International Journal of Law and Psychiatry xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Law and Psychiatry

Quality of criminal responsibility reports submitted to the Hawaii judiciary ☆

Kristen D. Fuger [a], Marvin W. Acklin [a,*], Annie H. Nguyen [a], Lawrie A. Ignacio [a], W. Neil Gowensmith [b]

[a] Argosy University, Hawaii Campus, Honolulu, HI, United States
[b] University of Denver, Denver, CO, United States

ARTICLE INFO

Available online xxxx

Keywords:
Quality of forensic reports
Forensic assessment
Forensic mental health evaluations

ABSTRACT

This paper is the third in a series of research reports on quality of forensic mental health evaluations submitted to the Hawaii judiciary. Previous studies examined quality of reports assessing competency to stand trial (CST) and post-acquittal conditional release, in felony defendants undergoing court-ordered examinations. Utilizing a 44-item quality coding instrument, this study examined quality of criminal responsibility reports in a sample of 150 forensic mental health evaluations conducted between 2006 and 2010 by court-appointed panels. Raters attained high levels of agreement in training and quality coding. Similar to the previous studies, overall quality of reports was mediocre, falling below the .80 quality criterion score for report elements, regardless of evaluator professional identification or employment status. Level of agreement between evaluators and judicial sanity determinations was "fair" using Cicchetti's (1994) standards for interpretation of intra-class correlations. Level of agreement was lower than previously published findings for CST reports and better than conditional release reports. Reasons for mediocre report quality and "fair" inter-rater agreement are discussed, including the fact that criminal responsibility evaluations are complex, retrospective in nature, and involve significant degrees of inference. In contrast to CST evaluations, assessment of criminal responsibility involves a mental state at the time of the offense evaluation. Threats to reliability in forensic reports are discussed. Suggestions for improvement of report quality are proffered, including standardization of procedures and report format and use of forensic assessment instruments.

1. Introduction

The insanity defense has been the focus of intense public attention and misperception. Inaccurate knowledge of the insanity defense prevails (Golding, Skeem, Roesch, & Zapf, 1999; Hans & Slater, 1983). Common public conceptions view the insanity defense as a loophole allowing the guilty to avoid responsibility for their actions. The public overestimates legal application of the insanity defense; some studies have shown that laypersons believe the insanity defense to be raised in about one-third of all felony cases (Pasewark, McGinley, & Blau, 1989). In reality, criminal responsibility pleas are a relatively rare occurrence within the criminal justice system, with roughly one percent of criminal cases raising an insanity defense (Melton, Petrila, Poythress, & Slobogin, 2007). An insanity acquittal occurs in about one in four of these cases (Blau, McGinley, & Pasewark, 1993; Murrie & Warren, 2005; Quinsey, 2009; Zapf, Golding, & Roesch, 2006). Warren, Fitch, Dietz, and Rosenfeld (1991) reviewed 894 pre-trial reports for criminal responsibility in Virginia and found that 8% of the defendants were adjudicated not criminally responsible. In Hawaii, the rate of insanity pleas is slightly higher than the national average, with estimates of 1–3% of criminal cases raising an insanity plea, with the courts determining acquittal in about 25% of felony cases raising this defense (Gowensmith, 2008). There are approximately 300 felony criminal responsibility evaluations conducted in Hawaii annually, most of which occur in the First Judicial Circuit (Island of Oahu, Gowensmith, 2008).

1.1. Mental state at the time of the offense evaluations

In contrast to a competency to stand trial (CST) evaluation, criminal responsibility assessments require the clinician to conduct a retrospective evaluation of the defendant's mental state at the time of the offense (Acklin, 2007a; Melton et al., 2007; Roesch, Viljoen & Hui, 2004; Simon & Shuman, 2002). "The overriding goal of the insanity evaluation is a comprehensive reconstruction of the defendant's functioning at the time of the offense" (Rogers, 2008, p. 113). Given the retrospective and inferential nature of the examination, assessment of mental state at the time of offense is one of the most challenging forensic assessments (Acklin, 2007a; Melton et al., 2007).

Retrospectively, the forensic examiner must dissect the offense and examine and integrate the clinical and collateral data (Acklin, 2007a; Melton et al., 2007; Murrie & Warren, 2005; Simon & Shuman, 2002; Warren et al., 2004). This requires reconstruction and evaluation of events leading up to, during, and following the offense, thereby creating

☆ This report is based in part on the first author's dissertation submitted for partial requirements of the doctoral degree in psychology at Argosy University, Hawaii Campus.

* Corresponding author at: 850 W. Hind Drive Suite 203, Honolulu, HI 96821, United States.
E-mail address: acklin@hawaii.edu (M.W. Acklin).

0160-2527/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijlp.2013.11.020

"a reconstruction of the defendant's thought processes and behavior" (Melton et al., 2007, p. 201; Rogers, 2008; Weiner, 2006).

In assessing criminal responsibility, it is insufficient to opine that a defendant possesses a mental disease or disorder with impaired understanding, appreciation and/or control of their behavior. The mental disorder must also cause functional and legally definable impairments (Grisso, 2003). While most successful criminal responsibility cases involve psychotic disorders, psychosis is not synonymous with insanity (Acklin, 2007a; Melton et al., 2007). Psychiatric diagnoses do not drive the findings of insanity; rather, diagnoses provide the forensic evaluator with a framework for assessing pertinent clinical symptomatology and linking these factors with conduct at the time of the offense (Grisso, 2003; Rogers, 2008). Linking psychiatric factors to conduct at the time of the offense relevant to legal capacities carries weight in addressing the pending legal issue (Grisso, 2003; Melton et al., 2007).

1.2. Quality issues in criminal responsibility reports

Quality of forensic work products is a central component in the evolution of standards in the field of forensic mental health assessment (Heilbrun, DeMatteo, Marczyk, & Goldstein, 2008). Nicholson and Norwood (2000) observed, "The practice of forensic assessment falls far short of its promise" (p. 40). Warren et al. (2004) suggest that ordinary forensic practice typically falls short of professional aspirations. Wettstein (2005) reviewed the empirical research regarding perceived quality of forensic evaluations, highlighting problems in several areas: definitional criteria, process of quality assessment, quality indicators and measures, in the ongoing quality improvement enterprise. The developing literature emphasizes the need for improvement in the practice of forensic assessment, noting the wide variability and inconsistency in forensic evaluations across different jurisdictions (Gowensmith, Murrie, & Boccaccini, 2012; Heilbrun et al., 2008; Nguyen, Acklin, Fuger, Gowensmith, & Ignacio, 2011; Nicholson & Norwood, 2000; Otto & Heilbrun, 2002; Robinson & Acklin, 2010; Wettstein, 2005). Identified areas of quality variability include inconsistent use of psychological testing, failure to assess factors related to the legal issue at hand, lack of use of third party data, and lack of a linkage between clinical findings, capacities, and legal questions (Nicholson & Norwood, 2000).

Grisso (1986) outlined a variety of common criticisms that are often leveled against forensic evaluators, including lack of relevance of the opinions relative to the legal question, lack of confidence in expressed opinions, and opinions based on inadequate sources of information. Frequent faults include opinions that lack rationales, where the forensic purpose or referral question is unclear, organizational problems in report format, inadequate database (i.e., lacking data sources or use of irrelevant data), and overuse of clinical jargon (Grisso, 2010). Grisso characterizes competent forensic practice as: *accurate and accountable* (i.e., clear and logical explanations for what one does, what the data holds, and the conclusions), *specific* (i.e., efficiently answering the question that was asked by the referral source with necessary and logical clinical and forensic information), and *conceptually integrated* (i.e., the conclusions of the forensic evaluation should be consistent and based on logical and sound techniques, theories and information).

Heilbrun and Collins (1995) noted that forensic evaluators infrequently addressed important psycholegal components in criminal responsibility evaluations. Only 41% of the sample reported a conclusion regarding whether the defendant knew what they were doing, 27% addressed the awareness of consequences, and 29% addressed the awareness of wrongfulness in the defendant. In a study of 46 criminal responsibility reports, Borum (1994) found that 20% failed to mention criminal responsibility. In Hawaii, Acklin et al. (2005) found that 49% of criminal responsibility evaluations and 63% of CST evaluations failed to include a rationale for the psycholegal opinion. Robinson and Acklin (2010) found that 74% of CST evaluations included a rationale for their findings. In a study of conditional release report quality in Hawaii, Nguyen et al. (2011) found that only 35% of conditional release reports gave a complete rationale for the opinion of dangerousness. Only 60% gave a complete rationale for the conditional release recommendation. A necessary and sufficient forensic opinion links clinical and legal factors to the legal standard and standard of proof ("To a reasonable degree of psychological certainty, Mr. Doe's cognitive and volitional capacities were substantially impaired by schizophrenia"; Babitsky & Mangraviti, 2002).

Rogers and Shuman (2000) suggest that the state of insanity evaluations has "largely been an idiosyncratic process, reflecting the propensities and proclivities of the clinician" (p. 520). Grisso (2010) described failure to document the use of third party information as a common error occurring in forensic evaluations. In a study conducted by Otto et al. (1996), as cited in Nicholson and Norwood (2000), which reviewed 71 criminal responsibility reports, only 10% of the reports used any other data than the defendants' narratives from which they developed their opinions. Warren et al. (2004) found that clinicians tended to offer their opinions based on incomplete data. This study found that over half the examiners offered an opinion without review of the defendant's statement, criminal history or witness statements.

The professional literature defines parameters that represent best practices in forensic report quality. Heilbrun (2001) outlined 29 broad principles, grouped into four broad categories: preparation, data collection, data interpretation, and communication. These include ethical elements, absence of jargon and clarity of exposition (Allnutt & Chaplow, 2000; Giorgi-Guarnieri et al., 2002; Harvey, 1997; Melton et al., 2007), data elements in the forensic database; methodological elements, including procedures utilized in the assessment (Acklin, 2007a,b; Archer, Buffington-Vollum, Stredny, & Handel, 2006; Borum & Grisso, 1995; Lally, 2003; Melton et al., 2007; Nguyen et al., 2011; Robinson & Acklin, 2010), and opinion and rationale demonstrating linkages between clinical and legal impairments in relation legal standards (Gagliardi & Miller, 2008; Grisso, 2003; Melton et al., 2007; Zapf et al., 2006). The evaluator's justifications for an opinion should be clearly communicated (Gagliardi & Miller, 2008; Golding et al., 1999; Grisso, 2003; Hecker & Scoular, 2004; Melton et al., 2007; Wettstein, 2004). "Reports that only provide cursory psycholegal opinions or those that leap from a diagnosis to a psycholegal opinion no longer meet the standard in the field" (Conroy, 2006, p. 240). The forensic clinician is held by ethical standards to substantiate conclusions and provide the basis for the conclusions presented in a forensic report (Skeem & Golding, 1998; Specialty Guidelines, 1991, 2011). Determinations, opinions and diagnoses must be independent of other examiners and based on substantiated data and reasoning (Connell, 2008, HRS-704-404). Skeem and Golding (1998) state: "The most critical function involves advising the court about the defendant's specific abilities and deficits and explaining one's reasoned inference about the bases for these deficits" (p. 358).

1.3. Levels of agreement between forensic evaluators and the court

There is a high rate of agreement between a forensic opinions and ultimate judicial determination. Studies of CST evaluations report greater than 90% agreement rates (Greenberg & Wursten, 1988; Hecker & Steinberg, 2002; Warren et al., 2006; Zapf, Hubbard, Cooper, Wheeles, & Ronan, 2004). Research on forensic reports in Hawaii found agreement rates between forensic evaluators and judicial determination at approximately 90% (Acklin et al., 2005). Robinson and Acklin (2010) found that in 66% of these cases, evaluators and judges agreed on defendants' CST. Gowensmith et al. (2012) found "good" agreement rates among evaluators, with 70.9% agreement between all three evaluators in initial CST evaluations in Hawaii; the court and a consensus of evaluators (2 of 3) agreed on initial CST determinations 92.5% of the time. In contrast to CST reports, Nguyen et al. (2011) found that in conditional release evaluations (evaluations for post-acquittal release), all three of the evaluators and the judge reached unanimous agreement in only 39% of cases.

Limited data is available about the inter-rater agreement rates of insanity evaluations. Most recently, Gowensmith, Murrie, and Boccaccini

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

(2013) found that independent triads of sanity evaluators in Hawaii agreed unanimously in only 55% of cases, although chance agreement in that sample was estimated at 31%. Viljoen, Roesch, Ogloff, and Zapf (2003) cite two studies that have looked at inter-rater agreement by comparing the court decisions with evaluator decisions. Daniel and Harris (1981) identified an agreement rate of 88% between clinicians and courts on insanity. Fukunaga, Pasewark, Hawkins, and Gudeman (1981) found agreement rates of 93%; however, this rate was determined at a time in which evaluators could consult with each other prior to submitting their reports to the court. Research has also examined rates of clinical variation in insanity findings to determine whether there is a propensity towards one opinion or another. Murrie and Warren (2005) found that most clinicians found that defendants met the criteria for insanity in 5–25% of cases evaluated. This is consistent with other research that has found rates of 11–12% (Murrie & Warren, 2005). Warren et al. (2004) found a range of 13–20% insanity findings and reported that psychiatrists (20%) in their sample gave an opinion supportive of insanity significantly more than psychologists (13%).

1.4. Insanity defense and criminal responsibility evaluations in Hawaii

Hawaii's criminal responsibility statute follows the American Law Institute (ALI) standard, citing both cognitive and volitional prongs of criminal responsibility (Hawaii Revised Statute 704-400). Similar to many other ALI jurisdictions, Hawaii law states that an individual is not criminally responsible for their conduct if,

> at the time of the conduct as a result of physical or mental disease, disorder or defect the person lacks substantial capacity either to appreciate the wrongfulness of the person's conduct or to conform the person's conduct to the requirements of law.

Following the ALI standard, Hawaii statutes maintain that "the terms *physical or mental disease, disorder, or defect* do not include an abnormality manifested only by repeated penal or otherwise anti-social conduct" (HRS-704-400).

HRS section 704-404 outlines the parameters for court-ordered mental health examinations. The statute outlines the procedures for appointing examiners and specifies what must be included in communications with the court, including requirements for independence of the evaluations and opinions on the ultimate legal question. The court orders one examination (a "one-panel exam") for cases involving misdemeanors, and three concurrent examinations (a "three-panel exam") for felony cases. One of the examiners in all of these cases (the sole examiner for one-panel exams, and one of three examiners for three-panel exams) must be a designate of the State of Hawaii Department of Health (DOH); the department carrying this responsibility is known as "Courts and Corrections." For three-panel examinations, the other two examiners are drawn from a list of community-based mental health professionals certified by the DOH. The statute requires that at least one of these two independent examiners be a licensed psychiatrist, and that the third examiner be either a licensed psychiatrist or licensed psychologist. In practice, the third examiner is most often a psychologist. Forensic evaluators submit their independent evaluation findings to the court in the form of a written report which provides the court with an expert opinion on the legal question posed by the court.

The statute defines specific parameters for the content of the forensic report on criminal responsibility. The following items are mandated: a description of the examination, diagnosis of the defendant, an opinion as to whether the defendant had the capacity to appreciate the wrongfulness of conduct or conform behavior to the requirements of law, an opinion as to the defendant's capacity to embody a particular state of mind required to establish elements of the legal charge (when specifically ordered by court), and a statement that the opinion was arrived at independently from other examiners. Hawaii courts routinely mandate CST and criminal responsibility opinions in the same court order, especially in the initial evaluation of the defendant (Giorgi-Guarnieri et al., 2002; Gowensmith, personal communication, April 24, 2011).

This study is the third in a series of forensic report quality studies. Previous studies addressed the quality of CST evaluation reports (Robinson & Acklin, 2010) and conditional release evaluation reports (Nguyen et al., 2011). This study addresses quality of criminal responsibility evaluation reports as a contribution to the report quality literature.

2. Method

2.1. Sampling procedures

Data for the current study were collected from archival records at the First Circuit Court of Hawaii, the jurisdiction that covers the island of Oahu (the city and county of Honolulu is situated on Oahu, a 50-mile wide island, with a population of approximately 1,000,000 people). The records are housed in the courthouse and available to the public upon request. The data in this study were public, archived records for which there was no researcher–participant interaction. The evaluations reviewed for this study were conducted on defendants charged with felony offenses who were either incarcerated or hospitalized while undergoing a court-ordered criminal responsibility evaluation.

The first author was provided a list of court-ordered criminal responsibility evaluations for the years 2006–2010. Fifteen reports were selected for use in the inter-rater training trials. Fifty cases, which yielded 150 reports, were selected for review. Cases were selected according the following criteria: 1) independent court-ordered evaluations for criminal responsibility with a written forensic report from all three panel members, with an opinion on criminal responsibility; 2) evaluations conducted within between January 2006 and January 2010; 3) evaluations with a judicial determination of criminal responsibility; and 4) and conducted within the First Circuit Court of Hawaii.

2.2. Measures

Mental examination reports were coded using an objective survey instrument originally designed for evaluating juvenile forensic assessments by Sanschagrin (2006), and was modified for use with criminal responsibility evaluations. The instrument was previously utilized in forensic report quality research (Nguyen et al., 2011; Robinson & Acklin, 2010). Sanschagrin (personal communication, 10/12/2009) provided permission to modify the instrument for use in this study.

The survey instrument contained 43 items designed to assess report quality elements. Survey items were equally weighted and summed to yield a Quality Coefficient (QC) score. The survey contained 33 items that contributed to the QC score. Individual items were coded 2 if coded information was complete. Items were coded 1 if coded information was partial, vague, or incomplete. Items were coded 0 when the evaluator failed to include any information on the item. Quality Coefficient scores range from 0 to 66. In the case of an NGRI opinion, the court order mandated additional information including a dangerousness assessment and recommendations for the release or commitment of the defendant. For those reports that recommended NGRI, there are a total of 37 items with a range of 0–74 to account for the additional information. The remaining 6 coding items were used for classification purposes only (type of professional degree, board certification, type of third party data, type of FAI used, recommendation, and judicial determination) and were not included in the calculation of QC scores. The QC was calculated by dividing the total score of each report into the maximum possible score, which was converted to a percentage score. Report quality criterion was set at 80%, the standard used in the two previous report quality studies (Nguyen et al., 2011; Robinson & Acklin, 2010).

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

2.3. Procedure

Collection and data coding occurred in two phases. The first author trained a doctoral-level graduate research assistant on the use of the quality rating measure, included the element and individual items. Once the training was completed, the researcher and research assistant independently rated fifteen reports selected for inter-rater trials. Results from the ratings were analyzed using intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979) to assess level of inter-rater reliability. Interpretation of inter-rater analyses was applied using Cicchetti's (1994) interpretive criteria. The inter-rater reliability analyses were conducted on individual items and on the QC score in order to identify specific areas of disagreement and further need of training. The researcher and research assistant discussed discrepant items for batches of five reports to improve rigor and accuracy of coding, and minimize rater drift (Haynes, 1978). In inter-rater training trials 1–5, 24 items had perfect (1.0) agreement with an overall ICC of .88 ($p < .00$, 95% CI: .85–.91) reflecting "excellent" agreement (Cicchetti, 1994). In inter-rater training reports 6–10, 32 items achieved perfect agreement (1.0), and an overall ICC of .91($p < .00$, 95% CI: .89–.93), "excellent" agreement (Cicchetti, 1994). In trials 11–15, 33 items achieved perfect agreement (1.0) and an overall ICC of .85 ($p < .00$, 95% CI: .82–.88), also "excellent" agreement (Cicchetti, 1994). Agreement over all fifteen inter-rater reports was "excellent" (Cicchetti, 1994) with an overall ICC of .88 ($p < .00$, 95% CI: .87–.90.) As a result, training was deemed sufficient to move the formal coding phase of the study.

Fifty cases totaling one hundred and fifty reports were coded following the inter-rater training trials. The researcher and research assistant coded every tenth report in order to increase the reliability of the scoring, reduce potential rater bias, and to reduce rater drift. These tenth reports were analyzed using the same ICC standard utilized during the inter-rater trials. The range of achieved ICC in these reports was .82–.97. Mean ICC was .90 (SD = .04). This indicated "excellent" agreement (Cicchetti, 1994) and no evidence of rater drift.

2.4. Research hypotheses

Based on two previous studies of forensic report quality submitted to the Hawaii judiciary, it was hypothesized that: 1) Reports would not meet national standards of quality as defined in the forensic literature, mean QC scores would score below 80% of the total possible score. 2) Quality of reports submitted by community-based psychologists and community-based psychiatrists would not significantly differ. 3) Quality of reports submitted by Department of Health psychologists would be higher than reports written by community-based psychologists. 4) Third party information would be under-utilized (less than 80% of reports. 5) Psychologists would tend to utilize testing more than psychiatrists. It was predicted that psychologists would utilize testing (cognitive, personality, or forensic assessment instruments) more often than psychiatrists. 6) Rates of agreement between the evaluators on criminal responsibility would reach "good" levels of agreement (ICC ≥ .60, Cicchetti, 1994). 7) Rates of agreement between evaluator opinions and judicial determination would reach "good" levels of agreement (ICC ≥ .60, Cicchetti, 1994). Agreement between at least two evaluators and judicial determination would reach "good" levels of agreement (ICC ≥ .60, Cicchetti, 1994).

2.5. Data analysis

The two previous studies in Hawaii have utilized Cohen's kappa as the method for quantifying inter-rater agreement (Robinson & Acklin, 2010). The current study utilized intraclass correlation coefficient (ICC) for all inter-rater agreement analyses. ICC is a measure of the proportion of variance of an observation due to between-subject variability, and is a common measure of agreement in the behavioral sciences (McGraw & Wong, 1996; Shrout & Fleiss, 1979). The use of ICC in place of kappa is an improvement over the previous studies; it is more appropriate than kappa in situations involving more than two raters. Furthermore, an ICC not only accommodates multiple raters but also permits analysis of data at the item level and summary level (additive sum of item ratings; Meyer, G., personal communication, 05/25/2010).

The current study utilized Cicchetti's (1994) interpretation of inter-rater agreement values. Cicchetti's (1994) ICC interpretation guidelines are as follows: <.40 = "poor", .40 to .59 = "fair", .60 to .74 = "good", and .75 to 1.0 = "excellent". Values >.80 are interpreted as "nearly perfect", with 1.0 being "perfect." Each report was coded and evaluated using the modified coding instrument originally developed by Sanschagrin (2006). An item was coded 0, 1, or 2, with an assigned maximum score of 2 and a minimum score of 0. Higher scores represented higher quality and comprehensiveness, while lower scores indicated lower quality and incompleteness. After the report was evaluated, a total Quality Coefficient (QC) was calculated. Quality Coefficients were calculated by dividing the total score for each evaluation by the maximum possible to arrive at a percentage score. This study had a sample size of 50 cases (150 reports), a sample size sufficient to optimize statistical power ($1\text{-}\beta = .79$, $p = .05$ (Cohen, 1992).

3. Results

Hypothesis 1. Overall, only five (3.3%) reports met or exceeded the 80% quality criterion. Table 1 presents descriptive statistics for quality elements. The mean QC score for the sample (N = 150) was 60.67% (SD = 9.11) and ranged from 41% to 85%. The median QC score was 61.47%. Only two elements (Identification elements and Practical elements) achieved the 80% quality criterion. These findings confirm Hypothesis 1; quality of forensic reports fell below the selected quality criterion.

Examination of the 37 item-level QC scores revealed that 13 (35.13% of the items) met the 80% quality criteria. These included: defendant's name, case caption, evaluation date, report date, professional degree identified, independent statement, defendant's version of the offense, diagnosis, opinion provided, rationale for psycholegal opinion, plain language, sections, and the use of unbiased language.

Hypothesis 2. Community-based psychologists (n = 50) and community-based psychiatrists (n = 50) were compared for overall report quality. Reports completed by community-based psychologists had a mean QC score of 61.85 (SD = 7.04), with QC scores ranging from 44 to 76. Reports completed by community-based psychiatrists had a mean QC score of 58.12 (SD = 7.19), with QC scores ranging from 41 to 76. Community-based psychologists submitted higher quality reports than community-based psychiatrists ($t(98) = 2.61$, $p < .05$, two tailed). The effect size was moderate (*eta* squared 0.06, Cohen, 1988). These findings disconfirmed Hypothesis 2; community-based psychologist and psychiatrist's reports differed in quality.

Analysis at the element level quality scores revealed that community-based psychologists (M = 79.00, SD = 10.27) achieved higher QC scores on the Legal/Ethical element cluster compared to psychiatrists (65.67

Table 1
Quality scores for report and elements for entire sample (N = 150).

Quality element	Mean	SD
Total	60.67	9.11
Identification	84.39	12.59
Legal/Ethical	73.67	15.54
Historical	39.67	13.05
Assessment/Diagnostic	37.28	14.40
Psycholegal	76.67	24.13
Practical	96.89	8.71

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

mean QC, SD 14.44). Community-based psychologists included the Legal/Ethical cluster more frequently than psychiatrists ($t(98) = 5.31$, $p < .01$, two-tailed). The effect size was large (*eta* squared 0.22, Cohen, 1988). No other element level quality differences were found. Community-based psychologists more frequently included the Legal/Ethical items of identification and attribution of data to the source(s) and reason for referral than psychiatrists. Item level analysis revealed that psychiatrists utilized a forensic assessment instrument more frequently than the community-based psychologists. Community-based psychologists included the defendant's date of birth and a rationale statement for the psycholegal opinion more frequently than psychiatrists.

Hypothesis 3. Department of Health psychologists (DOH; $n = 50$) and community-based psychologists ($n = 50$) were compared on overall QC score. Reports completed by DOH psychologists had a mean QC score of 62.04 ($SD = 11.89$), with QC scores ranging from 41 to 85. Reports completed by community-based psychologists had a mean QC score of 61.85 ($SD = 7.04$), with QC scores ranging from 44 to 76. There was no significant difference in overall report quality between DOH and community-based psychologists ($t(79.65) = .097$, $p = .923$, two-tailed). These findings disconfirmed Hypothesis 3; report quality for DOH and community-based psychologists did not differ.

Several differences emerged for element level quality scores. DOH psychologists ($M = 86.1$, $SD = 15.9$) achieved higher QC scores for the Psycholegal Rationale element compared to community-based psychologists ($M = 76.6$, $SD = 25.9$ with $t(75.1) = 2.204$, $p < .05$, two tailed). The effect size was moderate (*eta* squared 0.04, Cohen, 1988). No other differences were found for element level quality scores.

Item level analysis of the Psycholegal Rationale element cluster revealed that this difference was isolated to the reports which recommended NGRI; specifically involving statements about dangerousness to self and others. DOH psychologists included statements about dangerousness items more frequently than community-based psychologists. No other differences were found for item level quality scores.

Hypothesis 4. Concerning sources of data utilized, a majority of the reports used between two to four sources of third party information ($N = 122$, 81.3%). Twenty reports (13.3%) utilized four or more sources. A small minority of reports (8 reports, 5.3%) used a clinical interview as a sole data source. Evaluators frequently referenced multiple sources of data in their reports (94.7%). These findings disconfirmed Hypothesis 4; evaluators utilized multiple sources of data in their reports.

Reports typically included criminal or probation records provided through the probation department (120 reports 80.0%). Most of the reports indicated the use of criminal and probation records, including police reports (123 reports, 82.0%). Collateral interviews were referenced in 98 reports (56.3%). Collateral interviews typically included family members, treating mental health professionals, and/or case managers. Medical records were referenced in 60 reports (40.0%); these mainly consisted of correctional medical records, which typically included psychiatric information. Evaluators often cited mental examination reports previously conducted by the evaluator (44 reports, 29.3%). Mental health records, such as psychiatric hospital records, were referenced in 31 reports (20.7%).

Community-based psychologists and Department of Health psychologists utilized two or more sources in 100% of reports. DOH evaluators utilized four or more third party data sources in 6 reports (12%); community-based psychologists used four or more sources in 5 reports (10%). Psychiatrists relied solely on a clinical interview in 8 reports (16.0%); four or more third party data sources were referenced in 9 psychiatrist reports (18%).

Hypothesis 5. Overall, psychological or forensic assessment instruments were utilized in less than a quarter of forensic reports (35 reports, 23.3%). In the 35 reports that utilized testing, cognitive assessment measures were utilized in 18 reports (12%). A personality assessment measure was used in only 1 report (less than 1%). In the 35 reports that utilized testing, forensic assessment instruments were utilized in 21 reports (14%). Five reports (3%) used a combination of cognitive testing and forensic assessment instruments.

Psychologists ($n = 100$) and psychiatrists ($n = 50$) were compared on the utilization of psychological assessment measures (including cognitive, personality, and forensic assessment instruments). Psychologists ($n = 100$) used clinical and/or forensic assessment measures in 21 reports (21%), including cognitive assessments; personality assessment measure (1 report); and 7 reports included a forensic assessment instrument. Psychiatrists ($n = 50$) used clinical and/or forensic assessment instruments in 14 reports (28.0%). One report written by a psychiatrist included a cognitive assessment measure; no reports included a personality assessment measure; and 14 reports included a forensic assessment measure. Psychologists and psychiatrists did not differ on overall assessment utilization ($t(148) = .95$, $p = .34$, two tailed). Psychologists used cognitive assessment more frequently than psychiatrists; however, psychiatrists used forensic assessment more frequently than psychologists. Hypothesis 5 was not supported. Neither psychologists nor psychiatrists routinely utilized assessment measures in their forensic reports, including forensic assessment instruments.

Psychologist examiners were disaggregated in order to compare community-based and Department of Health psychologists with psychiatrists for utilization of psychological assessments. No significant differences were found in any comparison between DOH and private practitioners, but DOH psychologists did use cognitive testing more often than psychiatrists.

Overall clinical and forensic assessment usage occurred in less than 25% of reports (23.3%) a disappointingly low rate of utilization. DOH psychologists tested most frequently (16 reports) followed by psychiatrists (14 reports). Forensic assessment instruments were utilized in a minority of evaluations (21 reports, 14.0%). Cognitive assessment instruments were used in 18 reports (12%); personality assessment (e.g., use of MMPI-2) was virtually nonexistent in the sample.

Hypothesis 6. Level of agreement between evaluators was calculated using interclass correlation coefficients (ICC; Shrout & Fleiss, 1979). Inter-evaluator level of agreement was "fair" (Cicchetti, 1994), $ICC = .51$, $p < .01$. Agreement rates between all three mental examiners were "fair." This level of agreement fell short of the .60 criterion set for Hypothesis 6, disconfirming Hypothesis 6.

In 23 cases (69 reports, 46%), all three examiner evaluators reached consensus: In 26 cases (78 reports, 52%), at least two evaluators reached consensus: Thus, in 98% ($n = 49$) of the cases, at least two of the three evaluators reached consensus on criminal responsibility. Only one case (3 reports, 2%) contained perfect disagreement between evaluators; one evaluator opined NGRI, one opined not NGRI, and the last evaluator refrained from an opinion.

Psychiatrists and community-based psychologists agreed 68.0% ($n = 34$) of the time on criminal responsibility recommendations. This represents a "fair" level of agreement ($ICC = .57$, $p < .01$). Community-based psychologists and Department of Health psychologists agreed in 64.0% ($n = 32$) of cases, also a "fair" level ($ICC = .54$, $p < .01$). Psychiatrists and DOH psychologists reached consensus with "fair" levels of agreement, in 58.0% ($n = 29$) of cases ($ICC = .42$, $p < .01$). These findings disconfirmed Hypothesis 6; levels of agreements failed to meet good levels of inter-evaluator agreement.

Hypothesis 7. Judicial determination of criminal responsibility was categorized as 1) not NGRI, 2) NGRI, or 3) not resolved. For the purposes of this study, if the individual was found guilty and sentenced, this accounted for a judicial determination of "not NGRI." Thirty cases (90 reports, 60%) were deemed not NGRI. If the defendant was acquitted

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

and committed or acquitted and released, this accounted for a judicial finding of NGRI. In nineteen cases (57 reports, 38%), judges determined defendants to be NGRI. One case was coded as not resolved (3 reports, 2%): the judicial determination was civil commitment and dismissal of the charges. This case was excluded from the analysis and left a sample size of 49 cases (147 reports).

Level of agreement between judicial determination and evaluators was calculated using interclass correlation coefficients (ICC; Shrout & Fleiss, 1979). Agreement for criminal responsibility between judicial determination and forensic examiners was calculated in two methods. The first method utilized judicial determination as a rater and compared agreement rates between four raters: the judge (judicial determination) and the three panel forensic examiners. The court and all three evaluators reached consensus in 40.8% of cases (N = 20, 60 reports). Level of agreement between judges (judicial determination) and all evaluators was "fair," ICC = .50, $p < .01$, $F = 5.11$, *df* (48–144), 95% CI = .36–.64.

The second method of calculating agreement used judicial determination as one rater and the majority opinion of the evaluators (defined by at least two of the three evaluators recommending the same opinion) as the second rater. In 77.5% (n = 38, 76 reports) of cases, the court and at least two of three evaluators agreed as to criminal responsibility. Agreement level was "fair," ICC = .49, $p < .01$, $F = 2.94$, *df* (48), 95% CI = .25–.68. Regardless of method of analysis, level of agreement failed to reach levels of agreement that are recognized as good.

The court and psychiatrists reached agreement in 63.2% of cases (n = 31, 62 reports) for "fair" levels of agreement, ICC = .47, $p < .01$. The court and community-based psychologists reached agreement agreed in 69.3% of cases (n = 34) for "fair" levels of agreement, ICC = .42, $p < .01$. In 77.5% of cases (n = 38), the court and DOH psychologists agreed on criminal responsibility recommendations. The court and Department of Health psychologists achieved "good" levels of agreement, ICC = .61, $p < .01$. The difference between levels of agreement between the court and DOH evaluators (.61) and the court and community-based evaluators (.42) was not statistically significant ($z = 1.27$, $p = .20$, two tailed). These findings failed to support Hypothesis 7; levels of agreement between evaluators and the court failed to reach "good" levels of agreement.

4. Discussion

The report quality of criminal responsibility reports in this sample can best be described as mediocre. These findings are consistent with our previous studies for CST and conditional release reports (Nguyen et al., 2011; Robinson & Acklin, 2010). Only five of 150 reports met the 80% criterion; 95% of reports submitted to the court failed to meet the 80% quality standard. In contrast to our previous studies, criminal responsibility report quality was higher than quality of conditional release reports (Nguyen et al., 2011), but lower than CST reports (Robinson & Acklin, 2010). Performance was generally equivalent to the CST study (Robinson & Acklin, 2010) and better than the conditional release study. Overall agreement rates were comparable to previously published criminal responsibility (Gowensmith et al., 2013) and CST studies (Gowensmith et al., 2012; Robinson & Acklin, 2010) and a substantial improvement over conditional release reports (Nguyen et al., 2011).

4.1. Reasons for mediocre criminal responsibility report quality

There may be legitimate reasons for differences in CST and criminal responsibility report quality. Competency to stand trial evaluations are conducted more frequently than criminal responsibility evaluations (Melton et al., 2007; Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008) and evaluators may be more familiar with standards and methods. There is a robust CST research literature and several generations of widely researched CST forensic assessment instruments (FAIs, Heilbrun, Rogers, & Otto, 2002). There are significant methodological differences between CST and mental state at the time of the offense (MSO) evaluations. Mental state at the time of the offense evaluations are retrospective in nature; they present a more challenging and complex psycholegal task (Acklin, 2007a; Melton et al., 2007; Roesch, Viljoen, & Hui, 2004), since they are more inferential and rely on sources of retrospective information. Nevertheless, levels of agreement between examiners and the court are comparable with previous studies. In this study, in 77.5% (n = 38, 76 reports) of cases, the court and at least two of three evaluators agreed on criminal responsibility. Robinson and Acklin (2010) found that in 90% of cases (N = 45), judges agreed with two or more evaluators on CST. Nguyen et al. (2011) found that in 67% (N = 33) of cases, the court agreed with at least two of the evaluators on release decisions.

Failure to link clinical information to expert psycholegal opinion is one of the most common weaknesses of forensic reports (Wettstein, 2005). Although Grisso (2010) identified opinions without sufficient explanations as a common error, this did not emerge as a major deficit in this study. A significant percentage of reports met the quality criterion for psycholegal opinion. The Psycholegal Rationale element cluster had a mean QC score of 76.67% with 74 reports meeting the 80% criterion. The QC score for the psycholegal opinion item was 94.67% and 84.00% for the Psycholegal Rationale item. This is a substantial improvement over Nguyen and colleagues' findings of 5% of reports meeting the 80% criterion on conditional release evaluations (2011). Inclusion of statements related to psycholegal impairment and rationale were more frequent when compared to Robinson's and Acklin's 2010 study of CST report quality. Robinson and Acklin (2010) found that 82% of reports included an explanation of the defendant's impairments and 74% included a rationale statement. In the current study, 91.3% of reports included a statement of cognitive and/or volitional impairments and 72.0% provided a complete rationale statement.

4.2. Agreement between evaluators

Our findings suggest that levels of agreement between evaluators and the court is a continuing concern. The panel of evaluators have access to the same information (i.e., records provided by the court), a scenario in which one might ordinarily expect high level of agreement in opinions (Acklin, 2007b; Robinson & Acklin, 2010). In our current study, unanimity was reached in 49% of criminal responsibility reports between all evaluators and in all but one case, at least two of three evaluators reached the same recommendation (98%). Agreement between all three evaluators and judiciary was achieved in 40.8% or 20 cases; agreement between the consensus of evaluators and judiciary occurred in 77.5% (N = 38 cases). The 49% rate of unanimous agreement is slightly lower than the 55% unanimous inter-rater agreement rate found in a similar sample of sanity evaluators in Hawaii (Gowensmith et al., 2013). In their study of CST evaluations in Hawaii, Gowensmith et al. (2012) found that 70.9% of evaluators agreed on CST opinions, with the court agreeing with a majority of evaluators (2 of 3) in 92.5% of cases reviewed. Similarly, Robinson and Acklin (2010) identified 70% (35 cases) agreement between all three evaluators on CST opinions and 94% (47 cases), two of the three evaluators agreed on CST opinions. These findings indicate that agreement among criminal responsibility evaluators in routine practice is lower than agreement on CST evaluations, and also supports the previously noted observation that criminal responsibility evaluations tend to be more complex and challenging than CST evaluations.

In this study, the expected "good" agreement rate (Cicchetti, 1994) was not achieved; instead "fair" levels of agreement (Cicchetti, 1994) were found. Robinson and Acklin (2010), using the more rigorous Landis and Koch (1977) criteria found "moderate" levels of agreement ($\kappa = .56$) between all evaluators and "substantial" levels of agreement ($\kappa = .67$) between evaluators and the court. Nguyen et al. (2011) found overall "poor" levels of agreement (Cicchetti, 1994) between all evaluators (ICC = .06) and the evaluators and court (ICC = .30) for conditional

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

release recommendation. Substantially better agreement was found in this study compared to Nguyen et al. (2011) and somewhat lower agreement compared to Robinson and Acklin (2010).

The consequences of poor inter-rater reliability are not merely academic. Facing felony charges is a high stakes situation that may permanently alter a defendant's life. Funder (1990) wrote that "the study of accuracy in judgment is exactly the same thing as measurement validity, where the measurements being validated are interpersonal judgments." From this perspective, "a personality judgment is accurate to the extent that it agrees with judgments of others and predicts relevant behaviors" (John & Robins, 1994, p. 208). Since reliability of judgment is validity, poor agreement decreases accuracy and increases the probability of error in decision-making. Accuracy of judgment insures that insane defendants are treated differently in the criminal justice system than guilty defendants. False positive and negative judgment errors (an insane defendant is wrongly prosecuted or a guilty defendant is wrongly exculpated) may have serious, adverse consequences in the lives of individuals. Judgments about criminal responsibility have serious consequences.

There are significant differences in complexity and task requirements between CST, criminal responsibility, and conditional release evaluations. As noted above, CST evaluations may be the least complex in comparison to retrospective MSO evaluations and conditional release evaluations which necessitate assessing risk of violence. In addition, the statutory language surrounding conditional release is less clear and less operationally defined than criminal responsibility and competency to stand trial (Nguyen et al., 2011).

Despite our expectation of good level of agreement, the finding of "fair" agreement (ICC = .49–.51, Cicchetti, 1994) is generally in accordance with the limited literature regarding agreement rates in forensic evaluations, most of which were conducted on CST evaluations. These studies typically found fair to high rates of agreement (Golding et al., 1999; Gowensmith et al., 2012, 2013; Nguyen et al., 2011; Robinson & Acklin, 2010; Skeem, Golding, Berge, & Cohn, 1998). Meyer, Mihura, and Smith (2005) performed a meta-analysis of inter-rater agreement across a wide range of common medical and scientific tests. They observed the relationship between complexity of the test and levels of agreement. Tasks involving "circumscribed judgment" (e.g., physical measurements) are generally more reliable than "complex tasks requiring synthesis of multiple, higher order inferences" (Meyer et al., 2005, p. 310). Judgment of static or unchangeable "objects" demonstrated clearer association with higher levels of inter-rater agreement than "dynamic tasks," such as interviewing on separate occasions (Meyer et al., 2005). Mental state at the time of the offense evaluations is complex, requiring the evaluator to retrospectively integrate a multitude of data and decipher a complicated psycholegal question (Acklin, 2007a). It is possible that our expectation of "good" agreement (>.80) was too stringent in a addressing as complex a forensic question as criminal responsibility.

4.3. Potential threats to reliability across forensic evaluations

The clinical and forensic literature describes the myriad factors that are potential threats to the inter-rater reliability, including changes in observation environment, different evaluation methods, training of raters, and reactive effects of observation (Acklin, 2007b; Haynes, 1978). Accuracy and the processes of accurate judgment have been a major concern of social psychology. Funder has identified four components and moderators that affect the likelihood of making an accurate judgment: "(1) properties of the judge, (2) properties of the target individual who is being judged, (3) properties of the trait that is being judged, and (4) properties of the information on which the judgment is made" (Funder & Fast, 2010, p. 683).

In the current study, examiners conducted their evaluations independently at different times over the course of days or weeks. The defendant may have refused to talk to some evaluators and not others; situational stressors or destabilizers, or variability in clinical condition could have occurred between evaluator visits; the clinical presentation of the defendant may change as an order effect in the evaluation process; there may be medication, medical, or dynamic inter-examiner or situational issues affecting presentation. Consequently, the defendant may have reported different narratives due to variable mental health symptomology (i.e., thought disorder) or in an attempt to present in a particular manner. Examiner and impression management effects cannot be ruled out. These factors contribute to "the dynamic nature of the judgment task" (varying defendant presentation) discussed by Meyer et al. (2005) where complexity of task is associated with lower inter-rater agreement.

Method variance (lack of standardization in the evaluation methodology) is a critical threat to inter-rater reliability. Method variance could be addressed by training (Robinson & Acklin, 2010; Gowensmith et al., 2012) and standardization in evaluation process (e.g., use of templates or checklists as guides; Robinson & Acklin, 2010; Witt, 2010) and use of FAIs. The use of FAIs, for example, the Rogers Criminal Responsibility Scales (RCRAS, Rogers, 1984), was low in this study. In the process of structured professional judgment, FAIs aid the clinician in organizing relevant information in relation to pertinent legal standards (Goldstein, 2007). Increased use of FAIs has the potential to improve report quality and level of agreement (Robinson & Acklin, 2010; Gowensmith et al., 2012; Otto & Heilbrun, 2002). Variability within examiners (within group variance) appears to have played an important role in the current study. In complex evaluations, sources of variability are numerous. Acklin (2007b) highlighted the following threats to reliability: complexity and definitional specificity of the target construct, training of raters, and idiosyncratic factors associated with the evaluation process (environmental effects, examiner–defendant interaction effects, examinee dispositional effects, specific target behaviors, and rater bias). Only a small minority of evaluators used psychological assessment measures of any kind (8 out of 25 evaluators). Drawing group wide generalizations obscures important sources of variability within the comparison groups. Findings are related more to the characteristics of a subset of evaluators within each group as opposed to true group differences.

In a recent study examining decision models in forensic examiners, Mossman (2013) examined statistical models of decision-making and concluded "this article has shown why imperfect accuracy, random error, and different decision thresholds should be included among potential explanations" for examiner disagreement (p. 53). "Even when evaluations are performed by unbiased examiners who generate well-correlated internal ratings of defendants' capacities, random errors will make some disagreement" about forensic capacities inevitable (p. 53).

Procedurally, criminal responsibility evaluations in Hawaii are similar to MSO evaluations recommended in the literature or conducted elsewhere (Giorgi-Guarnieri et al., 2002; Melton et al., 2007). Like Hawaii, MSO evaluations conducted across the nation take place in similar settings (e.g., state hospitals, jails, and outpatient settings). Many of the quality elements examined here were relevant to forensic report quality in general. The use of sections, inclusion of historical information and informed consent are not specific to criminal responsibility but related to general forensic report quality.

Forty-four reports (29%) were repeat evaluations, with the previous report cited as a data source. There is the potential that information previously included may not have been included a second time, thus lowering report quality. Data was not collected on whether or when the previous evaluation occurred, whether it was related to the same case, same psycholegal question, or time frame. Subsequent reports may have represented more of an ongoing narrative with the court regarding the defendant's psycholegal dispositions, especially if the evaluation occurred in the same case or psycholegal question. Even if it would be different cases or several years apart, certain information may have been omitted from subsequent reports. Historical items, for example, are one area that may not have been repeated particularly

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

because of the static nature of the information (i.e., schooling, vocational history, psychiatric history). In these cases, several evaluators chose to present either a summary of data from a previous report, an update for "pertinent" information, or a reference back to the previous report.

This study excluded certain examiner or report characteristics. Coding for date of the evaluations, charges, offender characteristics (gender, ethnicity, age), and psychiatric diagnoses may have permitted a more nuanced and complete picture. This is a limitation in comparing Hawaii reports to other report quality studies.

The QC scale used in this study utilized equal item weights; items equally contributed to the overall and element QC scores. This is consistent with previous studies utilizing this coding instrument (Nguyen et al., 2011; Robinson & Acklin, 2010; Sanschagrin, 2006). The question of element or item rating—what the prioritization of items or element weights in report quality should be—is an issue for further research. At this point of time, one cannot say with empirical certainty that the inclusion of the psycholegal opinion and a rationale is more important, for example, than a clearly described history, use of plain language, clearly delineated sections, or even the name and age of the defendant. Intuitively, it would seem that items such as diagnosis and psycholegal opinion would be of primary importance to judges and attorneys (Grisso, 2003). Research is consistent in showing that judges are specifically interested in clinical diagnosis, psycholegal opinion, rationale, and conceptual linkages (Redding, Floyd, & Hawk, 2001). To address the weighting of quality items, a useful extension to the current study would be surveys of judges to rate the most important elements.

4.4. Improving the quality of forensic evaluation reports

As far as we are aware, this is the first study in the forensic report quality literature to specifically address criminal responsibility. Other types of forensic mental health evaluators have yet to be evaluated for quality, including discharge from conditional release, sentencing mitigation, guardianship assessments, or child custody evaluations (Wettstein, 2005). These investigations would likely require modification of the quality coding instrument utilized in the current study.

Quality improvement can be facilitated by self-monitoring of forensic clinicians conducting forensic evaluations. Murrie and Warren (2005) describe the following self-monitoring steps as: 1) keeping current on research in insanity field and comparing own rates to average; 2) reviewing personal methodology and variation within own cases and determinations; 3) monitoring partisan bias—how many times did the decision agree with the attorney or disagree. Self-monitoring tasks would serve to increase the reliability and accuracy of insanity evaluations (Murrie & Warren, 2005). Feedback has been shown to increase rater accuracy and agreement (Acklin, 2007b). Feedback from the court to the evaluators regarding judicial outcome and opinions of other three panel members could further assist the evaluator rigor and accuracy. Feedback on the utility of forensic reports submitted would allow the forensic clinician to evaluate whether they are including information useful to the court, such as clear rationales for opinions (Redding et al., 2001; Robinson & Acklin, 2010). Research has shown that observers are more accurate when they are aware that their performance is being observed and monitored (Taplin & Reid, 1973). Wettstein (2005) discusses peer review as a component of credentialing for forensic experts. Peer review of evaluator work samples is part of the process for specialty certification through the American Board of Forensic Psychology (Grisso, 2010). Peer review could serve to increase consistency, report quality, and evaluator agreement. Formal quality assurance programming (Wettstein, 2005) and performance feedback (Farkas, DeLeon, & Newman, 1997) would likely improvement the standardization and quality reports.

Witt (2010) suggests that clinicians use a forensic report checklist, as an *aide memoir* to guide the clinician thought the main points. Witt (2010) proposed that forensic clinicians can use identified common errors, such as lack of third party data, lack of explanation for opinion, lack of plain language and organization as a starting point for the checklist. Ethical and statutory requirements would also be beneficial to include in a checklist. Common criticisms, such as inconsistent use of psychological testing, failure to assess factors related to the legal issue at hand, lack of use of third party data, and lack of a linkage between clinical findings and legal questions and capacities, could be addressed through specific training and standardized processes, such as templates or checklists (Grisso, 2010; Nicholson & Norwood, 2000). Development of templates and standardized report formats would increase consistency of reports, thereby increasing communication to the courts. Templates are not meant to decrease individual evaluation styles, writing styles or preferences, rather to increase standardization to aid the court more effectively. Standardization of report format, or at least a general checklist, would increase report quality through increased likelihood of appropriate attention and inclusion of forensic report elements (Robinson & Acklin, 2010; Grisso, 2010; Witt, 2010).

Rogers and Shuman (2000) suggest that the state of "insanity evaluations have largely been an idiosyncratic process, reflecting the propensities and proclivities of the clinician" (p. 520). Exposure to periodic training could serve to improve report quality and decrease evaluator idiosyncratic methods. Robinson and Acklin (2010) reported increases in report quality following training in CST reports. Mandated training may be an essential element in quality improvement, including exposure to statutory language, court mandates, and generally accepted forensic best practices and standardized procedures of evaluations and forensic reports. Training provides exposure to case law developments and developing practice standards described in the developing forensic mental health literature.

References

Acklin, M. W. (2007a). The Rorschach test and forensic psychological evaluation: Psychosis and the insanity defense. In C. Gacano, & B. Evans (Eds.), *The handbook of forensic Rorschach assessment* (pp. 157–174). New York, NY: Routledge/Taylor & Francis Group.

Acklin, M.W. (2007b). *Quantifying consensus: methodological issues in competency to stand trial evaluations*. Unpublished manuscript.

Acklin, M., Kennedy, R., Robinson, R., Dunkin, B., Dwire, J., & Lees, B. (2005). Quality of forensic reports: An empirical investigation of three panel reports. *Paper presented at the Annual State of Hawaii Forensic Mental Health Examiner Training Conference, Honolulu, HI.*

Allnutt, S. H., & Chaplow, D. (2000). General principles of forensic report writing. *Australian and New Zealand Journal of Psychiatry*, *34*(6), 980–987.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*(1; 1), 84–94.

Babitsky, S., & Mangraviti, J. (2002). *Writing and defending your expert report*. Falmouth, MA: SEAK, Inc.

Blau, G., McGinley, H., & Pasewark, R. (1993). Understanding the use of the insanity defense. *Journal of Clinical Psychology*, *49*(3), 435–440.

Borum, R. (1994). Standards and practices in forensic evaluation. *American Psychology Law Society News*, *14*(2), 2–4.

Borum, R., & Grisso, T. (1995). Psychological test use in criminal forensic evaluations. *Professional Psychology: Research and Practice*, *26*(5), 465–473.

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

Cohen, J. W. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Connell, M. (2008). Ethical issues in forensic psychology. In R. Jackson (Ed.), *Learning forensic assessment* (pp. 55–72). New York, NY: Routledge/Taylor & Francis Group.

Conroy, M.A. (2006). Report writing and testimony. *Applied Psychology in Criminal Justice*, *2*(3), 237–260.

Daniel, A. E., & Harris, P. W. (1981). Female offenders referred for pre-trial psychiatric examination. *Bulletin of the American Academy of Psychiatry and the Law*, *9*, 40–47.

Farkas, G. M., DeLeon, P. H., & Newman, R. (1997). Sanity examiner certification: An evolving national agenda. *Professional Psychology: Research and Practice*, *28*(1), 73–76.

Fukunaga, K. K., Pasewark, R. A., Hawkins, M., & Gudeman, H. (1981). Insanity plea: Inter-examiner agreement and concordance of psychiatric opinion and court verdict. *Law and Human Behavior*, *5*, 325–328.

Funder, D. (1990). Process versus content in the study of judgmental accuracy. *Psychological Inquiry*, *1*, 207–2090.

Funder, D., & Fast, L. (2010). Personality in social psychology. In S. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed.). New York: Wiley & Sons.

Gagliardi, G. J., & Miller, A. K. (2008). Writing forensic psychological reports. In R. Jackson (Ed.), *Learning forensic assessment* (pp. 539–563). New York, NY, US: Routledge/Taylor & Francis Group.

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020

Giorgi-Guarnieri, D., Janofsky, J., Keram, E., Lawsky, S., Merideth, P., Mossman, D., et al. (2002). AAPL practice guidelines for forensic psychiatric evaluation of defendants raising the insanity defense. *The Journal of American Academy of Psychiatry and the Law*, *30*(2), S3–S40.

Golding, S., Skeem, J., Roesch, R., & Zapf, P. (1999). The assessment of criminal responsibility: Current controversies. In A. Hess, & I. Weiner (Eds.), *The handbook of forensic psychology* (pp. 379–408) (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Goldstein, A. (2007). Forensic psychology: Towards a standard of care. In A. Goldstein (Ed.), *Forensic psychology: Emerging topics and expanding roles* (pp. 3–41). Hoboken, NJ: John Wiley & Sons, Inc.

Gowensmith, W. N. (2008). *Overview of forensic evaluations in Hawaii*. Paper presentation at the 5th Annual State of Hawaii Forensic Examiner Training, Kaneohe, HI.

Gowensmith, W. N. (2011, April 24) personal communication.

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*. http://dx.doi.org/10.1007/s10979-010-9259-8 (December 2010).

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.

Greenberg, J., & Wursten, A. (1988). Psychologist and the psychiatrist as expert witnesses: Perceived credibility and influence. *Professional Psychology: Research & Practice*, *19*(4), 373–378.

Grisso, T. (1986). *Evaluating competencies. Forensic assessments and instruments*. New York: Plenum.

Grisso, T. (2003). *Evaluating competencies. Forensic assessments and instrument* (2nd ed.) New York: Kluwer Academic/Plenum.

Grisso, T. (2010). Guidance for improving forensic reports: A review of common errors. *Open Access Journal of Forensic Psychology*, *2*, 102–115 (Retrieved from www.forensicpsychologyunbound.ws/-2010.2:102-115)

Hans, V., & Slater, D. (1983). John Hinckley, Jr. and the insanity defense: The public's verdict. *Public Opinion Quarterly*, *47*, 202–212.

Harvey, V. S. (1997). Improving readability of psychological reports. *Professional Psychology: Research and Practice*, *28*(3), 271–274.

Hawaii Revised Statute HRS 704-400-404 (na). Retrieved from http://www.capitol.hawaii.gov/site1/hrs/default.asp

Haynes, S. N. (1978). *Principles of behavioral assessment*. Oxford, England: Gardner.

Hecker, J. E., & Scoular, R. J. (2004). Forensic report writing. In W. T. O'Donohue, & E. R. Levensky (Eds.), *Handbook of forensic psychology: Resource for mental health and legal professionals* (pp. 63–81). New York, NY, US: Elsevier Science.

Hecker, T., & Steinberg, L. (2002). Psychological evaluation at juvenile court disposition. *Professional Psychology: Research and Practice*, *33*(3), 300–306.

Heilbrun, K. (2001). *Principles of forensic mental health assessment*. New York: Kluwer Academic/Plenum Publishers.

Heilbrun, K., & Collins, S. (1995). Evaluations of trial competency and mental state at time of offense: Report characteristics. *Professional Psychology: Research and Practice*, *26*(1), 61–67.

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

Heilbrun, K., Rogers, R., & Otto, R. (2002). Forensic assessment: Current status and future directions. In J. R. P. Ogloff (Ed.), *Taking psychology and law into the twenty-first century* (pp. 119–146). New York, NY: Kluwer Academic/Plenum Publishers.

John, O., & Robins, R. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206–219.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research & Practice*, *34*(5), 491–498.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (2007). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (3rd ed.) New York NY: Guilford Press.

Meyer, G. J., Mihura, J. L., & Smith, B.L. (2005). The interclinician reliability of Rorschach interpretation in four data sets. *Journal of Personality Assessment*, *84*, 296–314.

Mossman, D. (2013). When forensic examiners disagree: Bias, or just inaccuracy? *Psychology, Public Policy, and Law*, *19*(1), 40–53.

Murrie, D., Boccaccini, M., Zapf, P., Warren, J., & Henderson, C. (2008). Clinician variation in findings of competence to stand trial. *Psychology, Public Policy & Law*, *14*(3), 177–193.

Murrie, D., & Warren, J. (2005). Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice*, *36*(5), 519–524.

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.

Nicholson, R. A., & Norwood, S. (2000). The quality of forensic psychological assessments, reports, and testimony: Acknowledging the gap between promise and practice. *Law and Human Behavior*, *24*(1), 9–44.

Otto, R. K., & Heilbrun, K. (2002). The practice of forensic psychology: A look to the future in light of the past. *American Psychologist*, *57*(1), 5.

Pasewark, R. A., McGinley, H., & Blau, G. (1989). Insanity plea: Influence of psychiatric opinion. *Journal of Police and Criminal Psychology*, *5*, 29–32.

Quinsey, V. (2009). Are we there yet? Stasis and progress in forensic psychology. *Canadian Psychology*, *50*(1), 15–21.

Redding, R. E., Floyd, M. Y., & Hawk, G. L. (2001). What judges and lawyers think about the testimony of mental health experts: A survey of the courts and bar. *Behavioral Sciences & the Law*, *19*(4), 583–594.

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

Roesch, R., Viljoen, J., & Hui, I. (2004). Assessing intent and criminal responsibility. In W. T. O'Donohue, & E. R. Levensky (Eds.), *Handbook of forensic psychology: Resource for mental health and legal professionals* (pp. 157–174). New York, NY: Elsevier Science.

Rogers, R. (1984). *Rogers criminal responsibility assessment scales*. Lutz, FL: Psychological Assessment Resources Inc.

Rogers, R. (2008). Insanity evaluations. In R. Jackson (Ed.), *Learning forensic assessment* (pp. 109–128). New York, NY: Routledge/Taylor & Francis.

Rogers, R., & Shuman, D. W. (2000). *Conducting insanity evaluations*. New York: Guilford.

Sanschagrin, K. A. (2006). The quality of forensic mental health assessments of juvenile offenders: An empirical investigation. ProQuest Information & Learning. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *66*(11), 6292.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Simon, R. I., & Shuman, D. W. (2002). *Retrospective assessment of mental states in litigation: Predicting the past*. Washington, DC: American Psychiatric Press.

Skeem, J. L., & Golding, S. L. (1998). Community examiners' evaluations of competence to stand trial: Common problems and suggestions for improvement. *Professional Psychology: Research and Practice*, *29*(4), 357–367.

Skeem, J. L., Golding, S. L., Berge, G., & Cohn, N.B. (1998). Logic and reliability of evaluations of competence to stand trial. *Law and Human Behavior*, *22*(5), 519–547.

Specialty guidelines for forensic psychologists (1991). *Law and Human Behavior*, *15*(6), 655–665.

Specialty guidelines for forensic psychology (2011). *Adopted by American Psychological Association Council of Representatives*.

Taplin, P., & Reid, J. (1973). Effects of instructional set and experimental influence on observer reliability. *Child Development*, *44*(3), 547–554.

Viljoen, J., Roesch, R., Ogloff, J., & Zapf, P. (2003). The role of Canadian psychologists in conducting fitness and criminal responsibility evaluations. *Canadian Psychology*, *44*(4), 369–381.

Warren, J. I., Fitch, W. L., Dietz, P. E., & Rosenfeld, B.D. (1991). Criminal offense, psychiatric diagnosis, and psycholegal opinion: An analysis of 894 pretrial referrals. *Bulletin of the American Academy of Psychiatry and the Law*, *19*, 63–69.

Warren, J. I., Murrie, D. C., Chauhan, P., Park, B.S., Dietz, M.D., & Morris, J. (2004). Opinion formation in the evaluating sanity at the time of the offense: An examination of 5175 pre-trial evaluations. *Behavioral Sciences & the Law*, *22*, 171–186.

Warren, J. I., Murrie, D. C., Stejskal, W., Colwell, L. H., Morris, J., Chauhan, P., et al. (2006). Opinion formation in evaluating the adjudicative competence and restorability of criminal defendants: A review of 8,000 evaluations. *Behavioral Sciences & the Law*, *24*(2), 113–132.

Weiner, I. B. (2006). Writing forensic reports. In I. B. Weiner, & A. K. Hess (Eds.), *The handbook of forensic psychology* (pp. 631–651) (3rd ed.). Hoboken, NJ, US: John Wiley & Sons Inc.

Wettstein, R. M. (2004). The forensic examination and report. In R. I. Simon, & L. H. Gold (Eds.), *The American psychiatric publishing textbook of forensic psychiatry* (pp. 139–164). Washington, DC, US: American Psychiatric Publishing, Inc.

Wettstein, R. M. (2005). Quality and quality improvement in forensic mental health evaluations. *Journal of the American Academy of Psychiatry and Law*, *33*(2), 158–175.

Witt, P. (2010). Forensic report checklist. *Open Access Journal of Forensic Psychology*, *2*, 233–240 (Retrieved from www.forensicpsychologyunbound.ws/-2010.2:233-240)

Zapf, P., Golding, S., & Roesch, R. (2006). Criminal responsibility and the insanity defense. In I. Weiner, & A. Hess (Eds.), *The handbook of forensic psychology* (pp. 332–363) (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

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

Please cite this article as: Fuger, K.D., et al., Quality of criminal responsibility reports submitted to the Hawaii judiciary, *International Journal of Law and Psychiatry* (2013), http://dx.doi.org/10.1016/j.ijlp.2013.11.020