# IMPLICATIONS OF STATIC-99 FIELD RELIABILITY FINDINGS FOR SCORE USE AND REPORTING

MARCUS T. BOCCACCINI
*Sam Houston State University*

DANIEL C. MURRIE
*University of Virginia*

CYNTHIA MERCADO
STEPHEN QUESADA
*John Jay College of Criminal Justice*

SAMUEL HAWES
AMANDA K. RICE
*Sam Houston State University*

ELIZABETH L. JEGLIC
*John Jay College of Criminal Justice*

The Static-99 is a well-researched measure used in many courtroom and correctional settings to help make decisions about sexual offenders. But, as with most forensic assessment measures, we know much more about interrater agreement for the Static-99 in formal research studies than in routine forensic and correctional practice. This study describes "field reliability" for the Static-99 in two states that use the Static-99 for routine correctional procedures, Texas ($N = 600$) and New Jersey ($N = 135$). Rater agreement coefficients were strong for Static-99 total scores (intraclass correlations = .79 and .88), but the total scores assigned by pairs of evaluators differed for approximately 45% of offenders in each state. Because each individual Static-99 score has a unique interpretation, and a 1-point difference in a Static-99 score can have substantial practical implications for decision making, these findings suggest the need for administration procedures or interpretation methods that acknowledge and account for measurement error in Static-99 total scores.

***Keywords:*** Static-99; field reliability; rater agreement; sexually violent predator; risk assessment

The Static-99 (Hanson & Thornton, 2000) is an actuarial risk assessment instrument designed to predict sexual recidivism among sex offenders. Comprising 10 items that address an offender's age, prior living arrangements, and several aspects of his offense history, the Static-99 can be completed by clinicians or nonclinical correctional staff on the basis of file review. According to Static-99.org, the Static-99 is "the most widely used sex offender risk assessment instrument in the world, and is extensively used in the United

**AUTHORS' NOTE:** *This research was supported in part by Grant No. NIJ 2007-IJ-CX-0037 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The project was also supported by the Texas Department of Criminal Justice, New Jersey Department of Corrections, and the New Jersey Department of Human Services. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the institutions that supported this research. Correspondence concerning this article should be addressed to Marcus T. Boccaccini, Psychology Department, Sam Houston State University, Huntsville, TX 77341; email: Boccaccini@shsu.edu.*

CRIMINAL JUSTICE AND BEHAVIOR, Vol. 39 No. 1, January 2012 42-58
DOI: 10.1177/0093854811427131
© 2012 International Association for Correctional and Forensic Psychology

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

States, Canada, the United Kingdom, Australia, and many European nations." At least 30 U.S. states use the Static-99 in formal procedures (Interstate Commission for Adult Offender Supervision, 2007), and 1 state even mandates use of the Static-99 in statute (Virginia Code Ann. § 37.2–903).

Static-99 scores influence decisions at various stages of criminal and civil proceedings as well as correctional procedures. Perhaps most visibly, nearly all evaluators administer the Static-99 in proceedings that address the civil commitment of sexual offenders as *sexually violent predators* (SVPs; Jackson & Hess, 2007). Several studies have found that those offenders whom SVP evaluators recommend for civil commitment, and those against whom states initiate commitment proceedings, have significantly higher Static-99 scores than those who are not recommended or pursued for civil commitment (Boccaccini, Murrie, Caperton, & Hawes, 2009; Levenson & Morin, 2008). But less visibly, Static-99 scores often influence a variety of decisions in correctional and community management contexts, such as risk level classification and community notification requirements. In criminal proceedings, Static-99 scores are common in most sex offender risk assessments, such as those requested to inform sentencing decisions (see, e.g., Bengston & Långström, 2007). In short, the Static-99 is ubiquitous in formal decisions about sex offender management.

The Static-99 is also by far the most researched sex offender risk assessment measure, with more than 60 available validity studies (see Hanson & Morton-Bourgon, 2009). Static-99 scores are moderate predictors of sexual recidivism, with the most recent Static-99 meta-analysis reporting a median Cohen's *d* effect size of .67 (Hanson & Morton-Bourgon, 2009).

Clinicians and systems that rely on Static-99 scores usually do so in one of three ways (see Static99.org). First, clinicians might report the observed recidivism rates among men in the instrument's normative samples with a particular score. Second, clinicians may report a "relative risk" interpretation for a Static-99 score, which reflects the reoffense risk associated with a particular score as compared to the "typical" score of 2. Third, systems may use cut scores to sort offenders into risk level groups, with offenders who score at or above a certain cut score being placed into a high-risk community supervision group or referred for an SVP evaluation.

## BACKGROUND

### THE MEANING OF A 1-POINT SCORE DIFFERENCE ON THE STATIC-99

One important feature of the Static-99, and its revision, the Static-99R (Helmus, Thornton, Hanson, & Babchishin, in press), is that each possible score has a unique interpretation. For example, forensic evaluators who refer to normative sample recidivism rates would note that a score of 4 is associated with a 5-year sexual recidivism rate of 7.7%, whereas a score of 5 is associated with a recidivism rate of 10.2%.[1] Likewise, a score of 4 is associated with a relative risk ratio of 1.89, whereas a score of 5 is associated with a relative risk ratio of 2.42. At the policy level as well, a 1-point difference in a Static-99 score can have substantial implications. For example, Static-99 scores are used in some settings to place offenders into one of several risk level groups. Although interpretation is the same for those whose scores fall within the same group, a 1-point difference could change the risk level grouping for offenders with scores near risk score cut points. For

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

example, a score of 4, but not 3, makes an offender with certain offenses eligible to be considered for civil commitment as an SVP in Virginia. Until November of 2008, a score of 4, but not 3, made an offender subject to community notification in Texas. Texas now uses results from a group of measures to determine risk level status, but a Static-99 score of 6 or higher, regardless of scores on other measures, usually makes offenders subject to community notification.

There are other measures for which a 1-point difference may at first appear to have substantial implications for legal decisions; for example, defendants with mental retardation, which usually requires an IQ score of 70 or below, are not eligible for the death penalty (*Atkins v. Virginia*, 2002). But standard reporting procedures for IQ scores include the use of confidence intervals (CIs), which acknowledge the potential impact of measurement error on an obtained IQ score. For example, the 95% CI for a Full Scale IQ score of 71 on the Wechsler Adult Intelligence Scale–IV (Wechsler, 2008)—which seems too high to qualify for a diagnosis of mental retardation—is 68 to 76. Thus, persons who receive a score of 71 may still qualify for the diagnosis if CIs are included in the interpretation. Authorities strongly encourage clinicians to report CIs for IQ scores (Sattler, 2008) and all test scores used to make decisions about individuals (Nunally & Bernstein, 1994). In contrast, clinicians typically report Static-99 scores without reference to measurement error or CIs.

CIs are the most common method for incorporating rater agreement findings into test score interpretation. Evaluators use CIs when interpreting test scores because observed scores "are not perfectly accurate" (Sattler, 2008, p. 112), and CIs "serve as a reminder that measurement error is inherent in all test scores" (Wechsler, 2008, p. 126). CIs are based on the standard error of measurement (SEM), which is calculated using a reliability coefficient (e.g., rater agreement coefficient) and the standard deviation of scores on the measure. The larger the rater agreement coefficient, the smaller the SEM and the smaller the CI.

## RATER AGREEMENT FOR STATIC-99 SCORES

How strong does rater agreement need to be for a measure, such as the Static-99, for which each score has its own unique interpretation? Although popular guides categorize agreement of .75 and higher as "excellent" (see Cicchetti, 1994), this level of agreement may not be sufficient for measures like the Static-99. Nunnally and Bernstein (1994) argue that when "decisions depend on very small score differences . . . it is difficult to accept any measurement error" and that if "important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum and a reliability of .95 should be considered the desirable standard" (p. 265). The most recent review of rater agreement coefficients for the Static-99 found a median rater agreement value of .90 (see Hanson & Morton-Bourgon, 2009; Helmus, 2008), suggesting that agreement for Static-99 scores often meets these higher standards.

Nevertheless, whereas reliability research typically addresses how often scores are *similar,* reliability research on the Static-99 must also consider how often total scores are *identical.* Static-99 interpretive resources do report rater agreement coefficients but do not provide users with information about how often evaluators disagree by 1 point or more other than by noting that raters "rarely disagree by more than one point in a Static-99 score" (Harris, Phenix, Hanson, & Thornton, 2003, p. 73).

Only two studies, both unpublished, have examined how common it is for two evaluators to assign identical total scores on the Static-99. Hanson (2001) examined agreement

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

between California evaluators who scored the Static-99 as part of SVP evaluations for 55 different offenders. The single-rater intraclass correlation coefficient (ICC) for absolute agreement ($ICC_{A,1}$) was .87, typically interpreted as indicating strong to excellent agreement (Cicchetti, 1994). However, raters assigned identical total scores in only 42.5% of the cases. Austin, Peyton, and Johnson (2003) described a reliability study conducted for the Pennsylvania Board of Probation and Parole, in which 220 offenders were randomly selected to be scored twice on the Static-99. Evaluators assigned identical total scores in only 40.9% of the 220 cases. These score disagreements would have led to differences in risk level classifications for 26.8% of the Pennsylvania offenders.

## FIELD RELIABILITY OF STATIC-99 SCORES

As with most psychological measures, most of the available information about the reliability of Static-99 scores comes from studies in which the measure was scored for research purposes rather than “real-world” use. Because most Static-99 studies focus on the predictive validity of Static-99 scores, they rarely report detailed information about the raters who provided scores and for whom agreement values were reported. However, it seems likely that many of the raters in these studies received uniform training and practiced scoring with supervision until they attained strong interrater reliability (as typical in most instrument validity studies). Examining reliability among trained research raters is important, because it reflects the degree of reliability that is possible for a measure in controlled circumstances and provides an optimal context to examine the predictive validity of a measure.

However, reliability among raters who have received identical training and who have practiced scoring a measure for research purposes may not generalize to evaluators in the field, who may have varied training and practice before scoring the measure for clinical, administrative, or forensic purposes. Field reliability, or reliability among professionals scoring the instrument as part of routine practice in the field, is important because the scores from evaluators in the field are the scores that influence actual decisions about offenders. Compared to the many Static-99 studies that report agreement among research raters, we have fewer studies that report reliability among pairs of evaluators assigning scores for real-world legal or correctional proceedings. All three Static-99 field reliability studies we could identify examined agreement among SVP evaluators, with the first being Hanson’s (2001) unpublished study from California ($ICC_{A,1}$ = .87, identical scores in 42.5% of cases). In a similar, but larger, field reliability study, Levenson (2004) reported a strong rater agreement coefficient (ICC = .85) for Static-99 total scores among 281 offenders evaluated for SVP commitment in Florida. In both of these studies, all Static-99 scores came from state-contracted evaluators who were making determinations as to whether an offender should be considered by the state for commitment. In the most recent field reliability study, Murrie et al. (2009) examined agreement among Texas evaluators who scored the Static-99 for SVP cases that went to trial. Agreement for Static-99 total scores was modest ($ICC_{A,1}$ = .58 to .64) for 27 offenders scored by opposing evaluators (state vs. respondent) and for 30 offenders scored by two evaluators testifying for the state ($ICC_{A,1}$ = .61; reported in Boccaccini et al., 2009). The Static-99 study that comes closest to being a full field reliability study (i.e., both scores come from field evaluators) outside of the SVP context found strong agreement (ICC = .91) between field scores from community supervision officers and those from “expert raters” (e.g., R. K. Hanson, A. J. R. Harris; see Hanson, Harris, Scott, & Helmus, 2007, p. 10).

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

Although two of the existing field reliability studies suggest relatively strong agreement among field evaluators (i.e., Hanson, 2001; Levenson, 2004), the other (Murrie et al., 2009) suggests rater agreement may be weaker in some instances, such as high-stakes SVP cases that proceed to trial. But even Hanson's (2001) findings show that a strong rater agreement coefficient of .87 should not lead to the assumption that evaluators usually assign identical scores. Of course, none of these studies describe field reliability for the most commonly used Static-99 scores, such as those assigned by correctional staff, who score the measure to assist departments of corrections and parole boards in making release, supervision, and community management decisions. Indeed, in Texas alone, the Static-99 is scored for all sex offenders who may qualify for sex offender registration, only a subset of whom are screened or formally evaluated for civil commitment. We know relatively little about rater agreement in these contexts.

## PRESENT STUDY

We examined the field reliability of Static-99 item and total scores from more than 700 sex offenders, including 600 evaluated by correctional staff in Texas for civil commitment screening and community notification purposes and 135 evaluated for civil commitment by doctoral-level evaluators in New Jersey. We calculated rater agreement coefficients (ICCs), SEM values, and 95% CIs for Static-99 total scores, and we examined the frequency with which both sets of total scores are identical and, thus, lead to identical interpretation. A finding that total scores are identical in most cases would support the current practice of reporting Static-99 scores "as is," without reference to measurement error. However, a finding that scores are often not identical might bode for using CIs or other methods to account for measurement error when reporting Static-99 scores.

We also examined rater agreement for item scores to determine whether scoring discrepancies could be prevented by removing a less reliable item or items. The two existing (but unpublished) studies that report agreement for each Static-99 item report relatively strong agreement for all items, with percentage agreement values above 76% for all items and above 85% for most items (Austin et al., 2003; Hanson, 2001), suggesting that removing an item or items will not alleviate the need to account for measurement error in Static-99 total score interpretation.

This study represents, by far, the largest rater agreement study of Static-99 scores. It may also reflect the most ecologically valid study of Static-99 reliability, in that we examine reliability "in the field," in conditions comparable to routine practice. There are, however, two inevitable limitations to the field data used for this study. First, in both states, the two Static-99 evaluations often occurred at different points in time, because the Static-99 was scored when policy or procedure mandate it must be scored. The amount of time that elapsed between sets of field scores raises the issue of whether scores in our study may be different for legitimate scoring reasons. Because offenders were not released into the community during the time between the first and second score, this seems unlikely, unless an offender had his 25th birthday, committed an additional sex offense while in custody, or information on a new offense became available between the time of the first and second administration. Each scenario usually leads to a higher score.[2] In the current study, we control for age-related changes when possible, and examine whether disagreements appear to reflect that scores from the second administration tend to be higher than those from the first. A second limitation is that the second evaluator may have seen the first evaluator's

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

scores. Although this is a limitation in terms of obtaining a truly independent reliability estimate, this scenario reflects routine practice in the field and must be considered in any discussion of “field reliability.” If field evaluators were not blind to scores from prior evaluators, this might inflate agreement, and we should not be surprised to find higher reliability in such conditions.

## METHOD

### SAMPLES AND PARTICIPANTS

*Texas.* We obtained Static-99 item and total scores from Texas Department of Criminal Justice (TDCJ) files for 1,202 offenders who were screened for SVP civil commitment between 1999 and 2003 but not committed. Approximately half (*n* = 635, 52.8%) had been scored on the Static-99 on two occasions. We excluded 11 offenders who passed their 25th birthday between evaluations, which would have required a scoring change on the Static-99 item addressing age. We also removed 24 offenders because we were missing information about either date of birth or date of evaluation, which made it impossible for us to consider whether a score change might have been attributable to a change in age. The final sample consisted of 600 male offenders, with a mean age of 42.89 years (*SD* = 11.75). Offenders were identified as Caucasian (*n* = 275, 45.8%), Hispanic (*n* = 206, 34.3%), African American (*n* = 109, 18.2%), or Other (*n* = 10, 1.7%).

The average time between the first and the second Static-99 administration was 13.67 (*SD* = 9.74) months. The Static-99 scores in offender files came from three different sources: SVP screening, Risk Assessment Review Committee (RARC), and Institutional Parole Office (IPO) evaluations. Each offender was scored on the Static-99 by correctional staff as part of the SVP screening process, using offenders’ official criminal and correctional records. All offenders who may qualify for SVP commitment (i.e., have been convicted of two or more contact sexual offenses) are screened for SVP commitment. These screening administrations typically occur 18 months prior to release. For approximately 80% of the offenders, the second Static-99 score came from the RARC. During the time frame of this study, the RARC rescored the Static-99 for offenders whose most recent Static-99 score was 3 or lower, to ensure that offenders who should have been classified for high-risk level status (i.e., Static-99 of 4 or higher) had not been misclassified as moderate or low risk. Offenders who are classified as high risk are subject to community notification requirements (if convicted on or after January 1, 2000), whereas those classified as moderate or low risk are not. The RARC’s rescoring procedures mean that low-scoring offenders will be overrepresented in the Texas sample, because the RARC would not have automatically rescored offenders who had already received a score of 4 or higher. This case selection process could lead to an attenuated reliability coefficient attributable to range restriction. It could also lead to a systematic increase in scores over time, because subsequent scores may tend to regress toward the mean. The remaining scores came from IPO evaluations. IPO officers score the Static-99 for sex offenders who are eligible for parole, which usually occurs prior to SVP screening and RARC evaluations.

All Static-99 scores were assigned by TDCJ staff. We do not have information about the education and training of specific staff members who provided the Static-99 scores used in this study. Staff members who score the Static-99 come from many different types of

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

educational and professional backgrounds. Although some are mental health professionals, others are administrative staff or parole officers. All have bachelor's degrees, and some have master's degrees in psychology or criminal justice. The amount of Static-99 training and experience among staff also varies and has changed over time. Prior to 2002, training was conducted "in house" by TDCJ staff and emphasized reading and learning the scoring instructions provided in the instrument manual. In 2002, several TDCJ staff attended a 2-day workshop by David Thornton, and these staff became in-house trainers. Since that time, all staff who score the Static-99 received training from the in-house trainers. Historically, some staff members associated with the SVP screening and RARC process have scored as many as 300 Static-99s per month, although most IPO officers score approximately 1 to 15 per month.

For data analysis, we categorized scores into initial and subsequent scores according to administration date. We used this procedure as opposed to evaluation type (e.g., SVP screening vs. RARC vs. IPO) because we were not always able to determine the exact source of each score for an offender, mostly because of changes in evaluators over time and variations in how individual evaluators or committees identified themselves on the Static-99 scoring sheets. Moreover, categorizing scores as initial and subsequent allowed us to consider whether disagreements in scoring might be attributable to a systematic increase or decrease in scores over time.

*New Jersey.* In New Jersey, multiple sets of Static-99 scores were available for 361 of 3,168 offenders who were included in a large-scale sex offender recidivism study (see Mercado, 2010). Offenders with multiple scores were those who had been evaluated for SVP civil commitment between 2000 and 2007. In New Jersey, offenders who may be eligible for civil commitment are scored on the Static-99 by a doctoral-level psychologist approximately 6 months prior to release. Approximately 15% of those screened with the Static-99 are referred for an SVP evaluation. Because Static-99 scores are administered to assist in the SVP evaluation referral process, it is likely that high-scoring offenders will be overrepresented in the New Jersey sample, which could lead to attenuated reliability attributable to range restriction and a systematic decrease in scores over time attributable to regression to the mean.

Offenders referred for an SVP evaluation are evaluated by two medical doctors, usually psychiatrists. These two evaluations usually occur within days of one another. The psychiatrists rescore the Static-99 for some offenders but not others. When they do not rescore the Static-99, they often refer to the initial screening evaluator's score in their evaluation report. Of the 361 offenders with multiple Static-99 scores, 135 had two sets of item scores. Thus, for 226 offenders, it seemed possible that the second evaluator had not rescored the measure but had accepted the previous evaluator's score. Of course, this also could have happened for offenders with two sets of item scores, if the second evaluator accepted (and transcribed) both the item and total scores, but it seemed especially likely when an evaluation listed only a total score. Although we cannot know the extent to which subsequent scores may have been transcribed from earlier evaluations, we focus on results from the sample of 135 offenders with two sets of item scores because these seem less likely to have been transcribed.

With regard to race, nearly half (48.1%, $n = 65$) of the New Jersey offenders were identified as Black. The remaining offenders were identified as White (31.9%, $n = 143$), Latino (17.8%, $n = 24$), or Other (2.2%, $n = 3$). We used age at release as an approximation for age

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

**TABLE 1: Descriptive and Rater Agreement Statistics for Static-99 Total Scores**

| *Sample/Evaluator* | M | SD | *Standard Error of Measurement* | *95% CI* | $ICC_{A,1}$ |
|---|---|---|---|---|---|
| Texas (*N* = 600) | | | | | .79 |
| Evaluation 1 | 2.16 | 1.51 | .69 | ±1.35 | |
| Evaluation 2 | 2.25 | 1.66 | .76 | ±1.49 | |
| New Jersey: Item scores (*N* = 135) | | | | | .88 |
| Evaluation 1 | 3.87 | 2.30 | .80 | ±1.57 | |
| Evaluation 2 | 3.92 | 2.28 | .79 | ±1.55 | |

*Note.* CI = confidence interval; $ICC_{A,1}$ = absolute agreement intraclass correlation coefficient for a single rater.

**TABLE 2: Static-99 Total Score Patterns Among Pairs of Evaluators (in percentages)**

| *Agreement Between Pairs of Scores* | *Texas (*N *= 600)* | *New Jersey (*N *= 135)* |
|---|---|---|
| Both scores were identical | 55.0 | 54.1 |
| Scores differed by 1 point | 32.5 | 33.3 |
| Scores differed by 2 or more points | 12.5 | 12.6 |
| Second score higher than first | 23.2 | 25.2 |
| Second score lower than first | 21.8 | 20.7 |

at evaluation, because evaluation dates were missing from many Static-99 scoring sheets. The average age at release from the department of corrections was 40.52 (*SD* = 10.95). The data collection process for New Jersey required research assistants to record “first” and “second” Static-99 scores but did not track specific Static-99 evaluation dates. Thus, it was not possible to calculate the amount of time between administrations, although routine assessment procedures suggest that most pairs of scores were assigned within 6 months of one another.

## RESULTS

### TOTAL SCORE AGREEMENT

Rater agreement for Static-99 total scores was remarkably similar in both the Texas and New Jersey samples (see Tables 1 and 2), despite differences in the types of offenders who tended to receive a second evaluation (i.e., lower risk in Texas, higher risk in New Jersey) and the types of evaluators scoring the measure (i.e., bachelor’s- and master’s-level correctional staff in Texas, doctoral-level evaluators in New Jersey). For example, absolute-agreement single-evaluator ICC values were strong in both Texas ($ICC_{A,1}$ = .79, 95% CI [.76, .82]) and New Jersey ($ICC_{A,1}$ = .88, 95% CI [.83, .91]), although the nonoverlapping CIs indicate that agreement was significantly higher in New Jersey than in Texas.

Despite strong $ICC_{A,1}$ values, Static-99 total scores from pairs of evaluators were identical for only approximately half (55.0% in Texas, 54.1% in New Jersey) of the offenders (see Table 2). Of course, it is possible for two evaluators to come to the same total score while disagreeing about some items scores. In Texas, the two evaluators agreed in their scoring of all items for 298 (49.7%) offenders. In New Jersey, the two evaluators agreed in their scoring of all items for 58 (43.0%) offenders.

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

In most instances of total score disagreement, the second evaluator's score differed from the initial evaluator's score by only 1 point (see Table 2). Thus, in most cases, the evaluators' scores were either identical or within 1 point of each other (87.5% in Texas, 87.4% in New Jersey). However, the two evaluators disagreed by 2 or more points for approximately 1 out of every 10 offenders (12.5% in Texas, 12.6% in New Jersey).

There was no evidence of a systematic increase or decrease in scores over time (i.e., from first score to second score) in either sample. For example, the difference in mean scores between first and second evaluations was less than one tenth of a point in each state (see Table 1). In Texas, 23.2% of offenders received a higher score on their second evaluation, whereas 21.8% received a lower score on their second evaluation. Similarly, in New Jersey, 25.2% received a higher score on their second evaluation, whereas 20.7% received a lower score. There was no evidence that the length of time between evaluations was associated with a systematic increase or decrease in scores. In Texas, where we could most accurately calculate the amount of time between evaluations, the correlation between the amount of time between evaluations and the difference in Static-99 total scores (subsequent minus initial) was –.05 ($p = .60$).

## CONFIDENCE INTERVALS FOR STATIC-99 TOTAL SCORES

SEM values for Static-99 total scores ranged from .69 to .80 (see Table 1), which we calculated using the standard deviation and $\mathrm{ICC}_{\mathrm{A,1}}$ values for each evaluation (i.e., Evaluation 1, Evaluation 2) in each state: $\mathrm{SEM} = SD\sqrt{(1 - \mathrm{ICC}_{\mathrm{A,1}})}$. We then multiplied the SEM values by 1.96 to calculate 95% CIs for Static-99 total scores. The 95% CIs are ±1.35 points (Evaluation 1) and ± 1.49 points (Evaluation 2) for scores from Texas and ±1.57 (Evaluation 1) and ±1.55 (Evaluation 2) for scores from New Jersey.

## RISK LEVEL AGREEMENT

In Texas, correctional staff used the Static-99 scores in this study to determine risk level status. At that time, a score of 4 or higher was used to qualify an offender for high-risk status. We used the Texas Static-99 scores to examine the extent to which scoring disagreements would have affected risk level status. In other words, how often did the score from one evaluator indicate high risk (4 or higher) whereas the score from the other evaluator indicated low risk (3 or lower)? Scoring disagreements would have affected the risk level status of 72 (12.0%) of the offenders, whereas risk level status would have been unchanged for the remaining 528 (88.0%) offenders. The Cohen's Kappa value for this pattern of agreement was .53, indicating only fair agreement (Cicchetti, 1994) for risk level classifications.

## ITEM SCORE AGREEMENT

Table 3 provides percentage agreement and kappa values for Static-99 items scores. Kappa is a chance-corrected measure of agreement for categorical ratings, but Kappa must be interpreted in some caution in this context. Kappa assumes independence between raters (i.e., that the second rater did not know the first rater's findings), which cannot be assumed in this study. Kappa values can also yield paradoxically low underestimates of reliability when a very high level of agreement is expected by chance, such as for very common or very uncommon characteristics (see Packard & Levenson, 2006). We nevertheless report kappa coefficients in addition to percentage agreement to facilitate comparisons with existing

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

**TABLE 3: Rater Agreement for Static-99 Items**

| | *Texas (*N *= 600)* | | *New Jersey (*N *= 135)*[a] | |
|---|---|---|---|---|
| *Static-99 Item* | *% agree* | *Kappa* | *% agree* | *Kappa* |
| 1. Young | 99.3 | .91 | 95.6 | .70 |
| 2. Ever lived with lover | 91.8 | .79 | 81.8 | .63 |
| 3. Index nonsex violence | 98.2 | .64 | 89.5 | .68 |
| 4. Prior nonsex violence | 92.7 | .72 | 88.1 | .75 |
| 5. Prior sex offenses | 81.7 | .63 | 78.4 | .65 |
| 6. Prior sentencing dates | 86.2 | .62 | 84.3 | .69 |
| 7. Noncontact sex convictions | 94.0 | .53 | 93.9 | .68 |
| 8. Unrelated victims | 89.8 | .77 | 94.1 | .68 |
| 9. Stranger victims | 94.7 | .77 | 89.5 | .78 |
| 10. Male victims | 97.7 | .92 | 99.3 | .97 |

a. *N* ranges from 133 to 135 for items in the New Jersey sample because of missing data.

studies that report kappa values for Static-99 items (e.g., Hanson, 2001). Cicchetti (1994) suggests that kappa values greater than .75 be interpreted as indicating excellent agreement, those between .60 and .74 as indicating good agreement, those between .40 and .59 as indicating fair agreement, and those lower than .40 as indicating poor agreement.

Focusing on the agreement values in Table 3, agreement was above 78% for all items across both states and close to or above 90% for many items. Kappa values were in the good-to-excellent range, with the exception of the noncontact sex convictions item in Texas (kappa = .53). However, the percentage agreement value for this item was 94.0% for this relatively low base rate item (7.3% base rate for first evaluation, 6.3% base rate for second evaluation).

The two items that had somewhat lower levels of agreement in both states were prior sex offenses (81.7% and 84.3%) and prior sentencing dates (86.2% and 84.3%), which are the only two Static-99 items that are not scored using a present-versus-absent categorization. Each of these items is scored on the basis of a count of incidents and then categorized using cut scores. However, kappa values for both of these items were in the *good* range in both states and were not markedly lower than those for other items.

## DISCUSSION

At first glance, it appears that two conscientious raters should rarely disagree about an offender's Static-99 total score. After all, Static-99 item scores are based on counts of offenses and the presence or absence of offender and victim characteristics. The information needed to score these items is available in official records available to both raters, and the items appear to require little, if any, subjective judgment to score. Scoring disagreements should be rare and are probably attributable to raters' miscounting or overlooking information, unless one rater has different records than the other. The rater agreement coefficients reported in the Static-99 manual (e.g., Harris et al., 2003) and reviews of empirical studies (Hanson & Morton-Bourgon, 2009) all suggest that strong rater agreement on the Static-99 is commonplace. Nearly all studies report rater agreement coefficients (e.g., ICC, *r*) of at least .85 for the Static-99 total score, indicating excellent agreement according to

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

commonly used guidelines (see Cicchetti, 1994). In the current study, we examined rater agreement for Static-99 scores from field settings in two states and also found rater agreement coefficients in the *excellent* range ($\text{ICC}_{A,1}$ = .79 and .89).

The ICC values in our samples were strong, even in circumstances in which we might expect lower agreement. Raters in our samples came from different educational and professional backgrounds and likely had varying levels of formal Static-99 training and scoring experience. In addition, the offenders who were evaluated twice were not a randomly selected subset of offenders, and they may have been among the most difficult to score—especially in New Jersey, where rescored offenders likely had more extensive offending histories. Rater agreement coefficients for the Static-99 have been relatively low in some other studies when evaluators appear to have differed in Static-99 training and experience, even when the second rater, as in our samples, likely had access to scores from prior raters (Boccaccini et al., 2009; Murrie et al., 2009; cf. Levenson, 2004). Moreover, the nonrandom rescoring practices in each state meant that rater agreement coefficients would likely be attenuated because of range restriction. Given these barriers to strong agreement, the .79 and .88 $\text{ICC}_{A,1}$ values for Static-99 total scores from the Texas and New Jersey field settings reinforce the status of the Static-99 as a measure that appears quite reliable across raters.

Given these strong ICC values, it may seem surprising that Static-99 total scores from evaluators rating the same offender were identical for only approximately 55% of offenders in each state. In nearly half of the cases, evaluators—who were highly reliable according to common reliability metrics—assigned different Static-99 total scores. In other words, *excellent* ICC values for the Static-99 cannot be interpreted to mean that two evaluators will necessarily assign the same score to the same offender.

### TEXAS AND NEW JERSEY FIELD RELIABILITY FINDINGS IN THE CONTEXT OF PRIOR STATIC-99 RESEARCH

Many evaluators and researchers may be surprised that raters assigned the same Static-99 total score for only approximately 55% of offenders and may question the extent to which these findings are generalizable to other contexts, especially considering the inherent limitations of using field scores for reliability research. There are, nevertheless, several reasons to suspect that our agreement findings are not unique to this study. First, rater agreement findings were remarkably consistent across the Texas and New Jersey samples, even though scores were calculated by raters with different levels of education (i.e., BA and MA level in Texas, doctoral level in New Jersey) and for different types of offenders (e.g., low scorers in Texas, high scorers in New Jersey). Second, findings from both states were similar to those from two unpublished reliability studies, each reporting that evaluators assign the same Static-99 total score in approximately 40% to 45% of cases (Austin et al., 2003; Hanson, 2001), even when rater agreement coefficients are strong (e.g., ICC = .87; Hanson, 2001). Finally, there was no evidence of a systematic increase or decrease in scores over time, suggesting that the overall pattern of findings was not attributable to appropriate score increases (e.g., offenses while in custody, new information about prior offenses). Thus, data from both samples in this study (i.e., Texas and New Jersey), like data from prior research in California (Hanson, 2001), all suggest that two evaluators assign different scores to the same offender in roughly half the cases they score.

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

## IMPLICATIONS FOR POLICY AND PRACTICE

Findings from this study, like those from unpublished field reliability studies (i.e., Austin et al., 2003; Hanson, 2001), suggest that evaluators and systems should consider procedures for describing or accounting for measurement error in Static-99 total scores. Static-99 users report scores "as is," and two similar scores may prompt very different courses of action (e.g., referral for civil commitment versus release to the community). So how might Static-99 users—whether individual evaluators, institutions, or criminal justice systems—thoughtfully use Static-99 scores, in light of the finding that many of those scores could be slightly different had another rater assigned them? In this section, we identify several possible procedures. The procedures we propose are illustrative, not exhaustive. That is, we describe examples of the types of procedures that Static-99 users may want to employ, given findings that raters in the field often assign slightly different scores. No procedure is perfect, and the exact application of these procedures will probably be best informed by local reliability studies that provide detailed information about the nature and frequency of total score disagreements in the relevant setting.

*Enhance reliability to the extent possible.* One possible reaction of evaluators and systems to field reliability findings like ours is to "fix" the problem of score disagreements, perhaps through increased training. Although systems should take all possible steps to formally train raters and encourage a high level of reliability, it seems unlikely that training can improve reliability to the extent that procedures to address measurement error become unnecessary. Many scoring disagreements on the Static-99 appear to be attributable to evaluators' missing or overlooking information in files (Quesada, Mercado, & Jeglic, 2009). Although training may help raters recognize information needed for scoring, training cannot eliminate carelessness in scoring. Moreover, approximately one quarter of scoring disagreements appear to be attributable to situations in which the appropriate score is not obvious and raters must use their best judgment in assigning a score (e.g., when file has some, but limited, information related to an item; see Quesada et al., 2009). Training cannot prevent all of these disagreements.

*Report CIs for Static-99 total scores.* Ultimately, findings from Texas, New Jersey, California, and Pennsylvania all indicate that field reliability for the Static-99 is very good. But it is not *so* good that we can neglect to acknowledge measurement error. One way to incorporate information about measurement error into test score interpretation is to report CIs. In the current study, 95% CIs for Static-99 total scores ranged from 1.35 to ±1.57 points. Using the ±1.35 value for initial Texas evaluations as an example, the rater would obtain the lower bound of the 95% CI for an assigned score by subtracting 1.35 points from the score and an upper bound by adding 1.35 to the score. For example, the 95% CI for a score of 3 would be 1.65 to 4.35. The 95% CI for a score of 5 would be 3.65 to 6.35.

It is important to conduct local reliability studies to determine the size of CIs because their size is influenced by the standard deviation of scores and rater agreement coefficient, both of which can vary from setting to setting. Nevertheless, it seems likely that 95% CIs for Static-99 total scores will be at least ±1.0 point in most contexts. The standard deviation of Static-99 total scores is close to 2.0 points in most studies (see Helmus, 2009, Table 4), so even with a rater agreement coefficient of .95 and a standard deviation of 2.0, the 95%

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

CI for the Static-99 would be ±0.88 points. And the 95% CI would be greater than ±1.0 points when the rater agreement coefficient dipped below .93, as it has in many studies.

These types of observed score CIs are especially important for systems that use cut scores to assist in decision making. For example, if Texas were still using a cut score of 4 to make decisions about risk level status, this means that offenders with scores of 3 or 5 may have true scores that include the cut score of 4. We would recommend rescoring the Static-99 for these offenders. However, offenders with scores of 2 (95% CI = 0.65 to 3.35) and 6 (95% CI = 4.65 to 7.35) do not include 4, and rescoring would not be required.

*Use multiple raters for each offender.* One way to improve the reliability of an instrument score is to base the score on ratings from multiple raters. In other words, two or more raters score the same offender to enhance reliability. The more raters, the more reliable the score, the smaller the standard error of measurement, and the smaller the confidence interval. For example, using the New Jersey sample from this study, we can use the Spearman-Brown prophecy formula to calculate a rater agreement coefficient of .94 for scores if they are averaged across two raters. The 95% CI for Evaluation 1 scores would narrow from ±1.57 points to ±1.10 points if offender scores were based on the average of two raters' scores.

It is important to note that there is little to no social science evidence to suggest that having evaluators discuss and come to a consensus about an offender's score offers any benefit over and above simply averaging evaluators' scores. Having a group come to a consensus about an offender's score may be especially useful for identifying more obvious scoring errors, such as miscalculation and overlooking clearly relevant information. However, group judgments are usually, at best, only as accurate as those based on the average judgments from individuals, and when groups do make errors, they tend to be of larger magnitude than errors from individuals (see Gigone & Hastie, 1997). Furthermore, requiring evaluators to come to a consensus score can be much more costly than simply averaging scores, given the time raters must spend discussing cases to come to a consensus.

*Use multiple raters for offenders near cut scores.* Of course, it may seem cost-prohibitive to rescore every offender in every context, particularly in contexts where the Static-99 is administered on a widescale basis (e.g., scoring every sexual offender in a state department of corrections). When this is the case, it seems appropriate to focus rescoring resources on those for whom a relatively small score difference may have a large impact. For example, in states, such as Virginia, where Static-99 scores influence which offenders are pursued for civil commitment (a process that is costly in terms of state resources and offender civil liberties), it may be especially important to rescore offenders who score 1 or 2 points above or below the cut score used to make SVP referral decisions. Likewise, when Static-99 scores determine whether an offender qualifies for community notification (e.g., Texas), it may make sense to rescore offenders whose confidence interval includes the cut score.

*Considerations for recidivism rate interpretations.* Current norms for the Static-99 (and Static-99R) provide confidence intervals for recidivism rates. For example, the estimated 5-year recidivism rate for sexual recidivism among routine sample offenders is 7.7%, with a 95% CI of 5.9% to 10.0%. These confidence intervals around the recidivism rate estimates are not the same as the CIs around the observed score, which are based on the SEM.

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

Rather, the CIs around the Static-99 recidivism rates are based on findings from logistic regression analyses examining the relation between Static-99 scores and recidivism. They capture the extent to which predictions based on the regression equation tend to match up with the observed data used to derive the equation. The better the observed data match up with the prediction equation, the smaller the CI.

Measurement error is one reason why data for many offenders do not match perfectly with expectations based on the regression equation. The size of the recidivism rate CIs should increase as measurement error increases and decrease as measurement error decreases. The important question that evaluators must consider is whether the normative recidivism rate CIs adequately account for the amount of measurement error in their field settings. Ideally, evaluators could consider the extent to which rater agreement in their setting matches rater agreement for scores in the normative sample. Unfortunately, rater agreement values are not available for all of the samples that were combined to create the recidivism rate norms and CIs, likely because the norms are based on findings collapsed across multiple studies and rater agreement was not examined in all of those studies. Moreover, most evaluators have no way to estimate rater agreement for their own field scores. What should evaluators do in the absence of this information? If they chose to report recidivism estimates, they must also report the CIs for the recidivism rate estimates because these do account for the measurement error that existed in the normative sample. However, evaluators should recognize that those CIs may or may not adequately account for measurement error in their field settings. The risk estimate confidence intervals may be too narrow if agreement in the field setting is especially poor or too wide if agreement in the field setting is excellent. Of course, there are a number of other factors that could influence the extent to which the normative recidivism rate estimates apply to a field setting, such as the overall predictive effect of the Static-99 and the base rate of recidivism, but evaluators should recognize that measurement error is one of these factors.

## LIMITATIONS

Again, we emphasize that using field data involves accepting some study limitations that we would not accept if we were using research staff to generate data solely for research purposes. For example, some second raters may not have been blind to scores from the first rater, cases were not randomly selected for rescoring, and there were typically several months that elapsed between the two evaluations. Although we cannot know with certainty the effect that the these field study characteristics had on agreement in the Texas and New Jersey samples, this study does not appear to be an unfair test of the Static-99. Our agreement findings fall squarely within the range of those reported in the published and unpublished Static-99 research literature. The possibility that subsequent raters were not blind to earlier scores is perhaps the biggest limitation of the study, because it might inflate agreement. Nevertheless, the Texas and New Jersey ICC values were lower than those obtained in other studies with this same limitation (i.e., ICC = .91; Hanson et al., 2007).

At times, our findings suggest somewhat stronger agreement than other studies, such as total score agreement being stronger in Texas and New Jersey (approximately 55%) than in other reports (40% to 45% in Austin et al., 2003; Hanson, 2001). At other times, our findings suggest somewhat weaker agreement, such as the $\mathrm{ICC}_{A,1}$ value of .79 in Texas

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

compared to the median rater agreement coefficient of .90 reported by Hanson and Morton-Bourgon (2009). Of course, our findings may not generalize to other field settings, and we encourage all systems that rely on the Static-99 to conduct their own local reliability studies.

Finally, although most findings for the Static-99 clearly are related to the Static-99R, the one item scoring change (young age) has arguably made the Static-99R measure somewhat more difficult to score. The revised age item has four, as opposed to two, scoring options and includes two scoring options with negative values. In both Texas and New Jersey, agreement was lower for items that contained more than two scoring options, and the use of negative values could lead to an increase in calculation errors. Moreover, the age used to score this item varies depending on whether the offender's current offense is or is not a sexual offense. Ultimately, these issues may lead to somewhat lower agreement for the Static-99R than the Static-99. If this were the case, we would expect to find somewhat lower ICC values for total scores, which would increase the size of CIs and lower rates of total score agreement.

## CONCLUSIONS

The purpose of this study was to examine how often pairs of evaluators in field settings, who both assign scores for the purpose of clinical or correctional use (as opposed to research use), report the same score. Across two state samples, raters assigned the same total score for approximately 55% of offenders, even though ICC rater agreement coefficients were excellent. Thus, although pairs of scores were rarely more than 1 point apart, as the 2003 Static-99 scoring manual informs users (Harris et al., 2003), they were often *at least* 1 point apart. This finding raises important questions about how evaluators and policies should acknowledge and account for measurement error in Static-99 scores. We have offered several possible options, including reporting confidence intervals and rescoring cases near cut scores. Static-99 users should think carefully about these options and how each may best address the types of errors that could have a major impact on the offenders they evaluate. Ultimately, evaluators and agencies must make decisions on how to best address score disagreements by conducting local reliability studies.

## NOTES

1. Recidivism rate examples are based on Routine Correctional Services of Canada sample norms. There are several different sets of Static-99 norms, based on different samples of offenders. Although we use the routine samples norms here, the same principle applies to each set of norms: Each score is associated with a different recidivism rate.

2. In rare circumstances, a new offense can lower a Static-99 score (see Phenix, Doren, Helmus, Hanson, & Thornton, 2008, p. 3). For example, if an initial score includes a point for an index offense that also involved nonsexual violence, but an offender commits a new index offense that does not include a point for nonsexual violence (and this additional offense did not increase his score on the "prior offenses" item), his new offense could actually result in a slightly lower Static-99 score.

## REFERENCES

Atkins v. Virginia, 536 U.S. 304 (2002).

Austin, J., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the Static-99/RRASOR sex offender risk assessment instruments.* Retrieved from National Institute of Corrections website: http://nicic.gov/Library/022957

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

Bengtson, S., & Långström, N. (2007). Unguided clinical and actuarial assessment of re-offending risk: A direct comparison with sex offenders in Denmark. *Sex Abuse*, *19*, 135-153. doi:10.1007/s11194-007-9044-5

Boccaccini, M. T., Murrie, D. C., Caperton, J., & Hawes, S. (2009). Field validity of the STATIC-99, and MnSOST-R among sex offenders evaluated for commitment as sexually violent predators*. Psychology, Public Policy, and Law*, *15*, 278-314. doi:10.1037/a0017232

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

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*, 149-167.

Hanson, R. K. (2001). *Note on the reliability of Static-99 as used by California DMH evaluators.* Unpublished report, California Department of Mental Health, Sacramento, CA.

Hanson, R. K., Harris, A. J. R., Scott, T. L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (Corrections Research User Report No. 2007-05). Ottawa, ON: Public Safety Canada.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis. *Psychological Assessment*, *21*, 1-21. doi:10.1037/a0014421

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119-136.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Coding rules for the Static-99 Revised-2003. Corrections research: Manuals and forms.* Ottawa, ON: Department of the Solicitor General of Canada.

Helmus, L. (2008). *Static-99 replications: Descriptive information.* Retrieved from Static-99 website: http://www.static99.org/pdfdocs/static-99replicationsdescriptives.pdf

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* (Master's thesis). Carleton University, Ottawa, Ontario, Canada. Retrieved from http://www.static99.org/pdfdocs/helmus2009-09static-99normsmathesis.pdf

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (in press). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment.* doi:10.1177/1079063211409951

Interstate Commission for Adult Offender Supervision. (2007). *SO Assessment Information Survey 4.* Retrieved from http://www.interstatecompact.org/

Jackson, R. L., & Hess, D. T. (2007). Evaluation of civil commitment of sex offenders: A survey of experts. *Sexual Abuse: A Journal of Research and Treatment*, *19*, 425-448. doi:10.1177/107906320701900407

Levenson, J. S. (2004). Sexual predator civil commitment: A comparison of selected and released offenders. *International Journal of Offender Therapy and Comparative Criminology*, *48*, 638-648. doi:10.1177/0306624X04265089

Levenson, J. S., & Morin, J. W. (2008). Factors predicting selection of sexually violent predators for civil commitment. *International Journal of Offender Therapy and Comparative Criminology*, *50*, 609-629. doi:10.1177/0306624X06287644

Mercado, C. C. (2010, October). *An examination of treatment, SVP commitment, and recidivism in a statewide sample of sex offenders.* Symposium presented at the Annual Convention of the Association for the Treatment of Sexual Abusers, Phoenix, AZ.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Packard, R. L., & Levenson, J. S. (2006). Revisiting the reliability of diagnostic decisions in sex offender civil commitment. *Sexual Offender Treatment*, *1*(3). Retrieved from http://www.sexual-offender-treatment.org/50.html

Phenix, A., Doren, D., Helmus, L., Hanson, R. K., & Thornton, D. (2008). *Coding rules for the Static-2002.* Retrieved from http://www.static99.org/pdfdocs/static2002codingrules.pdf

Quesada, S. P., Mercado, C. C., & Jeglic, E. (2009, March). *The reliability of the Static-99: A comparison of researcher and clinician ratings*. Poster presented at the annual meeting of the American Psychology-Law Society, San Antonio, TX.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

Wechsler, D. (2008). *WAIS-IV: Technical and interpretive manual.* San Antonio, TX: Pearson.

**Marcus T. Boccaccini** is an associate professor of psychology at Sam Houston State University. His research focuses on the agreement between forensic evaluators in the field, and the validity of their opinions and the scores they assign on clinician-scored measures.

**Daniel C. Murrie** is an associate professor of psychiatry and neurobehavioral sciences at the University of Virginia School of Medicine and director of psychology at the Institute of Law, Psychiatry, and Public Policy. His research examines forensic evaluations in the field, with the goal of improving forensic practice.

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016

**Cynthia Mercado** is an associate professor of psychology at the John Jay College of Criminal Justice in New York. Her work focuses on establishing empirical evidence for use in sex offender policy and sexual violence prevention.

**Stephen Quesada** holds a master's degree from John Jay College of Criminal Justice in forensic psychology. He currently works with the homeless population as a social worker for a nonprofit organization, NaNa's House.

**Samuel Hawes** is a doctoral candidate in the clinical psychology PhD program at Sam Houston State University. His research focuses on the predictive validity of measures used in forensic assessment.

**Amanda K. Rice** is a doctoral student in the clinical psychology PhD program at Sam Houston State University. Her research focuses on the reliability of risk assessment measures.

**Elizabeth L. Jeglic** is an associate professor of psychology at the John Jay College of Criminal Justice in New York. Her research interests include sex offender assessment and treatment and their relationship to public policy.

Downloaded from cjb.sagepub.com at JOHN JAY COLLEGE on October 28, 2016