Field Reliability of Competence to Stand Trial Opinions: How Often Do Evaluators Agree, and What Do Judges Decide When Evaluators Disagree?

W. Neil Gowensmith · Daniel C. Murrie · Marcus T. Boccaccini

Abstract Despite many studies that examine the reliability of competence to stand trial (CST) evaluations, few shed light on “field reliability,” or agreement among forensic evaluators in routine practice. We reviewed 216 cases from Hawaii, which requires three separate evaluations from independent clinicians for each felony defendant referred for CST evaluation. Results revealed moderate agreement. In 71% of initial CST evaluations, all evaluators agreed about a defendant’s competence or incompetence (kappa = .65). Agreement was somewhat lower (61%, kappa = .57) in re-evaluations of defendants who were originally found incompetent and sent for restoration services. We also examined the decisions judges made about a defendant’s CST. When evaluators disagreed, judges tended to make decisions consistent with the majority opinion. But when judges disagreed with the majority opinion, they more often did so to find a defendant incompetent than competent, suggesting a generally conservative approach. Overall, results reveal moderate agreement among independent evaluators in routine practice. But we discuss the potential for standardized training and methodology to further improve the field reliability of CST evaluations.

Does the outcome of a forensic psychological evaluation depend on which evaluator performed the evaluation? Many of the systems that require forensic evaluations appear to take reliability for granted, essentially assuming that evaluators are interchangeable. For example, state psychiatric hospitals often rely on a single clinician to evaluate a defendant’s trial competence and mental state at the time of offense, and prisons often rely on a single clinician to assess a potential parolee’s risk of violence. In contrast, many judges, attorneys (e.g., Krafa et al., 2002; Shuman, Whitaker, & Champagne, 1994), and legal scholars (e.g., Bernstein, 2008; Perlin, 2004) are skeptical that forensic mental health professionals can reach reliable and objective opinions.

Some skepticism is understandable. Forensic evaluators’ opinions might be influenced by differences in their social or political attitudes (Homant & Kennedy, 1987), the party requesting their services (Murrie et al., 2009), or other unidentified individual differences (Boccaccini, Turner, & Murrie, 2008; Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008; Murrie & Warren, 2005). Although these threats to reliability may be especially common in field settings where evaluators have varied training and experience, most studies addressing the reliability of forensic evaluations come from controlled research settings. For many forensic issues, we know more about uniformly trained research assistants and case vignettes than about practicing forensic evaluators and criminal defendants. Yet the “field reliability” of forensic evaluations has tremendous implications for fairness and accuracy in the justice system.
When asking questions about the field reliability of forensic evaluations, the best place to start is to examine evaluations of competence to stand trial (CST). CST evaluations are the most common form of criminal forensic evaluations (Golding, 1992; Melton, Petrila, Poythress, & Slobogin, 2007), with an estimated 60,000 defendants evaluated each year in the United States (Bonnie & Grisso, 2000), or approximately 5% of all felony defendants (Hoge, Bonnie, Poythress, & Monahan, 1992). CST evaluations are also among the most reliable types of forensic evaluations, authorities suggest (Melton et al., 2007; Zapf & Roesch, 2006). So CST evaluations may represent the best-case scenario for evaluators to show high levels of field reliability.

Research Addressing the Reliability of CST Evaluations

Many studies report excellent reliability in evaluations of CST, but they probably reveal more about CST instruments than about CST evaluators. When similarly trained raters use structured measures of CST (usually based on the same interview), rates of agreement range from good to excellent (Rogers & Johansson-Love, 2009), albeit with stronger agreement for overall competent/incompetent status than for individual scales of the CST instruments (e.g., Golding, Roesch, & Schreiber, 1984; Roesch & Golding, 1980). Instrument-focused studies are important in order to understand the psychometric properties of particular CST instruments, and to gauge whether clinicians can come to reliable conclusions when using the same instrument. But none examine “field reliability,” or whether clinicians do come to reliable conclusions in routine practice.

Of the few studies that do examine clinicians in practice, the first two revealed excellent agreement among clinicians who offered opinions about whether or not a particular defendant was competent to stand trial (Goldstein & Stone, 1977; Poythress & Stock, 1980). But, clinicians in these early studies appear to have worked and trained together in the same setting, with the same institutional standards of practice. Thus, they were probably not independent evaluators, in the practical or statistical sense of the term. In at least one of the studies (Poythress & Stock, 1980), colleagues conducted the CST interview together, ensuring that they considered identical interview content. In the other (Goldstein & Stone, 1977), evaluators may have consulted with each other to arrive at opinions. Even authorities who are generally optimistic about the reliability of CST evaluations caution that reliability may be weaker among independent clinicians across different institutions or “evaluation centers” (Melton et al., 2007, p. 144).

A more recent glimpse of reliability in CST evaluations comes from a study that reviewed 100 CST reports—two evaluations for each of 50 defendants—produced by a sample of 18 Utah clinicians (Skeem, Golding, Cohn, & Berge, 1998). Evaluators demonstrated fair agreement regarding whether a defendant was competent or incompetent (82% agreement), and even regarding the general diagnostic category that best applied to the defendant (79% agreement). However, agreement was much weaker (around 25% on average) regarding whether a defendant was impaired on specific psycho-legal skills necessary for trial competence. Indeed, evaluator agreement fell below 10% regarding almost half of these skills (Skeem et al., 1998).

Most recently, Mossman et al. (2009) reported reliability data from five experienced forensic clinicians who rated 156 redacted CST reports (authored by other clinicians) on the degree to which the defendant described in each report demonstrated the capacities necessary for CST. When offering ratings about psycho-legal abilities—rather than the dichotomous conclusion of competent or incompetent—the five participating clinicians tended to respond similarly. But as the authors emphasize, the study findings may not answer questions about the field reliability of CST determinations in wide-scale practice. Even if clinicians make similar appraisals about a defendant’s underlying psycho-legal abilities, they may not offer the same dichotomous conclusions about competence. This is because the ultimate competent-versus-incompetent decision involves value judgments and perhaps external pressures, Mossman et al. (2009) stated.

The observation from Mossman et al. (2009) that clinicians may achieve weaker reliability on overall CST opinions than on underlying CST abilities differs from the observation from Skeem et al. (1998), who observed fair agreement (82%) regarding the overall CST opinion, but weaker agreement (around 25% on average) regarding whether a defendant was impaired on specific underlying abilities. However, the Mossman et al. (2009) and Skeem et al. (1998) observations are not necessarily contradictory: The former reflects how participant clinicians agreed about a defendant’s capacities documented in another evaluator’s report; the latter reflects researcher appraisal of what two different clinicians described in two different reports documenting two different CST interviews. Nevertheless, these contrasting observations underscore just how little we know about how and why clinicians (dis)agree in routine practice.

Towards Better Understanding the Field Reliability of CST Evaluations

The studies reviewed above suggest some scenarios under which reliability for CST evaluations can be strong: when
clinicians who work and train in the same setting evaluate
the same defendants simultaneously (Poythress & Stock,
1980), when clinicians form opinions based on another
clinician’s written report of an evaluation (Mossman et al.,
2009), and when clinicians use the same instrument to
score the same evaluation (Rogers & Johannsson-Love,
2009). However, other studies suggest that clinicians may
differ in terms of the opinions they tend to reach (Murrie
et al., 2008); agreement may be less strong in routine
practice (Skeem et al., 1998), and especially poor in
ambiguous cases (Morris, Haroun, & Naimark, 2004).
Given these mixed findings, it remains difficult to draw
conclusions about the field reliability of CST opinions
under typical, routine circumstances (i.e., when evaluators
conduct separate evaluations with the same defendant and
report independent findings via separate reports).
A study that might better help us draw conclusions about
field reliability would require identifying a “real world”
context in which defendants are routinely evaluated by
independent evaluators, who do not work together, and
who separately submit their evaluation findings. Of course,
if such a context revealed any disagreement among eval-
uators, it would also prompt questions about how courts
decide cases when evaluators reach different conclusions
about a defendant’s CST.

Study Context

Hawaii has arranged forensic evaluation services in ways
that make the state ideal for a naturalistic study of reli-
ability among forensic evaluators. According to statute
(Hawaii Revised Statutes 2003, sections 704-404 and 704-406),
the court orders a “three-panel exam” for felony
cases in which questions about CST arise: that is, three
concurrent and independent evaluations, each requiring a
written evaluation report. Statute requires that one evaluation
report be completed by a Department of Health
(DOH) state evaluator (all of whom were psychologists at
the time of this study), while the other two are to be
completed by licensed and “certified” independent practi-
tioners not employed by the DOH.¹ One of the two
independent practitioners must be a psychiatrist, while the
other evaluator may be either a psychiatrist or psychologist.
In practice, the third evaluator is almost always a psy-
chologist. The court chooses the independent evaluators
from a list of certified evaluators supplied by the DOH, and
then determines the location for the evaluation. Evaluation

¹ Evaluators become “certified” if the state forensic administrator
approves their background credentials. Evaluators may also attend
DOH-sponsored annual trainings on forensic mental health
assessment.

locations include correctional facilities, inpatient hospitals,
or community settings. All three evaluators concurrently
and independently review pertinent records, interview the
defendant, complete testing as needed, and submit the
reports directly to the court. Evaluators are prohibited by
statute from collaborating or sharing information with each
other during the evaluation process. More so than any other
state, Hawaii has arranged their forensic service delivery
system to rely on multiple, concurrent, independent eval-
uations (Gowensmith, 2010) in a way that is uniquely
conducive to studying field reliability.

Recently, Robinson and Acklin (2010) provided a
glimpse into CST reliability in Hawaii by examining 150
CST reports (50 CST cases) submitted to Oahu’s Circuit
Court. Although their primary goal was to study the quality
of evaluation reports, they briefly reported evaluator
agreement in their sample. All three evaluators agreed
about a defendant’s trial competence in 35 of the 50 cases
they reviewed (70%). The study authors did not discuss the
implications of agreement findings, explore reasons for
disagreement, or distinguish between agreement rates for
initial versus subsequent CST evaluations. Thus, their
results suggest a level of evaluator agreement we might
expect in a context where defendants are routinely evalu-
ated by independent evaluators, but also prompt questions
that warrant further study.

In this study, we examine in detail a more recent
statewide sample of 729 CST evaluations to capitalize on
Hawaii’s unique forensic service delivery system, which
allows for the largest available study of CST “field reli-
ability” to date.

Method

Sample

During our study period (September 1, 2007–December 31,
2008), Hawaii felony courts received multiple concurrent
CST evaluation reports for 247 defendants. We were able
to obtain at least two of the concurrent CST reports for 216
of these defendants (see “Discussion” section for impli-
cations of missing reports). Of course, the court can order an
initial competence evaluation (i.e., the first time CST is
addressed for a particular defendant facing a particular
charge) as well as subsequent competence evaluations (i.e.,
CST evaluations ordered after a defendant who was origi-
really found incompetent has received some competence-
restoration services, and requires re-evaluation). This study
separately examines both initial and (when pertinent)
subsequent evaluations that were submitted during the
study period. When assigning evaluators to subsequent
evaluations of a defendant, courts in Hawaii usually
(though not always) reassign the same evaluators that completed the initial evaluations on that defendant.

Of the 216 defendants with multiple evaluator opinions available for review, 182 had results from an initial evaluation. Of these, 157 had results from an initial evaluation only, but 25 had results from an initial and a subsequent evaluation. During this study period, 34 defendants had multiple evaluator reports from a subsequent evaluation, but not from an initial evaluation. Thus, there were 59 defendants with at least one subsequent evaluation.

Overall, across the 216 defendants in this study, there were 251 cases (i.e., CST evaluation orders that resulted in separate, concurrent evaluation reports completed by up to three evaluators). In this study, “case” refers to the collection of the three separate evaluation reports (i.e., one case of a three-panel evaluation for a particular defendant) that address CST for the same charge(s) in response to the same court order. “Evaluation report” refers to one of the evaluation reports within a “three-panel” case.

There were three evaluator opinions available for 227 cases, and only two available for 24 cases; thus, a total of 729 CST evaluation reports available for review. We opted to include the 24 cases with only two evaluator reports because they were originally ordered as three-panel evaluations, thus not qualitatively different from the cases in which all three reports were available for review. In these 24 cases, the court could not locate the third report, but this apparently occurred at random, with no discernible pattern or explanation. Conceivably, including these cases could slightly inflate estimates of agreement (e.g., we found that two evaluators agreed in 17 of the 24 cases, but it is possible the missing evaluator disagreed in some of these). Nevertheless, we found similar results whether or not we included these cases in the study analyses. The 729 reports were authored by a total of 35 evaluators.

Most of the defendants undergoing CST evaluations (n = 154, 71.3%) were male, and most (n = 182, 84.3%) spoke English as their primary language. Mean age was 38.87 (SD = 12.70) years. The sample was ethnically diverse, with 34.3% (n = 74) identified by collateral data or self-report as White/Caucasian, 16.2% (n = 42) as Asian, 13.4% (n = 29) as Native Hawaiian/Pacific Islander, 2.8% (n = 6) as Black/African American, 0.9% (n = 2) as Hispanic, and 6.0% (n = 13) of an unknown ethnicity. The remaining 23.1% (n = 50) defendants were identified as multiethnic, with most (n = 35) identifying both Native Hawaiian/Pacific Islander and another ethnic background.

Evaluators

The group of DOH evaluators included seven licensed psychologists who were employed as forensic evaluators by the Hawaii DOH during the study period. As a group they completed 251 evaluation reports (M = 36.14, SD = 22.21). The group of independent evaluators included 28 certified evaluators who maintained private practices and operated as part-time forensic evaluators. This category of evaluators comprised 13 licensed psychologists who completed 221 reports (M = 16.85, SD = 14.39), and 15 licensed psychiatrists who completed 257 reports (M = 17.07, SD = 21.01).

Measures

Data for this study were collected from evaluation reports using a form that addressed: (i) defendant characteristics such as date of birth, ethnicity, gender, level of charge(s), and English as a primary language, (ii) evaluator characteristics such as professional discipline and affiliation (i.e., state DOH evaluator versus independent evaluator), (iii) evaluation characteristics such as date of evaluation report, referring court, and sequence of the evaluation, and (iv) content and conclusions of the evaluation report such as assigned diagnoses, rule-out qualifiers, and recommendations on competency.

The data was collected by two doctoral-level psychologists and one clinical psychology graduate research assistant. Inter-rater agreement was initially measured by reviewing a sample of 10 evaluation reports. The evaluation report coders identified areas of ambiguity and refined the form before coding another 10 evaluation reports, which yielded 100% coding agreement on all items. All three coders reviewed every hundredth evaluation report to ensure continued inter-rater agreement throughout the data collection, and maintained 100% agreement.

The research team collected the 729 evaluation reports from existing court records and state DOH evaluation report files. The team identified these 729 reports by reviewing the grand total of 1651 forensic evaluation reports submitted during the study period. We excluded 922 reports from the study because they did not reference competency to stand trial (n = 341) or they were single evaluator “one-panel” misdemeanor cases (n = 581; typically, courts order only one evaluation for cases involving misdemeanor charges).

Diagnoses

We recorded the diagnoses that evaluators had assigned and documented in evaluation reports, following the DSM-IV-TR classification system. Diagnoses were limited to psychotic disorders, mood disorders, substance-related disorders, cognitive disorders, and mental retardation, because these are the diagnoses most often related to CST opinions (Pirelli, Gottdiener, & Zapf, in press).
Evaluation Procedures

We recorded which evaluators used formal forensic assessment instruments to assess trial competence. Of 35 evaluators in the sample, five (14.3%, two DOH evaluators and three independent evaluators) used formal competency assessment measures. Those evaluators were among the most prolific of the group, so 169 of the 729 evaluation reports in the sample (23.2%) documented at least one formal competency assessment measure. These included the Competency Assessment Instrument (CAI; Laboratory of Community Psychiatry, 1973); Evaluation of Competency to Stand Trial-Revised (ECST-R; Rogers, Tillbrook, & Sewell, 2004), and the Fitness Interview Test-Revised (FIT-R; Roesch, Zapf, & Eaves, 2006).

Charge Type

We coded charges into three categories: “Felony A” (including murder, attempted murder, manslaughter, rape, and attempted rape), “other violent charge,” and “non-violent charge.” For defendants with multiple charges, we used the category for the most serious charge. For initial CST evaluations, more than half of the defendants (n = 106, 58.2%) were charged with other violent charges, and about a third (n = 63, 34.6%) were charged with non-violent offenses. There were only 10 (5.5%) initial evaluation defendants with Felony A charges, and three for whom we could not identify the type of charge. The charge type distribution was similar for subsequent CST evaluations (52.5% other violent, 37.3% nonviolent, 10.2% Felony A).

Psycho-legal Recommendations

We coded CST opinions as competent, incompetent, or unknown. A total of 13 evaluation reports were coded as “unknown,” because the evaluator either declined to offer an opinion or was so unclear that we could not code the opinion with certainty.2

Court Disposition

We coded the court disposition that the judge offered immediately after the hearing that addressed the evaluation report, even if the court made a different decision at a later time (i.e., a later finding of competence after an initial finding of incompetence). Court dispositions were collected and coded from existing court records or through a public online judicial database.

Results

Overall Rates of Evaluator Opinions and Court Dispositions of Competency

In initial evaluations, evaluators opined that 72.7% of defendants were competent to stand trial, 25.4% were incompetent, and 1.9% unknown. For these initial cases, courts ruled 68.1% CST, 30.2% IST, and 1.5% unknown. In subsequent evaluations, evaluators opined 48.0% of defendants were CST, 48.0% were IST, and 4.0% unknown. On these subsequent cases, courts found 35.6% CST and 62.7% IST, and again 1.5% unknown.

Evaluator Agreement and Case Dispositions for Initial Evaluations

In 70.9% of initial CST evaluation cases, all evaluators agreed about the defendant’s CST status (i.e., competent versus incompetent), as detailed in Table 1. The free-marginal multirater kappa coefficient for initial evaluations was .65 (Randolph, 2008), which falls within the range typically considered “good,” but not “excellent,” reliability (Cicchetti & Sparrow, 1981). Free-marginal kappa is appropriate when evaluators are not forced to assign a certain number of cases to each category (Brennan & Prediger, 1981; Randolph, 2008), which is the case for CST evaluations (i.e., evaluators are not required to find 50% of defendants competent). Fixed-marginal kappa, which is more commonly reported (e.g., Fleiss, 1980) and assumes evaluators assign a certain number of cases to each condition, was .54 (95% CI [.14, .94]) for initial evaluations.

Evaluators more often agreed that the defendant was competent than incompetent: 58.8% of all initial evaluations ended with all three evaluators agreeing the defendant was competent, whereas 12.1% ended with all three concluding the defendant was incompetent. These proportions appear consistent with the overall base rate of opinions for initial evaluations in the dataset (73.5% competent vs. 26.5% incompetent). When evaluators disagreed about a defendant’s CST status, the disagreement was similarly likely to have resulted from only one evaluator finding the defendant to be incompetent (13.7% of all initial evaluations) as only one evaluator finding the defendant to be competent (11.5% of all initial evaluations).

How did courts decide a defendant’s CST status when evaluators reported contradictory opinions? Court dispositions were generally consistent with the majority opinion among evaluators (see Table 1). Overall, judges agreed

2 Cases in which one or more evaluator opinions were coded as “unknown” contributed to the “other disagreement” categories listed in Tables 1 and 2.
with the majority evaluator opinion (i.e., the opinion that at least 2 of 3 evaluators offered) in 92.5% of cases (97.7% of unanimous agreement cases, 78.3% of disagreement cases). However, there was a trend in the dispositions which suggested that courts were more willing to go against a majority opinion of competence than a majority opinion of incompetence. For example, courts found defendants to be competent in only 4.8% of cases in which two of three evaluators concluded the defendant was incompetent, but found 32.0% of defendants to be incompetent when two of three evaluators concluded that the defendant was competent. Courts never found a defendant to be competent when all three evaluators unanimously concluded that the defendant was incompetent, but courts did find two defendants (1.9%) incompetent when all evaluators agreed the defendant was competent.

### Evaluator Agreement and Case Dispositions for Subsequent Evaluations

Whereas evaluators agreed about CST in 70.9% of the initial CST evaluations, agreement was slightly lower, at 61.0% (see Table 2) for subsequent competence evaluations (i.e., those re-evaluations ordered for defendants who were initially found incompetent and then received competency restoration services). The free-marginal multirater kappa coefficient for subsequent evaluations was .57 (Randolph, 2008), as was the fixed-marginal kappa coefficient (95% CI [.07, 1.00]), which is considered “fair” agreement (Cicchetti & Sparrow, 1981). The percentage of cases in which all three evaluators unanimously agreed that the defendant was competent was much lower for subsequent evaluations (32.2%) than initial evaluations (58.8%). Similarly, there were more subsequent evaluation cases in which all three evaluators unanimously concluded that the defendant was incompetent (28.8%) compared to initial evaluations (12.1%). Again, this finding is fairly consistent with the overall base rate of opinions in subsequent evaluations (50.6% competent vs. 49.4% incompetent). As with initial evaluations, disagreement for subsequent evaluations was nearly as likely to have resulted from only one evaluator finding the defendant to be incompetent (13.6% of all subsequent evaluations) as only one evaluator finding the defendant to be competent (15.3% of all subsequent evaluations).

As with initial evaluations, court dispositions were generally consistent with the majority opinion among evaluators (see Table 2). Judges agreed with the majority opinion in 77.4% of cases (86.1% of unanimous agreement cases, 58.8% of disagreement cases). The lone exception to this trend occurred when courts went against majority opinion to rule a defendant incompetent in six of the eight cases (75.0%) in which two of the three evaluators concluded that the defendant was competent. Consistent with the general trend in initial CST evaluations, this finding from subsequent evaluations also suggests that courts were more willing to go against a majority opinion of competence than a majority opinion of incompetence. For example, courts never found a defendant to be competent when all three evaluators unanimously concluded that the defendant was incompetent, but courts did find the defendant to be incompetent (21.0%) when all three evaluators unanimously agreed the defendant was competent.

### Reasons for Disagreement

**Evaluator Characteristics.** Disagreement among evaluators did not appear to be attributable to evaluator discipline (psychologist vs. psychiatrist) or evaluator
Table 2  Evaluator agreement and case dispositions for subsequent competence to stand trial evaluations

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Cases</th>
<th>% of cases</th>
<th>Court disposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>% CST</td>
</tr>
<tr>
<td>Agreement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All agree competent</td>
<td>19</td>
<td>32.2</td>
<td>73.7</td>
</tr>
<tr>
<td>All agree incompetent</td>
<td>17</td>
<td>28.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Disagreement</td>
<td>23</td>
<td>39.0</td>
<td></td>
</tr>
<tr>
<td>Two competent, one incompetent</td>
<td>8</td>
<td>13.6</td>
<td>25.0</td>
</tr>
<tr>
<td>One competent, two incompetent</td>
<td>9</td>
<td>15.3</td>
<td>11.1</td>
</tr>
<tr>
<td>One competent, one incompetent</td>
<td>2</td>
<td>3.4</td>
<td>50.0</td>
</tr>
<tr>
<td>Other disagreement</td>
<td>4</td>
<td>6.8</td>
<td>75.0</td>
</tr>
</tbody>
</table>

Note. \( N = 59 \). For offenders with more than one subsequent evaluation, only the earliest evaluation was considered for this table. CST Competent to stand trial, IST incompetent to stand trial. Other disagreement cases were those in which one or two evaluators reported that they could not determine whether the defendant was or was not competent. Underlined values in the court disposition columns indicate values for which the disposition was inconsistent with the majority opinion of the evaluators. Free-marginal kappa for cases with three opinions is .57. Fixed-marginal Kappa is .57, 95% CI [.07, 1.00]

Employer (state DOH vs. private practice). Across the 46 initial evaluations involving disagreement, the disagreeing expert was the state DOH evaluator in 41.3% (\( n = 19 \)) of cases, the independent psychologist in 26.1% (\( n = 12 \)) of these cases, and the independent psychiatrist in 32.6% (\( n = 15 \)) of these cases \( \chi^2 (2, N = 46) = 1.61, p = .45 \).

We also considered whether some individual evaluators may have been more likely than others to be the disagreeing expert (i.e., offer a minority opinion). To examine this possibility, we compared the 10 evaluators who performed more than 20 initial evaluations. Evaluators’ individual base rates for being the disagreeing evaluator ranged from 0.0% (\( n = 22 \) evaluations) to 14.6% (\( n = 41 \) evaluations). The mean disagreement base rate was 7.66% (\( SD = 4.56 \)), with very little skewness (.08), indicating that there was a relatively normal distribution of disagreement rates (i.e., no evaluator had an exceptionally high or low rate of disagreement with other evaluators).

Characteristics of Judges. Of the 25 judges who made dispositions, only five did so in more than 15 cases. Each of these five judges disagreed with the majority evaluator opinion in at least one case, with rates of disagreement ranging from 3.2% to 15.2% per judge, and most falling between 10.0 and 15.0%. Thus, no judges seemed unusually likely to disagree with the majority evaluator opinion.

Use of Competency Instruments. We attempted to examine whether evaluators who used competency assessment instruments were more or less likely to agree about CST. However, only nine cases featured two evaluators using a CST instrument (\( n = 4 \) initial evaluation cases, \( n = 5 \) subsequent evaluation cases); this low number prevented us from examining whether agreement between evaluators improved when both used competency assessment instruments.

Defendant Diagnosis. Did evaluators disagree about competence because they disagreed about the defendant’s diagnosis? Because psychotic disorders are the most commonly diagnosed disorders among those found incompetent (Pirelli et al., in press), it may be that evaluators disagreed about competence when they disagreed about whether the defendant qualified for diagnosis of a psychotic disorder. To examine this issue, we calculated the number of cases in which the evaluators agreed that the defendant qualified for a psychotic disorder diagnosis, the number in which they agreed that the defendant did not qualify for a diagnosis, and the number in which they disagreed about whether the defendant qualified for the diagnosis. We then used chi-square analyses to examine the relationship between this three category diagnostic agreement variable and a three category competence agreement variable (all agree competent, all agree incompetent, disagree about competence).

The findings for initial evaluations are summarized in Table 3. Disagreement about competence was not associated with disagreement about psychotic disorder diagnoses. Although evaluators disagreed about psychotic disorder diagnoses in 20.7% of the cases in which they disagreed about competence, they also disagreed about psychotic disorder diagnoses in about the same proportion of cases in which they agreed that the defendant was competent (26.2%) or agreed that the defendant was incompetent (27.3%). Perhaps the most compelling number in Table 3...
is that all evaluators agreed that the defendant qualified for a psychotic disorder diagnosis in 50.9% of the cases in which they disagreed about the defendant’s competence. Thus, agreeing upon diagnosis of a psychotic disorder did not necessarily lead those evaluators to agree on competency.

The statistical significance of the chi-square statistic was mainly attributable to the finding that evaluators who agreed that the defendant was incompetent were more likely to have agreed that the defendant qualified for a psychotic disorder diagnosis (54.5%) than did not qualify for a psychotic disorder diagnosis (18.2%). Likewise, evaluators who agreed that the defendant was competent were more likely to have agreed that the defendant did not qualify for a psychotic disorder diagnosis (59.8%) than did qualify for a psychotic disorder diagnosis (14.0%). This is not surprising, given that psychotic disorders are the most commonly diagnosed psychiatric disorders in defendants found incompetent (Pirelli et al., in press).

We performed similar analyses across other diagnostic categories commonly associated with CST; however, no clear pattern emerged for any diagnostic category with respect to evaluators’ disagreement on CST. Evaluators disagreed about the presence of a mood disorder in 32% of cases, a substance-related disorder in 22% of cases, and a cognitive disorder or mental retardation in 22.7% of cases, but these disagreements were similarly common whether evaluators agreed or disagreed about trial competence.

**Defendant Ethnicity.** The three most common ethnic categories of defendants were Caucasian (n = 60), Asian (n = 41), and Native Hawaiian/Pacific Islander descent (n = 54). A chi-square analysis between these ethnic groups and evaluator agreement (all agree CST, all agree IST, disagree) indicated that agreement did not differ significantly depending on defendant ethnicity ($\chi^2 (4, N = 155) = 5.52, p = .24$, Cramer’s $V = .19$).

**Severity of Alleged Offense.** We used chi-square analyses to examine the relation between offense severity (Felony A, other violent felony charge, nonviolent felony charge) and evaluator agreement. We examined agreement as a dichotomous variable (agree vs. disagree) and as a three category variable (all agree competent, all agree incompetent, disagree). None of the analyses revealed statistically significant relations, although these analyses were limited in power due to the small number of Felony A cases. For initial evaluations, there was an overall trend for disagreement (dichotomized) to increase as crime severity increased. Evaluators disagreed in 50% of the 10 Felony A cases, 30.1% of the 106 other violent felony cases, and 20.6% of the 63 nonviolent offense cases [$\chi^2 (2, N = 179) = 4.52, p = .10$, Cramer’s $V = .16$]. For subsequent evaluations, evaluators disagreed in 50.0% of the six Felony A cases, 71.0% of the 31 other violent felony cases, and 50.0% of the 22 nonviolent offense cases [$\chi^2 (2, N = 59) = 2.72, p = .26$, Cramer’s $V = .22$].

Were judges more likely to disagree with the majority opinion in cases involving more serious charges? For initial evaluations with known dispositions, judges disagreed with the majority opinion in 30.0% of the 10 Felony A cases, 9.9% of the 101 other violent felony cases, and 10.0% of the 60 nonviolent offense cases [$\chi^2 (2, N = 171) = 4.32, p = .12$, Cramer’s $V = .16$]. In two of the three Felony A cases, judges went against majority opinions of competency, instead finding the defendants incompetent. For subsequent evaluation cases, judges disagreed with the majority opinion in 20.0% of the five Felony A cases, 6.9% of the 29 other violent felony cases, and 16.7% of the 18 nonviolent offense cases [$\chi^2 (2, N = 52) = 1.43, p = .49$, Cramer’s $V = .17$].

**Discussion**

The primary contribution of this study is to shed light on the reliability of CST evaluations in routine forensic practice. Overall, the study revealed that agreement among evaluators is the rule, not the exception. Evaluators agreed unanimously on their CST opinions in 70.9% of initial CST cases, and in 61.0% of subsequent CST cases. Also, courts tended to rule on CST cases in the direction of the majority evaluator opinion (89.0% of cases). Despite this general pattern of agreement, some disagreement among evaluators, as well as court decisions that were contrary to majority evaluator opinions, underscore important

---

*Note. Percentage values calculated separately for each row*
questions about how judges and policy makers might best respond to the reality of imperfect reliability among forensic evaluators in the field.

**Evaluator Agreement in the Field**

First, these results do not seem to support the bleakest view of CST evaluations, which compares them to “flipping coins in the courtroom” and asserts that the “defendant’s fate depends only upon who performed the evaluation” (Morris et al., 2004, p. 216; see Perlin, 2004 for discussion). The finding that 71% of initial evaluations yielded perfect agreement across three evaluators reveals that agreement among evaluators is possible, and even commonplace, in “real world” CST cases. In short, evaluators in routine practice appear to agree about a defendant’s trial competence in roughly two-thirds of cases. Practicing forensic evaluators might not be surprised by these findings, which seem to suggest that the majority of CST evaluations are fairly straightforward (such that different evaluators should arrive at the same clear conclusion), but a substantial minority are more difficult, ambiguous, or otherwise complicated (such that different evaluators may arrive at different conclusions).

Of course, it is this minority of cases that result in evaluator disagreement—around one-third in the current sample—that raises concerns. Not surprisingly, our sample revealed poorer agreement than an early study of co-working evaluators who co-conducted interviews (Poithress & Stock, 1980). But our results appear comparable to those from the most similar available study: a review of 50 cases in Utah (Skeem et al., 1998). Disagreement among mostly “trios” of evaluators in Hawaii (29% disagreement in initial evaluations) was somewhat more common than disagreement among pairs of evaluators (18%) in Utah. Although jurisdictions likely differ, and more studies of this sort are necessary, we might estimate that around 20–30% of all CST cases result in divergent opinions among independent evaluators.

**What Explains Disagreement about CST?**

We could identify no idiosyncratic evaluator(s) who disagreed significantly more often than others. Likewise, we could identify no obvious evaluator characteristics (e.g., discipline or employer) that appeared to explain the disagreement among evaluators. Neither the defendant’s ethnicity nor the severity of their charge(s) explained evaluator disagreement. Finally, disagreement among evaluators was not simply a matter of disagreeing about diagnosis. Early reviews of forensic evaluation (e.g., Roesch & Golding, 1980) lamented that evaluators erroneously equated diagnoses (rather than functional abilities) with CST status; in these situations disagreements about diagnosis would likely lead to disagreements about CST status. However, evaluators in our study did not appear to simply equate CST status with diagnostic status. Their rates of disagreement about the presence of a psychotic disorder were about the same whether or not they disagreed about that defendant’s competency. Moreover, in about half of the cases in which evaluators disagreed about a defendant’s competency, they agreed that the defendant had a psychotic disorder. Thus, evaluators—as we might hope—were likely considering the functional abilities necessary for trial, rather than simply diagnosis.

Although we could not identify any clear or consistent explanations for disagreement among evaluators, there are potential explanations we could not examine. For example, we did not code the extent to which evaluators addressed the defendant’s functional abilities that underlie trial competence (i.e., understanding charges, appreciating charges, reasoning, assisting counsel, making rational decisions). Skeem et al. (1998) reported that Utah evaluators often disagreed about these underlying abilities. Similarly, although we examined the general diagnoses evaluators assigned, we could not examine the exact psychiatric symptoms that evaluators considered most salient in a given case. Examining the link between psychiatric symptoms and functional abilities is the crucial task underlying CST evaluations (Zapf & Roesch, 2009). So evaluators who disagree about a defendant’s psychiatric symptoms, a defendant’s functional abilities, or the link between the two, may therefore disagree about the defendant’s overall trial competence. In this study, we could not examine these complex inferences. Indeed, these inferences are difficult to examine in any naturalistic study because evaluators often fail to describe them in their reports (Christy, Douglas, Otto, & Petrila, 2004; Skeem & Golding, 1998). Finally, differences in evaluator experience or methodology may have contributed to some disagreements. Although some degree of evaluator disagreement is probably inevitable, all of these potential explanations for disagreement warrant more in-depth study.

**What Might Improve Evaluator Agreement?**

**CST Instruments?** In this sample, formal competence assessment instruments were not common enough for us to examine whether they improved evaluator agreement. Proponents of formal CST assessment instruments may be quick to notice that evaluator agreement in this naturalistic reliability study was much poorer than evaluator agreement in reliability studies designed to validate CST instruments (see, e.g., review by Rogers & Johannsson-Love, 2009). However, there are far more
differences between this naturalistic study and most instrument-development studies than the use of instruments alone. Evaluators in our sample performed evaluations separately, but raters in instrument-development studies typically score instruments simultaneously based on the same interview. Evaluators in our sample comprised a heterogeneous group, but raters in instrument-development studies typically share similar education or training. Thus, we cannot claim that the strong reliability values from CST instrument-development studies would necessarily generalize to the field.

But more widespread use of formal CST instruments certainly might improve reliability among evaluators. If some disagreement in this study is due to disagreement regarding the functional abilities that underlie trial competence (see above, and Skeem et al., 1998), using CST instruments that address and operationalize those abilities may enhance agreement. If some disagreement in this study is due to variability in evaluator methodology (see Robinson & Acklin, 2010), adopting formalized instruments system-wide would, of course, decrease variability in methodology and potentially improve reliability. However, adopting CST instruments may not substantially improve reliability unless there is also consistency in the ways evaluators are trained on instruments, implement these instruments, and integrate instruments into other components of a comprehensive assessment.

Generally, this study reveals the need for more research on CST instruments in the field. Researchers who study instruments will need to conduct more ecologically valid reliability studies, particularly among evaluators using these instruments in routine practice. Likewise, researchers who study forensic assessment “in the field” will need to better examine the role of instruments, and their potential for improving the field reliability of forensic evaluations. Future studies might examine whether reliability is stronger among evaluators who use CST instruments in the field versus evaluators who do not. If these instruments do improve field reliability, it becomes important to investigate barriers to their use and ways of promoting more widespread, consistent use.

**Mandated Forensic Training?** Is disagreement inevitable because some cases are inherently complex and some legal standards are imprecise? Or can we improve training, certification, and oversight of forensic evaluators to the extent that disagreements are exceedingly rare? There may be some merit in both perspectives. A few cases may be so unusually complex, or “grey” (Golding, 2008), that even skilled, reasonable professionals will disagree. Indeed, if we assume that the abilities necessary for trial competence lie along a continuum, we might expect to find some disagreement among clinicians regarding cases that fall near the midpoint of this continuum (Murrie et al., 2008). However, recent studies of forensic assessment in the field suggest more variability among evaluators than seems attributable to a few complex cases alone (e.g., Boccaccini et al., 2008; Murrie et al., 2008). Thus, increases in mandated training, certification, or oversight may be appropriate steps to reduce gross evaluator idiosyncrasies.

Evaluators in this sample were certified by the Hawaii Department of Health. But as in most states, Hawaii’s certification process has historically been a brief gate keeping function that cannot guarantee consistency across evaluators. There remains heterogeneity among evaluators in terms of education, specialized forensic training, prior forensic experience, and evaluation methodology. Current certification requirements include attending DOH forensic trainings and providing work samples for review. However, some veteran evaluators who predate current certification procedures were essentially “grandfathered” into the evaluator pool. New evaluators have completed updated certification requirements, but may not have the extensive experience common among veteran evaluators. To the extent that certification procedures become more rigorous and intensive, field reliability among evaluators is likely to improve. Indeed, the recent study of report quality in 50 Hawaii CST cases suggested that evaluators who attended a particular DOH training produced reports that better addressed critical content areas (Robinson & Acklin, 2010). As more evaluators attend similar trainings, agreement may improve.

**Judicial Response to Evaluator Disagreement**

Although a primary goal of this study was to address agreement among CST evaluators across cases, a second study goal was to shed light on how judges make decisions about a defendant’s trial competence when evaluators do not agree. Several studies reveal that courts routinely concur with the opinions forensic evaluators offer about a defendant’s trial competence. One study reported a 99.6% agreement rate between the court’s disposition regarding CST and the forensic evaluator’s opinion (Zapf, Hubbard, Galloway, Cox, & Ronan, 2004), and all others report at least 90% agreement (e.g., Cruise & Rogers, 1998; Hart & Hare, 1992). In our study, judges agreed with the majority opinion among evaluators in most initial evaluations (92.5% of cases), but the concordance rate fell to 77.4% for subsequent evaluations of competence. Reasons for the lower concordance rate are unknown, but probably reflect the challenging nature of cases in which a defendant has already been found incompetent and then provided restoration services. Evaluators less often agree in these cases, and judges face complicated choices about defendants with
substantial deficits who may never be “restored” to competence.

Faced with discrepant findings from evaluators, judges in our sample tended to take a conservative approach. That is, they leaned towards protecting the defendant’s right to participate meaningfully in trial, even at the risk of unnecessarily delaying trial and providing defendants with more hospitalization and restoration services than necessary. This was especially clear in those cases in which judges went against a unanimous opinion of competency from evaluators and ruled the defendant incompetent (1.9% of initial evaluations and 21.0% of subsequent evaluations); judges never went against a unanimous evaluator opinion of incompetency. Of course, we do not know if the tendencies we observed among judges in Hawaii generalize to other jurisdictions.

Policy Implications

Just as individual judges must decide cases with discrepant opinions from evaluators, policy makers must decide how to handle the reality that forensic evaluators sometimes disagree. Some cases are complex, and evaluators are not entirely interchangeable. Although the precise agreement rates we observed in Hawaii may not generalize to other jurisdictions, a similar pattern probably does. That is, independent evaluators probably agree about most cases, with a substantial minority of cases resulting in disagreement. To the extent that disagreement occurs, jurisdictions that rely on only one evaluator might “miss out” on an opposing perspective in some portion of cases. Indeed, to the extent that there is a factually correct answer to the question of whether or not a defendant is competent to meaningfully in that trial. Of course, majority opinion may not always be correct, and there are many practical limitations on how many evaluations a court can request. But, clearly courts that rely on only one evaluator will, at times, act upon evaluator opinions that would have been revealed as minority opinions if additional evaluations were available.

The justice system needs to continue to wrestle with how to handle these competing demands—efficient use of resources versus fair consideration of defendants’ right to due process—when issues of mental health come before the courts. Hawaii’s solution to this dilemma was to implement a “three-panel” system for forensic evaluations in felony cases, which belies an expectation that evaluators will disagree in some cases, and suggests a belief that courts can make better decisions when presented with multiple evaluations. However, few systems are likely to adopt such a resource-intensive approach. Another, more common, approach might be to ensure that a second evaluation is easily available when one party raises reasonable concerns about results from the first evaluation. Our finding (and others; Skeem et al., 1998) that evaluator disagreements are not uncommon certainly underscores the need for some mechanism to solicit a second opinion in complex cases.

Study Limitations and Future Directions

At the micro-level, there are a few limitations on the quality and quantity of the data we could examine for this study. First, 67 evaluation reports submitted to the court during the date range under study were not available for review due to a variety of logistical reasons. Including these missing reports in our larger sample of 729 reports would have been preferable, but we can identify no reasons to believe these missing data would systematically alter our results.

Second, we had little detailed information about the evaluators or their assessment techniques. Evaluator agreement may have been influenced by evaluator characteristics we could not analyze, such as specialized forensic training, extent of forensic evaluation experience, evaluation technique, or even broad evaluator characteristics such as socio-political values. Future research on field reliability should further examine evaluator characteristics that may contribute to evaluator differences.

At the macro-level, there are some limits to the generalizations we can offer. The very features that make Hawaii so ideal for this naturalistic study may suggest study results do not generalize well to all jurisdictions. The legal requirement for three CST evaluations suggests a willingness to invest liberally in evaluation services, as well as an expectation for some degree of disagreement among evaluators. Either of these might differ in other jurisdictions in ways that correspond to patterns of evaluator agreement. However, unless other systems invest in multiple independent evaluations, we simply cannot know the reliability among them.

A strength of this study was the evaluators’ independent affiliation (they were ordered by the court, rather than retained by prosecution or defense), which gives us a measure of evaluator agreement without the influence of an adversarial arrangement. However, research suggests that “adversarial allegiance,” a subtle pull for evaluators to
form opinions favorable to the party that retained them, may play a role in some forensic evaluator opinions (Murrie et al., 2008, 2009). So the reliability we observed among independent evaluators may not generalize to contexts in which evaluators are selected by opposing sides in a criminal case. This limitation is important because many systems allow for two CST evaluations only when each is initiated by opposing sides in adversarial proceedings. Future research should further explore reliability among forensic evaluators in adversarial settings.

Obviously, this study addressed only CST evaluations. It will be important to conduct similar studies that investigate field reliability in evaluations of legal sanity, violence risk, and other important criminal forensic evaluations. Indeed, naturalistic studies of field reliability are an essential first step in gauging wide-scale quality across all manner of forensic practice and targeting areas for improvement.

References


