rufino, 2013; Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009) has emerged as a concern in forensic behavioral science, defined as an extreme form of bias that favors the party who retained the forensic expert. Murrie and colleagues (2013) found that "risk scores assigned by prosecution and defense experts showed a clear pattern of adversarial allegiance" (p. 5). They note that even though their study examined only one kind of evaluation (i.e., sexual recidivism risk), "there is little reason to believe that this is the only kind of forensic psychological evaluation or forensic-science procedure vulnerable to allegiance effects" (p. 8).

Hawaii is unique in the United States in the manner in which mental health professionals are appointed by the court to perform forensic mental health evaluations (Mossman et al., 2007). The statutory basis for mental evaluations and procedures in the State of Hawaii is found in Hawaii Revised Statute (HRS), Chapter 704. When a defendant has been charged with a felony offense and a question of fitness to stand trial (CST), criminal responsibility (NGRI), or postacquittal (CR) release is raised, Hawaii statues provide for the appointment of a "three-panel" of mental health professionals to independently examine the defendant and report their opinions to the court. The statutes mandate a description of the examination, diagnosis of the physical or mental condition of the defendant, and opinions of the ultimate psycholegal construct in question.

Three types of examiners are appointed to conduct independent evaluations: community-based psychiatrists (PSY), community-based psychologists (CBP), and psychologists employed by the Courts & Corrections branch of the Department of Health (DOH). Community-based examiners are licensed to practice in their respective disciplines and are appointed from a court-approved list. This situation sets up an ideal naturalistic laboratory to study forensic decision-making processes in examiners and judges. The current study examined agreement and consensus in CST, NGRI, and postacquittal CR evaluations, and assessed Hawaii's system for forensic classification (Altman & Royston, 2000) for the determination of common but high-stakes psycholegal constructs.

The impetus for this study originated in separate CST, NGRI, and CR report quality studies involving 50 cases and 150 reports ($N = 450$) using a quality coding instrument (Fuger, Acklin, Nguyen, Ignacio, & Gowensmith, 2014; Nguyen, Acklin, Fuger, Gowensmith, & Ignacio, 2011; Robinson & Acklin, 2010). These studies found mediocre quality indices and low or absent use of forensic assessment instruments (FAIs; Otto & Heilbrun, 2002) in CST, NGRI, and postacquittal CR examinations. Preliminary interrater agreement for the report quality studies found considerable variability and, in many cases, substandard levels of agreement, according to published guidelines for interpretation of reliability estimators (e.g., Cicchetti, 1994; Fleiss, 1981; Landis & Koch, 1977; Reitveld & van Hout, 1993).

Gowensmith and colleagues examined 216 three-panel CST reports drawn from the same Hawaii population, noting that their study permitted an investigation of field reliability (Gowensmith, Murrie, & Boccaccini, 2012).[1] They found moderate levels of agreement between examiners. Seventy-one percent of the cases yielded consensus CST opinions with Cohen's kappa = .65. They emphasized the importance of examiner training in improving interrater reliability. Gowensmith and colleagues also examined 483 NGRI reports drawn from the same Hawaii population (Gowensmith, Murrie, & Boccaccini, 2013). They found consensus among the three opinions in only 55% of cases, with judges agreeing with the panel 91% of the time. They did not report coefficients of agreement in their study, but did remark that "reliability among practicing forensic examiners addressing legal sanity may be poorer than the field has tended to assume" (p. 98).

THE LOGIC BEHIND RELIABILITY

Cohen described the fundamental logic of interrater reliability studies (Cohen, 1960), writing that because categorizing of the units to be agreed on "is a consequence of some complex judgment process performed by a 'two legged meter,' it becomes important to determine the extent to which these judgments are reproducible, i.e., reliable" (p. 37). When two or more judges independently categorize a sample of units, "this quite parallels in its logic the concept of the coefficient of equivalence used with tests—the judges are analogous to alternate forms, and the nominal data are analogous to scores" (p. 38). In short, independent examiners may be viewed as "tests" for the purposes of reliability assessment ("parallel tests" in classical test theory; Suen & Ary, 1989). Haynes and colleagues (2011) defined "reliability as the degree to which measures taken by similar or parallel instruments, by different observers, or at different points of time yield the same or similar results" (p. 32). Suen and Ary (1989) stated that "reliability of the data is the result of using particular observers and a particular coding system under a particular set of circumstances" (p. 99). "Evidence of reliability of behavioral

[1] Field reliability involves "the applied measurement of natural behavior in a non-laboratory setting uninfluenced by knowledge that the coding will subsequently be reviewed" (Greg Meyer, personal communication, September 29, 2014).

observation data is needed to demonstrate to an external audience that the data reflect reality" (Suen & Ary, 1989, p. 100). Observer reliability reflects "consistency" amongst judgments (Suen & Ary, 1989), "interchangeability" of observers (Bakeman & Quera, 2011), and reproducibility of results (Cohen, 1960), and can be considered an index of the degree to which error is absent from the data (Nunnally, 1978). Kraemer, Periyakoil, and Noda (2002) wrote that "the reliability of a measure, so defined, indicates how reproducible that measure will be" and "how much error will be introduced into clinical decision-making based on the measure" (2002, p. 2112). Funder (1990) observed that "the study of accuracy in judgment is exactly the same thing as the study of measurement validity, where the measurements being validated are interpersonal judgments" (p. 208).

In a recent study, Mossman (2013) provided conceptual and empirical tools for accounting for variability in forensic judgments, including examiner bias and the inevitability of random error. Mossman utilized a data set drawn from the same population and a subset of the same examiners as our study (Gowensmith et al., 2012), on the issue of CST. In a situation involving binary judgments it is assumed that the construct is dimensional and has an underlying continuum. Outer areas of the continuum represent areas of the construct with low levels of disagreement. The middle zone represents an ambiguous, "hardest to decide zone" (Mossman, 2013) where disagreements and errors are more likely to occur.

Using the Hawaii CST data, Mossman posited four hypothetical "decision thresholds" (Swets, Dawes, & Monahan, 2000) that examiners may utilize in their judgments and illustrated the resulting impact on accuracy and error rates: most probable status, mild bias, clear and convincing bias, and fuzzy zone. These examiner decision thresholds are points along the decision axis and are associated with particular values of sensitivity and specificity. An examiner's opinion in a particular case reflects the examiner's implicit or explicit judgments about the costs and benefits of incorrect or correct judgments, combined with the examiner's thinking about the location of a particular case along the decision axis.

CRITIQUES OF TRADITIONAL RELIABILITY ESTIMATORS

A large literature on kappa and other presence–absence measures is found in a variety of sciences, such as wildlife biology, plant ecology, clinical medicine, and weather forecasting. Methods of agreement calculation and interpretation are controversial (Pontius & Millones, 2011). Some commentators have advocated the abandonment of kappa altogether given its idiosyncrasies (Pontius & Millones, 2011). Bakeman and colleagues (1997) and others argued that conventional guidelines for interpreting agreement coefficients are misleading (e.g., Cicchetti, 1994; Fleiss, 1981; Kraemer et al., 2002; Landis & Koch, 1977; Rietveld & Van Hout, 1993). Factors other than actual level of agreement influence the magnitude of Cohen's kappa, including prevalence (whether codes are equiprobable) and bias (whether marginal probabilities are equal or variable; Sim & Wright, 2005). Magnitude of kappa also is dependent upon the number of codes. Bakeman and colleagues note that, "There is no one value of kappa that can be regarded as universally acceptable" (p. 71). Furthermore, "it is misleading to claim, for example, that kappa values of .80 and above are acceptable whereas values below are not" (Bakeman & Quera, 2011, p. 83).

Other criticisms of kappa emerged in the 1970s, noting, for example, the limitations of kappa in multiple-rater situations (Conger, 1980). Some commentators have proposed alternatives that address the limitations of kappa (e.g., DiEugenio & Glass, 2004; Hallgren, 2012), including biases introduced by idiosyncratic properties of the data (prevalence and bias), suitability for fully crossed designs for more than two coders, and proposed alternatives including Fleiss's kappa (1981), Light's average kappa (1971), and Krippendorff's alpha (Hayes 2013; Hayes & Krippendorff, 2007). These methods accommodate multiple raters, noncrossed designs, and missing data. In the current study, comparisons between agreement and reliability estimators were examined, including multirater modifications of kappa.

In contrast to traditional interrater agreement and reliability estimates, Krippendorff's alpha (KALPHA) assesses the reliability of the construct being assessed, reflecting "the combination of the variable description and the categories, the information and background in the code book, and the instructions given during training" (De Swert, 2012, p. 5). In the current study, KALPHA was calculated to permit the evaluation of a forensic classification model for examiners alone and with ultimate judicial determination (Altman & Royston, 2000). In this context, judicial determination is not considered a separate or independent rating (i.e., a gold standard), but the outcome of the investigative or decision-making model.

In summary, recent interest in the reliability of forensic methods triggered our investigation of forensic mental health judgments and examiner and judicial decision-making processes. Prominent commentators have asserted that judges abdicate their independence to forensic examiners (Cox & Zapf, 2004; Zapf, Hubbard, Cooper, Wheeles, & Ronan, 2004). Overall, the study attempted to assess a judicial and forensic classification model (Altman & Royston, 2000), and addressed some of the conceptual and psychometric complexities in the application of reliability estimators.

THE CURRENT STUDY

The following research questions were investigated:

1. What level of performance in forensic judgments is reflected in agreement and reliability coefficients?

2. Does level of agreement and consensus differ by type of evaluation?
3. Are levels of complexity in types of evaluations empirically distinguishable?
4. Is professional discipline of examiner a relevant empirical factor in forensic evaluations?
5. To what degree are judicial determinations independent?
6. How should error be conceptualized in forensic mental health decision making?
7. Do these findings raise concerns about the quality of forensic decision making in legal situations where there is a high cost of errors?

METHOD

Three CST, NGRI, and CR report quality studies of 50 cases were coded utilizing three groups of examiners ($N = 450$ reports). Selection criteria included reports completed since 2006, all three reports present, and a judicial determination for each case. All reports were in the public domain and selected from files in the Circuit Court of the First Circuit, Oahu. Only reports submitted to the court after 2006 were coded to reflect recent standards and literature trends in forensic report writing. Examinees were felony defendants where a mental health issue pertinent to CST, NGRI, or postacquittal CR was raised by the defendant or *sua sponte* by the court. Examinations were typically conducted in the local jail or state hospital. Panel examiners included 7 community-based psychiatrists (PSY), 11 community-based psychologists (CBP), and 7 Courts & Corrections psychologists (DOH) employed by the State of Hawaii. These examiners are appointed by judges from a court-approved list. Examiners are licensed in their professional disciplines and are required to attend an annual training conference. Nine Circuit Court judges provided judicial determinations. Coders rated agreement between examiners and also between examiners and judicial determination. In case of CST or NGRI, court orders mandated ultimate opinions as to the defendant's capacity to understand the proceedings and assist in the defendant's own defense; or an opinion as to the extent, if any, to which the capacity of the defendant to appreciate the wrongfulness of the conduct or to conform the conduct to the requirements of law was impaired at the time of the conduct alleged (ALI insanity standard; HRS 704–404). For postacquittal CR, the statutes require ultimate opinions concerning whether the defendant is affected by a physical or mental disease, disorder, or defect and presents a risk of danger to self or others; also, whether the defendant is a proper subject for conditional release, presents a danger to self or others, and can be controlled adequately and given proper care, supervision, and treatment in a less restrictive setting than the state hospital. Similar to many jurisdictions, the majority of the evaluations of the CST and NGRI evaluations were combined (Chauhan, Warren, Kois, & Wellbeloved-Stone, 2015). Postacquittal CR evaluations were typically not combined.

Two coders used agreement, disagreement, and no opinion (Y/N/0) as response categories to establish the reliability of the coding system. Five classes of reliability estimators were calculated and compared between examiners and judicial determinations. Coders were graduate students in clinical psychology conducting doctoral dissertation research under supervision of the first author. Coders were trained to criteria under supervision of the first author. Coders trained to proficiency and demonstrated high levels of interrater agreement (Cohen's kappa > .90).

A priori sample size estimates were calculated using G*Power 3 (version 3.1.9.2 for Windows; Faul, Erdfelder, Lang, & Buchner, 2007). Setting parameters at alpha = 0.05, two-tailed test, and power at .80, a sample size of 82 was required to detect moderate effect sizes. Given the sample sizes used here (150 cases per analysis for CST, NGRI, and CR), our sample sizes had sufficient power. Post-hoc power estimates were calculated using G*Power 3, with alpha = 0.05, two-tailed test, moderate effect size, and sample size = 150. Obtained power was 0.96. Consequently the design had sufficient sensitivity to detect small, medium, and large effect sizes at sufficient levels of power.

Indices of agreement were calculated between examiners and judicial determination for CST, NGRI, and postacquittal CR. Given the mixed distribution of examiners and judges, the study utilized a "non-crossed" or "ill-structured" measurement design (ISMD; Hallgren, 2012; Nayaraynan, Greco, & Campbell, 2010; Putka, Le, McCloy, & Diaz, 2008). In contrast to fully crossed or nested designs, the ISMD utilized multiple raters where "different subjects are rated by different subsets of coders" (Hallgren, 2012, p. 3). Interrater agreement was estimated using Cohen's kappa, Fleiss's multirater kappa, and intraclass correlation coefficients (ICC (A, 1) = ICC_{abs}; two-way, absolute agreement, random effects model; McGraw & Wong, 1996). Absolute agreement ICCs are recommended over rater consistency ICC, "when judges are considered random effect, ICC_{abs} addresses the question of whether judges are interchangeable"; and "if we want to discriminate individuals (or sessions or codes) according to a certain criterion value, then we need to know whether observers discriminate identically (absolute agreement)" (McGraw & Wong, 1996, p. 91). In a 2 × 2 rating situation, kappa and ICC are equivalent (Bakeman & Quera, 2011). This study utilized an innovative reliability estimator, Krippendorff's alpha (KALPHA). Hayes & Krippendorff (2007) have clarified convincingly why Krippendorff's alpha should be the basic measure to apply for most researchers. Sample size, multiple (more than two) coders or missing data are not problematic for calculating KALPHA, and all measurement levels can be tested.

Interpretation of reliability coefficients has been long debated. Cicchetti and Sparrow (1981) proposed a conceptual rationale and guidelines for interpreting levels of clinical and practical significance. Cicchetti (1994) proposed a simple version of the guidelines introduced by Landis and Koch (1977). We adopt the Landis and Koch (1977) guidelines: < 0, poor agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement, accepting caveats about interpretive guidelines.

Ethics Statement

Each of the three studies that form the basis of this research was reviewed and approved by the Institutional Review of Argosy University, Hawaii campus.

Data Analysis

A priori sample size estimates were calculated using Cicchetti and Sparrow's formula (1981). Data were coded and analyzed using Microsoft Excel, Recal 3 (Freelon, 2010), Medcalc (http://www.medcalc.org/calc/diagnostic_test.php), Alan Fielding's online calculator for accuracy statistics (http://www.alanfielding.co.uk/multivar/accuracy.htm), and Hayes's SPSS macro (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html). Calculations used SPSS version 18 (Statistical Package for the Social Sciences, 2009). Recal3, an online calculator for the calculation of average pairwise percent agreement, was used for calculating Fleiss's multirater kappa, and average pairwise Cohen's kappa. Hayes's SPSS macro was utilized for calculation of Krippendorff's alpha (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html), including a bootstrapping procedure for cross validation.

RESULTS

The study examined agreement between forensic examiners and judges for CST, NGRI, and CR opinions. In our coding schema, "yes" codes indicate the presence of the condition: incompetency (IST), insanity (NGRI), and granted conditional release (CR). "No" codes indicate the absence of the condition: competent (not IST), sanity (not NGRI), and denied conditional release (not CR).

Table 1 presents summary forensic mental health opinions for 150 cases by type of evaluation. Examiners rated two of three defendants as CST, approximately 50% were NGRI, and almost two out of three as eligible for CR. A substantial portion of NGRI and CR reports, 12%, produced no opinions (see Table 1 for percentages for type of mental health opinions).

TABLE 1 Forensic Mental Health Examiner Opinions for 150 Cases/450 Reports

Study	Cases/Reports	Yes	No	No Opinion
CST	(n = 50/150)	30%/n = 45	68%/n = 102	2%/n = 3
NGRI	(n = 50/150)	47%/n = 71	41%/n = 61	12%/n = 18
CR	(n = 50/150)	61%/n = 91	27%/n = 41	12%/n = 18

Note. CST: Competency to Stand Trial. NGRI: Criminal Responsibility. CR: Postacquittal Release. Yes is the affirmative of the condition (incompetency, insanity, and granted conditional release). No codes indicate the negation of the condition (CST, not NGRI, and no CR).

Table 2 presents judicial CST, NGRI, and CR determinations for 148 cases. Judges ruled that approximately 2 out 3 defendants were CST, 6 out of 10 were not NGRI, and nearly 2 out of 3 were eligible for CR (see Table 2 for percentages of judicial determinations).

Table 3 presents forensic mental health opinions for 150 cases by professional discipline. Chi-square analyses revealed no consistent or significant differences among disciplines. These findings suggest that distinguishing examiners by professional discipline has no empirical justification (see Table 3 for opinions by professional discipline). A number of cases yielded "no opinion" findings by examiners. No opinion findings were lowest in CST (1), highest in CR (8), and intermediate in NGRI (5).

CST Results

Agreement Between Examiners

Examining three examiners (PSY, CBP, DOH), the following findings document levels of agreement. Average pairwise percent agreement was 78%. Average pairwise percent agreement, comparing PSY to CBP: 70%, PSY to DOH: 82%, and CBP to DOH: 82%. Fleiss's multirater kappa for all examiners was 0.508 for CST. Average pairwise Cohen's kappa was 0.511. Average pairwise Cohen's kappa between PSY and CBP was .0.343, PSY and DOH was 0.593, and CBP to DOH was 0.597. All of these coefficients indicate "fair to moderate" levels of agreement (Landis & Koch, 1977).

TABLE 2 Judicial Determination for 148 Cases/444 Reports

	Study Cases/Reports	Yes	No
CST	(n = 50/150)	32%/n = 16	68%/n = 34
NGRI	(n = 49/147)	38%/n = 19	60%/n = 30
CR	(n = 49/147)	64%/n = 34	30%/n = 15

Note. CST: Competency to Stand Trial. NGRI: Criminal Responsibility. CR: Post-acquittal Release. Yes is the affirmative of the condition (incompetency, insanity, and granted conditional release). No codes indicate the negation of the condition (CST, not NGRI, and no CR).

TABLE 3 Forensic Opinions for 150 Cases/450 Reports by Professional Discipline

	Study Cases/Reports	Yes	No	No Opinion
CST	(*n* = 50/150)	34%/*n* = 17	66%/*n* = 33	*n* = 0
NGRI	(*n* = 50/150)	56%/*n* = 28	32%/*n* = 16	12%/*n* = 6
CR	(*n* = 50/150)	70%/*n* = 35	30%/*n* = 15	*n* = 0
CBP				
CST	(*n* = 50/150)	32%/*n* = 16	68%/*n* = 34	*n* = 0
NGRI	(*n* = 50/150)	46%/*n* = 23	46%/*n* = 23	8%/*n* = 4
CR	(*n* = 50/150)	62%/*n* = 31	18%/*n* = 9	20%/*n* = 10
DOH				
CST	(*n* = 50/150)	24%/n = 12	70%/*n* = 35	6%/*n* = 3
NGRI	(*n* = 50/150)	40%/*n* = 20	44%/*n* = 22	16%/*n* = 8
CR	(*n* = 50/150)	50%/*n* = 25	34%/*n* = 17	16%/*n* = 8

Note. CST: Competency to Stand Trial. NGRI: Criminal Responsibility. CR: Postacquittal Release. Yes is the affirmative of the condition (incompetency, insanity, and granted conditional release). No codes indicate the negation of the condition (CST, not NGRI, and no CR). PSY: community-based psychiatrists. CBP: community-based psychologists. DOH: Courts & Corrections psychologists.

Agreement Between Examiners and Judges

When judges were included as fourth raters, average pairwise percent agreement was 81%. Average pairwise percent agreement, comparing JUD to DOH, was 82%, JUD to CBP was 88%, JUD to PSY was 82%, PSY to DOH was 70%, PSY to CBP was 82%, and CBP and DOH was 82%. Fleiss's kappa for all examiners with judges was 0.572. Average pairwise Cohen's kappa between comparing JUD to DOH was 0.597, JUD to CBP was 0.724, JUD to PSY was 0.593, PSY to DOH was 0.343, PSY to CBP was 0.593, and CBP to DOH was 0.597. These coefficients indicate "moderate" to "substantial" agreement (Landis & Koch, 1977).

Level of CST agreement was calculated using ICC (ICC_{abs}; 2 way, absolute agreement, random effects model; McGraw & Wong, 1996), with 95% confidence intervals were calculated. Level of agreement between PSY and CBP was .598 95% CI [.384, .751], between PSY and DOH .309 95% CI [.047–.534], and between CBP and DOH was .411 95% CI [.159, .614]. CST ICCs ranged from 0.309 to 0.598. Mean ICC for CST = 0.439 indicated "moderate" agreement. These coefficients indicate "fair" to "substantial" level of agreement (Landis & Koch, 1977).

Table 4 presents aggregated CST agreement coefficients for examiners and judges. With the exception of CBP, which fell into the "slight agreement" range, all of the other ICC were in "moderate" range (Landis & Koch, 1977; see Table 4 for aggregated ICC agreement coefficients for examiners and judges).

Table 5 presents judicial CST determinations by level of consensus. Judges utilized panel consensus (both consensus of 2 [CNS2] or 3 examiners [CNS3]) as a basis for CST decision-making. Of special interest were judicial

TABLE 4 Aggregated Agreement Coefficients Between All Examiners and Judges for CST, NGRI, and CR

Examiner	ICC	*p*	*F*	*df*	95% CI
CST	.567	<.001	3.624	149	.449–.667
NGRI	.509	<.001	3.074	146	.379–.691
CR	.264	≤.01	1.71	146	.1070–.408

Note. ICC_{abs}; one-way, absolute agreement, random effects model. PSY: community-based psychiatrists. CBP: community-based psychologists. DOH: Courts & Corrections psychologists. Jud 4th rater: judge as fourth rater.

TABLE 5 CST Examiner Consensus and Judicial Determination (N = 50)

Examiner	*N*	%	Judge
CNS3Yes (IST)	8	16	8 Yes 100%
CNS2Yes (IST)	6	12	5 Yes 83% 1 No 17%
CNS3No (CST)	26	52	25 No 96% 1 Yes 4%
CNS2No (CST)	9	18	7 No 78% 2 Yes 22%
CNS3Yes/No	34	68	
CNS2/3 Yes/No	49	98	
No CNS	1*	2	1 No

Note. Code: 1 = No (CST). 2 = Yes (IST). *No CNS = 1–2–0.

determinations when the panel provided “no consensus” or split opinions. For the single “no consensus” CST case (coded 1–2–0 by examiners), the judge ruled no (CST). Here the judge tended to maximize panel consensus in the judicial determination of CST (see Table 5 for judicial determinations by level of consensus). Levels of consensus for examiners and judges suggest that the CST construct may be evaluated with a moderate degree of reliability.

NGRI Results

Agreement Between Examiners

Average pairwise percent agreement was 63%. Average pairwise percent agreement, comparing PSY to CBP, was 58%, PSY to DOH: 68%, and CBP to DOH: 64%. Fleiss’s kappa for all examiners was 0.385 for NGRI. Average pairwise Cohen’s kappa was 0.391. Average pairwise Cohen’s kappa between PSY and CBP was .0.318, PSY and DOH was 0.454, and CBP to DOH was 0.401. All of these reliability coefficients fall into “fair” range of agreement (Landis & Koch, 1977) and reflect lower levels of agreement when compared to CST decisions.

When judges are included as fourth raters, average pairwise percent agreement was 67%. Average pairwise percent agreement comparing JUD to DOH was 78%, JUD to CBP was 68%, JUD to PSY was 64%, PSY to DOH was 58%, PSY to CBP was 68%, and CBP and DOH was 64%.

NGRI Agreement Between Examiners and Judges

Fleiss's kappa for all examiners with Judge was 0.427. Average pairwise kappa for NGRI was 0.511. Average pairwise Cohen's kappa between comparing JUD to DOH was 0.597, JUD to CBP was 0.724, JUD to PSY was 0.593, PSY to DOH was 0.343, PSY to CBP was 0.593, and CBP to DOH was 0.597. These coefficients indicate "fair" to "moderate" level of agreement (Landis & Koch, 1977).

Level of NGRI reliability was calculated using ICC (ICC_{abs}; two-way, absolute agreement, random effects model; McGraw & Wong, 1996), with 95% confidence intervals. Level of agreement between PSY and CBP was .576 95% CI [.357, .735], between PSY and DOH .422 95% CI [.172–.622], and between CBP and DOH was .547 95% CI [.323, .714]. All of the ICCs fall into the "moderate" range of agreement (Landis & Koch, 1977).

Level of NGRI agreement for examiners and judges was calculated using ICC (ICC_{abs}; two-way, absolute agreement, random effects model; McGraw & Wong, 1996), with 95% confidence intervals (see Table 4 for NGRI ICC agreement coefficients for examiners and judges). Not Guilty by Reason of Insanity ICCs ranged from 0.422 to -0.576. Mean ICC for NGRI = 0.515, in the "moderate" range of agreement (Landis & Koch, 1977).

Table 6 presents judicial NGRI determinations by level of consensus. Only 38% of cases achieved CNS3 (consensus of three examiners). Eighty-two percent of cases were rated as CNS2 (consensus of at least two examiners) and CNS3. Judges tended to favor CNS3 examiners opinions (100% consensus). Judges tended to diverge from panel in CNS2 cases.

TABLE 6 NGRI Examiner Consensus and Judicial Determination ($N = 49$)

Examiner	*N*	%	Judge
CNS3 Yes (NGRI)	14	28	13 Yes 93% 1 No 7%
CNS2 Yes (NGRI)	8	16	4 Yes 50% 4 No 50%
CNS3 No (not NGRI)	7	14	7 No 100%
CNS2 No (not NGRI)	15	30	13 No 87% 2 Yes 13%
CNS3 Yes/No	21	42	
CNS2/3 Yes/No	44	78	
No CNS	5*	10	4 No 80% 1Yes_20%_

Note. Code: 1 = No (not NGRI). 2 = Yes (NGRI). *No CNS: 1 = 0–0–2; 1 = 2–1–0; 1 = 0–0–0; 1 = 0-0-0; 2 = 2–0-0.

Of special interest were judicial determinations when the panel provided "no consensus" or split opinions. For the five no consensus NGRI cases, judges ruled no (not NGRI) on 0–0–2, no (not NGRI) on a split opinion (2–1–0), no (not NGRI) on the two no opinion (0–0–0) cases, and yes (NGRI) on 2–0–0. This evidence indicates that judges utilized consensus whenever possible, but in many cases judicial determinations diverged from three-panel consensus (see Table 5 for judicial NGRI determinations by level of consensus). These findings suggest that the determination of NGRI is comparatively more difficult than CST decision making.

CR Results

Agreement Between Examiners

Average pairwise percent agreement: 55%. Average pairwise percent agreement, comparing PSY to CBP was 60%; PSY to DOH: 60%; and CBP to DOH: 46%. Fleiss's kappa for all examiners was 0.177 for CR. Average pairwise Cohen's kappa was 0.195. Average pairwise Cohen's kappa between PSY and CBP was .0.270, PSY and DOH was 0.219, and CBP to DOH was 0.095. These coefficients reflect "slight" to "fair" agreement (Landis & Koch, 1977).

Agreement Between Examiners and Judges

When judges were included as fourth raters, average pairwise percent agreement was 61%. Average pairwise percent agreement, comparing JUD to DOH, was 62%, JUD to CBP was 56%, JUD to PSY was 84%, PSY to DOH was 60%, PSY to CBP was 60%, and CBP and DOH was 46%. Fleiss's kappa for all examiners with judges was 0.259. Average pairwise Cohen's kappa was 0.281. When between comparing, JUD to DOH was 0.315, JUD to CBP was 0.154, JUD to PSY was 0.631, PSY to DOH was 0.270, PSY to CBP was 0.219, and CBP to DOH was 0.095. The highly variable coefficients range from "fair" to "substantial" levels of agreement (Landis & Koch, 1977). This wide variability of agreement and consensus reflects the difficulty of making CR determinations.

Level of CR agreement between examiners using ICC (ICC_{abs}; two-way, absolute agreement, random effects model; McGraw & Wong, 1996), with 95% confidence intervals were calculated. Levels of CR agreement all fell into the "slight" agreement range (Landis & Koch, 1977). Level of agreement between PSY and CBP was .188 95% CI [-.71, .430], between PSY and DOH was .143 95% CI [-.99, .382], and between CBP and DOH was .098 95% CI [−188, .366].

Table 4 presents CR agreement data between examiners and judges using ICC (ICC_{abs}; two-way, absolute agreement, random effects model;

TABLE 7 CR Examiner Consensus and Judicial Determination ($N = 49$)

Examiner	N	%	Judge
CNS3 Yes (CR)	17	34	17 Yes 100%
CNS2 Yes (CR)	13	26	9 Yes 69% 4 No 31%
CNS3 No (no CR)	2	4	2 No 100%
CNS2 No (no CR)	9	18	6 No 67% 3 Yes 33%
CNS3 Yes/No	19	38	
CNS2/3 Yes/No	41	82	
No CNS	8*	16	5 Yes 63% 3 No 27%

Note. Code: 1 = No (no CR). 2 = Yes (CR). *For the eight no consensus CR cases, 1 = 1–2–0; 1–0–2; and 2–0–1; 2 = 2–1-0; 1-0–2; 2–0–1; 2-0–1; and 2-0-0.

McGraw & Wong, 1996), with 95% confidence intervals. CR ICCs ranged from 0.098 to .188. Using the Landis and Koch (1977) guidelines, agreement between all examiners was in the "slight" range. Mean ICC for CR = 0.143 indicated "slight" level of agreement (see Table 8 for CR agreement coefficients between examiners and judges). These data provide further evidence of the difficulty in reaching consensus in CR decision making.

We examined the manner in which judges utilized consensus among examiners as a basis for CR decision making. Of special interest were judicial determinations when the panel provided no consensus or split opinions. Table 7 presents judicial CR determinations by level of consensus. The CR group had the highest number of no opinion or split opinion cases. For the eight no-consensus CR cases, the judge ruled no (no CR) on the 1–2–0, 1–0-2, and 2–0-1 cases, and yes (CR) in five split-decision cases of 2–1–0,1–0–2,2–0–1, 2–0-1, and 2–0–0 (see Table 7 for judicial CR determinations by level of consensus). Overall, these data suggest the difficulties CR decisions pose for examiners and judges.

Krippendorff Reliability Estimates

Using an alternative model for examining CST, NGRI, and CR reliability, Table 8 presents Krippendorff's alpha with bootstrapping algorithm. KALPHA provides reliability evidence for the psycholegal construct. Results demonstrate three reliability models: single examiners only, all examiners, and judges as fourth raters. Bootstrapping allows for statistical inferences about the population when sample sizes are small and an estimation of cross-validation efficiency when the model is applied to new data. The bootstrapping procedure models 1,000 samples. Krippendorff's alphas are interpreted similarly to other reliability coefficients. These data indicate "substantial" level of agreement for CST, "fair" agreement for NGRI, and

TABLE 8 Cross-Validated Forensic Classification Model With Kappa, Examiner, and Judicial Determinations

	KALPHA			KALPHA		
Fleiss's	k	w/o judge	95% CI	w/judge	95%CI	(B)
CST single examiner						
PSY				.59	[.32, .82]	.55
CBP				.72	[.49, .90]	.13
DOH				.59	[.33, .82]	.55
All examiners	.51	.51	[.36, .65]	—	—	.85
Complete model	—	—	—	.57	[.45, .69]	.66
NGRI single examiner						
PSY				.30	[.07, .53]	.99
CBP				.41	[.15, .63]	.95
DOH				.61	[.40, .79]	.42
All examiners	.38	.34	[.220, .465]	—	—	1.000
Complete model	—	—	—	.39	[.30, .48]	1.000
CR single examiner						
PSY				.63	[.40, .86]	.40
CBP				.14	[−.13, .41]	.99
DOH				.30	[.05, .56]	.99
All examiners	.18	.18	[.036, .330]	—	—	1.000
Complete model	—	—	—	.26	[.14, .38]	1.000

Note. KALPHA may be interpreted like kappa. Krippendorff states that KALPHA < .80 is unacceptable for variables to be used in research. (B) Bootstrap = 1,000 cases. Probability of achieving KALPHA > .60.

"fair" agreement for CR (Landis & Koch, 1977) with wide confidence intervals, and uniformly unreliable bootstrapping findings (see Table 8 for forensic classification model with cross-validation). Krippendorff opined that KALPHA coefficients less than .80 indicate unsatisfactory reliability for research purposes (De Swert, 2012). Krippendorff's recommended .80 criterion for acceptable levels of reliability may be too stringent for field decision making.

Summarizing these findings, agreement between examiners tended to vary widely, and to differ between types of examination, with CST highest, CR lowest, and NGRI intermediate levels of agreement. A majority of the coefficients demonstrated a "moderate" level of agreement (ICC = 0.41–0.60; Landis & Koch, 1977) with substantially weaker levels of agreement for NGRI and CR. Coefficients of determination for ICCs at or below the .50 range indicate that less than half of the variance could be attributed to the defendant's actual legal standing; Murrie et al. (2009) found that *opposing* experts obtained ICC of .42 on the PCL-R. When examining the forensic classification model as a whole, the model appears to produce results that are highly variable, that is, inconsistent, ranging from slight to moderate consistency and reproducibility (Cohen, 1960), depending on type of examination.

Given the range of examiner variability and levels of agreement for examiners, judges tended to concur with majority opinions whenever possible, but not always. In most cases, where there were split or no consensus opinions, judges tended to make conservative choices, demonstrating mixed opinions. The use of consensus appears to be an intuitive rationale, which demonstrates the psychometric dictum that adding raters improves reliability. There is also a likelihood that judges utilized other information concerning defendants not available to examiners. These data suggest that judges demonstrate a moderate degree of independence in their determinations.

DISCUSSION

Forensic examiner performance varied widely with levels of agreement depending upon type of evaluation: generally moderate range for CST, fair to moderate in NGRI, and slight to fair range for CR. Many of the NGRI reliability coefficients were marginal and similar to CR. Standard guidelines for interpretations of agreement and reliability coefficients suggest that the examiner performance in our CST, NGRI, and postacquittal CR studies were typically below the desirable range (e.g., Cicchetti, 1994; Fleiss, 1981; Landis & Koch, 1977; Rietveld & Van Hout, 1993). The findings are similar to Gowensmith and colleagues (Gowensmith, Murrie, & Boccaccini, 2010; Gowensmith et al., 2013). The data reflect unsystematic variability between examiners and between examiners and judges, with the best performance in CST, the poorest performance in postacquittal CR, and intermediate performance in NGRI evaluations. Unsystematic variability, inconsistency, substandard levels of agreement, and overall mediocre performance of the classification model have implications for the quality of legal determinations. Considering error in forensic decision making, reliability estimators may be viewed as indices of "consistency in judgments" (Suen & Ary, 1989, p. 155) and "reproducibility" of opinions (Cohen, 1960). Error reflects inconsistency in opinions and inaccurate classification of defendants (Christensen, Crowder, Ousley, & Houck, 2014).

Complexity of Task Demands

A focus of the current study was empirical assessment of forensic task complexity. Meyer, Mihura, and Smith's (2005) meta-analysis of interrater agreement across a wide range of medical and scientific tests (e.g., diagnostic tests in radiology, cardiology, etc.) examined the association between complexity and rater agreement. Reflecting the "underlying computational demands of the decision task" (Sung, Johnson, & Dror, 2009 p. 321), variables contributing to task complexity include time pressure, heterogeneity of information, computational intensity, and working memory requirements. A mismatch between computational demands and cognitive resources is likely as complexity increases. The Meyer et al. (2005) meta-analysis found that circumscribed judgments (e.g., physical measurements) are more reliable than "complex tasks requiring synthesis of multiple, higher order inferences" (p. 310). Judgment of "static or unchangeable objects" demonstrated higher levels of interrater agreement than "dynamic tasks," such as interviewing on separate occasions. As task complexity increases, "individuals may use heuristic-based strategies, with associated increases in effort, confusion, error rate, and consequent reduction in performance" (Neal & Grisso, 2014; Sung et al., 2009, p. 321). Competency to stand trial, sanity, and release judgments are highly complex and inferential, requiring data integration and interpretation in contrast to single forensic assessment instruments, such as the HCR-20 or PCL-R (Murrie et al., 2013). While CST evaluations appear to be comparatively less inferential (and hence more reliable), assessment of retrospective states of mind (NGRI), or prediction of future violence (CR), are more complex and inferential.

Poor consensus and inconsistency in examiner judgments may reflect different assumptions and decision thresholds. Are examiner decision thresholds conscious and explicit, or unconscious and implicit? If implicit, do examiners utilize cognitive heuristics, for example, representativeness, availability, or anchoring, in their decision making (Neal & Grisso, 2014)? Does engaging in a formal a priori decision analysis for weighing costs of errors and tradeoffs improve examiner decision making? Each type of forensic examination has specific decision tradeoffs. It may be the case that examiners have no prevailing assumptions at all. Their decision rules may be implicit and their decisions random, especially in Mossman's "hardest to decide" or "fuzzy" zone where information is complex, equivocal, or ambiguous (Mossman, 2013). Despite the mediocre levels of performance, Hawaii's policy of independently court-appointed forensic clinicians is a corrective to biasing effects of partisan allegiance (Murrie et al., 2008, 2009, 2013).

Decision Costs in Forensic Evaluations

Forensic opinions are essentially risk assessments (Skeem & Monahan, 2011). Each predictive or postdictive legal scenario carries unique risks and costs of errors. Threats to reliability are inevitable and error appears to be unavoidable (Mossman, 2013). Reliability threats include a multitude of variables: physical and temporal changes in observation environment and conditions, nonstandardized evaluation methods, skill, training, and diligence of examiners, defendant-related factors such as variable clinical condition and level of compliance, presentational artifacts such as impression management, examiner-defendant interaction effects, examination order effects, and examiner decision thresholds contribute to task complexity.

Engaging in a formal a priori decision analysis of costs of errors may provide examiners a rational model for weighing the consequential impact of misclassifications. Considering CST, is it more problematic to try an IST defendant or to find a competent defendant IST? Trying and convicting an IST defendant denies the right of the defendant to the Constitutional guarantee of a fair trial (*Pate v. Robinson*, 383 U.S. 375 (1966) and a violation of Constitutional due process (*Bishop v. United States*, 350 U.S. 961. P. 383 U. S. 378). A misclassified IST opinion may involve going to a felony trial with an IST defendant. Trying, convicting, or executing an incompetent defendant is a denial of due process (*Ford vs. Wainwright*, 477 U.S. 1986). Conversely, a misclassified CST defendant may spend weeks or months in detention, with a delay in legal proceedings. Regarding NGRI, the costs of errors may be higher than CST; is it more problematic to find a sane defendant insane, or an insane defendant sane? In finding an insane patient "sane," the defendant faces the prospect of legal jeopardy for conduct that might have been exculpated under the law due to a mental disability. Here the decision tradeoffs involve loss of liberty, and potential involvement with the state judicial and mental health systems for years or even decades. The best-case scenario may involve a lengthy stay at the state hospital until the misclassification is corrected. One notorious type of forensic misclassification—malingered insanity acquitees—are well known for their poor, disruptive adjustment in state hospital settings (Gacono, Meloy, Sheppard, Speth, & Roske, 1995).

Forensic mental health assessment work also poses personal and professional risks to examiners, judges, and institutions (Miller & Brodsky, 2011). The postacquittal release of individuals with severe mental illness who have been acquitted by reason of insanity for a violent crime is perhaps the most consequential. Reichlin and Bloom (1993) describe the convulsive and damaging impact of publicity and public opinion on an Oregon state hospital facility where a mentally ill murderer escaped into the community ("Killer Flees State Hospital Program"; Reichlin & Bloom, 1993). CR evaluations highlight complex technical aspects of violence risk assessment, prediction, and management. Despite significant improvements in risk assessment methodology (Skeem & Monahan, 2011), prediction of future risk of dangerousness remains exceedingly difficult (Cooke & Michie, 2010; Hart & Cooke, 2013). The greatest fear of examiners and judges is releasing a patient who subsequently commits a highly publicized violent crime. For CR, alternatively, the costs of detaining a low-risk defendant are also high, requiring housing, supervision, and engagement with the criminal justice system.

Agreement and reliability analyses indicated the performance of examiners in postacquittal CR evaluations was particularly poor. In the CR quality study (Nguyen et al., 2011), CR reports also demonstrated the poorest quality. Structured approaches for data integration, hypothesis testing, and optimization of classifications (Swets et al., 2000) have been proposed to improve examiner agreement and reliability. In our study, only a few examiners utilized forensic assessment instruments (FAIs), for example, suggesting that they relied on unstructured clinical judgment. This may account for the low levels of examiner agreement and high percentage of no opinion and split decisions in the CR group. Ideally, postacquittal CR decision making involves application of structured assessments integrating remediation of mental illness, likelihood of postdischarge remediation, risk of dangerousness and substance use, and risk management (McDermott et al., 2008; Neal & Grisso, 2014). Structured professional judgment where examiners are systematically guided through the process of risk assessment, formulation, and management (Sutherland et al., 2012, p. 120), is the desired professional standard of practice. Deficiencies in the assessment model, implicit or explicit examiner biases, extraneous factors, and costs of errors inevitably influence examiner performance (Miller & Brodsky, 2011).

Quality Improvement in Forensic Evaluations

It is legitimate to ask if forensic evaluations conducted in the field necessarily impose a ceiling on rater performance. Miller and colleagues (Miller et al., 2012) found that field reliability of common risk assessment measures used in sexually violent predator proceedings found lower reliability than that reported in test manuals. A recent Psychopathy Checklist–Revised (PCL-R) study demonstrated similar reliability findings when the instrument was used in applied, naturalistic settings (Sturup, Edens, Sorman, Karlberg, Fredriksson, & Kristiansson, 2014). These researchers found wide variability between examiners and "reliability statistics were generally below what are reported in the PCL-R professional manual" (p. 318). Miller and colleagues wrote that, with the absence of reliability, "courts cannot have confidence that an offender's reported risk assessment score is a true representation" and "unreliable estimates and judgments are invalid estimates and judgments" (Miller et al., 2012).

Training of examiners has been found to be the single most important factor in improving rater performance (e.g., Robinson & Acklin, 2010; Sutherland et al., 2012). Robinson and Acklin (2010) found a posttraining effect on report quality measures in examiners who had attended a two-day forensic report writing conference. Based on their study of interrater reliability in sexual violence risk examiners, Sutherland and colleagues (2012) noted that training improved agreement and concordance with expert opinion. They also noted that experience and examiner confidence was not correlated with indices of agreement. They further note that examiners with less training "were also more likely to overestimate risk" (p. 131). Similarly, a recent quasi-certification process implemented in the state of Hawaii led to significant increases in both reliability and quality in CST, NGRI, and CR evaluation reports submitted after examiner certification as compared to those submitted prior to certification (Gowensmith, Sledd, & Sessarego, 2014). Miller and colleagues (2012) recommend that examiners be familiar with the strengths and limitations of the measures utilized in communicating with the courts.

The use of validated forensic assessment instruments (FAIs) has been proposed as a means to improve the quality of data and as a sound evidence base for forensic examination findings. Checklists have been proposed as a means for organizing report information and counteracting decision-making biases, including prereflective or implicit biases (Miller & Brodsky, 2011; Witt, 2010). Although structured professional judgment has been limited to the application of specific risk assessment methods (Skeem & Monahan, 2011), as an approach to the organization, weighing, and integration of data, it may have much to offer other types of evaluations (Sutherland et al., 2012). Curiously, the use of FAIS and structured professional judgment may be related to examiner age cohort: Younger clinicians (< 47 years of age) were found significantly more likely to utilize risk assessment tools when compared to older clinicians (Viljoen, McLachlan, & Vincent, 2010).

Standard laboratory recommendations for improving interrater agreement and reliability (Acklin, 2012) include clarifying definitional criteria, standardization of assessment methods, peer review, performance feedback, and periodic retraining to control "rater drift" (Haynes, 1978), typical in academic research, do not typically obtain in the field. Based on our current findings, four additional factors may be necessary in improving accuracy and agreement in forensic evaluations: (a) examination of examiner's tacit assumptions; (b) decision analysis of the costs of errors since sensitivity and specificity are reciprocal and a function of examiner decision thresholds; (c) application of structured professional judgment models in data integration and decision-making processes; and (d) utilization of forensic assessment instruments. Finally, methods for de-biasing judgments have been developed in the intelligence community, which permits rigorous hypothesis testing and analysis and permits the decision maker to backtrack the decision path. "Analysis of Competing Hypotheses" (ACH; Heuer, 1999) is an elegant model for how to think about a complex problem when the available information is incomplete, ambiguous, or complex. ACH procedures identify and test alternative hypotheses that counter common heuristics, such as conformation bias or the availability and representativeness heuristics.

Limitations, Generalizability, and Future Research

The failure to code for individual examiners and judges constitutes a limitation of the study's design. The study design was uncrossed, or "ill-structured," a limitation introduced by the naturalistic nature of the study. Preference for nested or fully crossed designs is unequivocal and "allows for systematic bias between coders to be assessed and controlled for in an IRR [interrater reliability] estimate" (Hallgren, 2012, p. 3). Putka et al. (2008) and Narayanan et al. (2010) define the necessary conditions for reliability studies with unbalanced, uncrossed, or fully nested designs. In comparing parallel reliability estimates in crossed and ill-structured designs, Putka et al. (2008) concluded that "all reliability estimators examined appeared to be relatively free of bias where there was little overlap between the sets of raters who rated each ratee" (2008, p. 978). Despite the sample sizes obtained here ($N = 450$), confidence intervals for many variables were quite wide. Finally, given the low frequency of examiner use of forensic assessment instruments (FAIs), the current study did not examine whether FAIs improve consensus and decision making.

Concerning external validity, we have no reason to believe that our current findings are not generalizable to typical field reliability in other jurisdictions across the United States that are utilizing community-based examiners. The professional performance model examined here demonstrated only mediocre performance in single-examiner situations, a situation that obtains in a majority of U.S. jurisdictions (Mossman et al., 2007). Miller et al. (2012) note the particularly disappointing performance of single-rater reliability. Mediocre levels of agreement and consensus in CST and NGRI evaluations and poor levels of CR reliability indicate significant areas of concern. Judges improve the overall examiner decision model, appear to utilize intuitive decisions in maximizing consensus when panels disagree, and demonstrate independence in hard-to-decide cases. These findings illustrate legitimate concerns about the quality of forensic evaluations across the United States. Hawaii's three-panel system, however, appears to place judges and defendants into a favorable position for an accurate and just judicial determination, counteracting the potential effect of adversarial allegiance and creating a multirater situation (Miller et al., 2012).

It may be the case that the highly complex, inferential nature of forensic evaluations in field settings inevitably imposes limits on reliability. The inevitably of error is unavoidable (Mossman, 2013). The question is whether actual improvement in examiner performance is possible through education, procedural reforms, and rigorous standards of practice. Substantial improvement in the quality, agreement, and consensus of forensic mental health evaluations is likely to require systematic reform. For forensic psychologists, the Specialty Guidelines for Forensic Psychology (APA, 2103) tend to be aspirational and nonspecific with respect to practice standards and procedural guidelines. The guidelines fall short of the type of reforms called for in *Strengthening Forensic Science in the United States: A Path Forward* (2009). The field of forensic mental health assessment currently lacks a standard of practice, a critical issue in quality improvement and the regulation of poor practice (Heilbrun, DeMatteo, Marcyzk, & Goldstein, 2008; Otto & Heilbrun, 2002). *Strengthening Forensic Science in the United States* (2009) advocates strong reforms, noting, for example, that "the accuracy of forensic methods resulting in classification and individualization conclusions needs to be evaluated in well-designed and rigorously conducted studies" (p. 184). The use of accreditation, standards and guidelines for quality control, proficiency testing to assess method performance, and certification as "a process specifically designed to insure of the competency of the individual examiner" (p. 208) is recommended to strengthen the quality of forensic behavioral science in the courtroom.

REFERENCES

Acklin, M. W. (2012). *Quantifying consensus: Methodological issues in competency to stand trial evaluations*. Unpublished manuscript.

Altman, D., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*, 453–473. doi:10.1002/(ISSN)1097-0258

American Psychological Association. (2013). *Specialty guidelines for forensic psychology*. Retrieved from http://www.apa.org/practice/guidelines/forensic-psychology.aspx

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*, 357–370. doi:10.1037/1082-989X.2.4.357

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York, NY: Cambridge University Press.

Bishop v. United States, 350 U.S. 961. P. 383 U. S. 378.

Blais, J., & Forth, A. (2014). Prosecution-retained versus court-appointed experts: Comparing and contrasting risk assessment reports in preventative detention hearings. *Law and Human Behavior*, *38*(6), 531–543. doi:10.1037/lhb0000082

Chauhan, P., Warren, J., Kois, L., & Wellbeloved-Stone, J. (2015). The significance of combining evaluations of competency to stand trial and sanity at the time of the offense. *Psychology, Public Policy, & Law*, *21*(1), 50–59. doi:10.1037/law0000026

Christensen, A., Crowder, C., Ousley, S., & Houck, M. (2014). Error and its meaning in forensic science. *Journal of Forensic Sciences*, *59*(1), 123–126. doi:10.1111/jfo.2013.59.issue-1

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Cicchetti, D., & Sparrow, S. (1981). Developing criteria for inter-rater reliability for specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. doi:10.1177/001316446002000104

Conger, J. A. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322–328. doi:10.1037/0033-2909.88.2.322

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*, 259–274. doi:10.1007/s10979-009-9176-x

Cox, M., & Zapf, M. (2004). An investigation of discrepancies between mental health professionals and the courts in decisions about competency. *Law and Psychology Review*, *28*, 109–132.

De Swert, K. (2012). *Calculating inter-rater reliability in media content using Krippendorff's alpha*. Retrieved from http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. doi:10.1162/089120104773633402

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Ford vs. Wainwright, 477 U.S. 1986

Freelon, D. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*(1), 20–33.

Fuger, K., Acklin, M. W., Nguyen, A., Ignacio, L., & Gowensmith, W. (2014). Quality of criminal responsibility reports submitted to the Hawaii judiciary. *International Journal of Law and Psychiatry*, *37*(3), 272–280. doi:10.1016/j.ijlp.2013.11.020

Funder, D. (1990). Process versus content in the study of judgmental accuracy. *Psychological Inquiry*, *1*(3), 207–209. doi:10.1207/s15327965pli0103_6

Gacono, C. B., Meloy, J. R., Sheppard, K., Speth, E., & Roske, A. (1995). A clinical investigation of malingering and psychopathy in hospitalized insanity acquittees. *Bulletin of the American Academy of Psychiatry and Law*, *23*(3), 387–397.

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*(2), 130–139. doi: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*(2), 98–106. doi:10.1037/lhb0000001

Gowensmith, W. N., Sledd, M., & Sessarego, S. (2014, August). *The impact of stringent certification standards on forensic evaluator reliability*. Paper presented at the 122nd annual meeting of the American Psychological Association, Washington, DC.

Hallgren, K. A., (2012). Computing inter-tater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods in Psychology*, *8*(1), 23–24.

Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences & the Law*, *31*, 81–102. doi:10.1002/bsl.2049

Hayes, A. F. (2013). *KALPHA*. Retrieved from http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. doi:10.1080/19312450709336664

Haynes, S. N. (1978). *Principles of behavior assessment*. New York, NY: Gardner Press.

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York, NY: Routledge.

Heilbrun, K., DeMatteo, D., Marczyk, G., & Goldstein, A. (2008). Standards of practice and care in forensic mental health assessment: Legal, professional, and principles-based considerations. *Psychology, Public Policy, & Law*, *14*(1), 1–26. doi:10.1037/1076-8971.14.1.1

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Hunsley, J., & Mash, E. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Kraemer, H., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, *21*, 2109–2129. doi:10.1002/(ISSN)1097-0258

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *1*(1), 30–46.

McDermott, B. E., Scott, C. L., Buss, D., Andrade, F., Zozaya, M., & Quanbeck, C. (2008). The conditional release of insanity acquitees: Three decades of decision-making. *Journal of the American Academy of Psychiatry and Law*, *36*(3), 329–336.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

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. doi:10.1207/s15327752jpa8403_09

Miller, C., Kimonis, E., Otto, R., Kline, S., & Wasserman, A. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment*, *24*(4), 944–953. doi:10.1037/a0028411

Miller, S. L., & Brodsky, S. L. (2011). Risky business: Addressing the consequences of predicting violence. *Journal of the American Academy of Psychiatry and the Law*, *39*(3), 396–401.

Mossman, D., Noffsinger, S. G., Ash, P., Frierson, R. L., Gerbasi, J., Hackett, M., . . . Zonana, H. V. (2007). Practice guideline: Evaluation of competence to stand trial. *Journal of the American Academy of Psychiatry and Law*, *4*, S59–S67.

Mossman, D. M. (2013). When forensic examiners disagree: Bias, or just inaccuracy? *Psychology, Public Policy, and Law*, *19*(1), 40–55. doi:10.1037/a0029242

Murray, J., & Thomson, M. (2010a). Clinical judgment in violence risk assessment. *Europe's Journal of Psychology* 6(1), 128–149. www.ejop.org

Murray, J., & Thomson, M. (2010b). Applying decision making theory to clinical judgments in violence risk assessment. *Europe's Journal of Psychology, 6*(2), 150–171. Retrieved from www.ejop.org

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. doi: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. 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

Narayanan, A., Greco, M., & Campbell, J. (2010). Generalisability in unbalanced, uncrossed, and fully nested studies. *Medical Education*, *44*, 367–378. doi:10.1111/med.2010.44.issue-4

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

Neal, T., & Grisso, T. (2014). The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy, and Law*, *20*(2), 200–211.

Nguyen, A., Acklin, M. W., Fuger, K., Gowensmith, W. N., & Ignacio, L. (2011). Freedom in paradise: Quality of conditional release reports submitted to the Hawaii judiciary. *International Journal of Law and Psychiatry*, *34*, 341–348. doi:10.1016/j.ijlp.2011.08.006

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Otto, R., & Heilbrun, K. (2002). The practice of forensic psychology: A look toward the future in light of the past. *American Psychologist*, *57*(1), 5–18. doi:10.1037/0003-066X.57.1.5

Pate V. Robinson, 383 US 375, 378. (1966).

Pontius, R. G., & Millones, M. (2011). Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407–4429. doi:10.1080/01431161.2011.552923

Putka, D., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. doi:10.1037/0021-9010.93.5.959

Reichlin, S. M., & Bloom, J. D. (1993). Effects of publicity on a forensic hospital. *Bulletin of the American Academy of Psychiatry and Law*, *21*(4), 473–483.

Rietveld, T., & Van Hout, R. (1993). *Statistical techniques for the study of language and language behavior*. New York, NY: Mouton de Gruyter.

Robinson, R., & Acklin, M. W. (2010). Fitness in paradise: Quality of forensic reports submitted to the Hawaii judiciary. *International Journal of Law and Psychiatry*, *33*(3), 131–137. doi:10.1016/j.ijlp.2010.03.001

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257–268.

Skeem, J., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, *20*(1), 38–42. doi:10.1177/0963721410397271

Statistical Package for the Social Sciences. (2009). *SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0*. Chicago, IL: SPSS Inc.

Sturup, J., Edens, J., 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*(4), 315–324. doi:10.1037/lhb0000063

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum.

Sung, M., Johnson, J., & Dror, I. (2009). Complexity as a guide to understanding decision bias: A contribution to the favorite longshot bias debate. *Journal of Behavioral Decision Making*, *22*, 318–337. doi:10.1002/bdm.v22:3

Sutherland, A., Johnstone, L., Davidson, K., Hart, S., Cooke, D., Kropp, P. R., . . . Stocks, R. (2012). Sexual violence risk assessment: An investigation of the interrater reliability of professional judgments made using the risk for sexual violence protocol. *International Journal of Forensic Mental Health*, *11*, 119–133. doi:10.1080/14999013.2012.690020

Swets, J., Dawes, R., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1–26. doi:10.1111/1529-1006.001

Viljoen, J., McLachlan, K., & Vincent, G. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment*, *17*, 377–395. doi:10.1177/1073191109359587

Witt, P. (2010). Forensic report checklist. *Open Access Journal of Forensic Psychology*, *2*, 233–240.

Zapf, P., Hubbard, K., Cooper, V., Wheeles, M., & Ronan, K. (2004). Have the Courts abdicated their responsibility for determination of competency to stand trial to clinicians? *Journal of Forensic Psychology Practice*, *4*(1), 27–44.