Published in final edited form as:

Psychol Assess. 2014 December ; 26(4): 1085–1094. doi:10.1037/pas0000009.

Does Field Reliability for Static-99 Scores Decrease as Scores Increase?

Amanda K. Rice,

Sam Houston State University

Marcus T. Boccaccini,

Sam Houston State University

Paige B. Harris, and

Sam Houston State University

Samuel W. Hawes

University of Pittsburgh

Abstract

This study examined the field reliability of Static-99 (Hanson & Thornton, 2000) scores among 21,983 sex offenders and focused on whether rater agreement decreased as scores increased. As expected, agreement was lowest for high-scoring offenders. Initial and most recent Static-99 scores were identical for only about 40% of offenders who had been assigned a score of 6 during their initial evaluations, but for more than 60% of offenders who had been assigned a score of 2 or lower. In addition, the size of the difference between scores increased as scores increased, with pairs of scores differing by 2 or more points for about 30% of offenders scoring in the high-risk range. Because evaluators and systems use high Static-99 scores to identify sexual offenders who may require intensive supervision or even postrelease civil commitment, it is important to recognize that there may be more measurement error for high scores than low scores and to consider adopting procedures for minimizing or accounting for measurement error.

Keywords

Static-99; risk assessment; sex offender; field reliability; rater agreement

---

The Static-99 (Hanson & Thornton, 2000) is the most commonly used and researched sex offender risk assessment instrument. Evaluators score the 10-item Static-99 by reviewing an offender's criminal records and assigning scores to items that consider offender and offense characteristics, including age at release, living with a significant other for 2 or more years, index nonsexual violence convictions, prior nonsexual violence convictions, number of prior sex offenses, number of prior sentencing dates, noncontact sexual offense convictions,

---

© 2014 American Psychological Association

Correspondence concerning this article should be addressed to: Marcus T. Boccaccini, Department of Psychology, Box 2447, Sam Houston State University, Huntsville, TX 77341. Boccaccini@shsu.edu.

Amanda K. Rice, Marcus T. Boccaccini, and Paige B. Harris, Department of Psychology, Sam Houston State University; Samuel W. Hawes, Department of Psychiatry, University of Pittsburgh.

sexual offending against unrelated victims, sexual offending against strangers, and sexual offending against males. Evaluators score nine of the 10 items as *yes* (present) or *no* (absent) but assign a score between 0 and 3 for the prior sex offenses item (Item 5). As a result, Static-99 total scores can range from 0 to 12. The Static-99 authors recently released a revised version of the measure—the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012). The two measures contain the same 10 items and identical scoring rules for nine of the 10 items, with the only scoring rule change relating to scores assigned based on the offender's age at release (Item 1).

Each Static-99 total score corresponds with an associated risk of recidivism, which can be communicated using several different methods, including a recidivism risk estimate, risk ratio, and categorical risk grouping (e.g., high, low). Each possible Static-99 total score has a different interpretation. For example, an evaluator referring to normative sample recidivism rates would note that a score of 4 is associated with a 5-year sexual recidivism rate of 7.7%,¹ while a score of 5 is associated with a rate of 10.2%, a score of 6 is associated with a 13.4% rate, and so on. A higher score will always be associated with a higher recidivism rate, regardless of whether evaluators compare offenders to the high risk, routine risk, or any other set of norms described in the Static-99 manual. Likewise, relative risk ratios for Static-99 scores are 1.89 for a score of 4, 2.42 for a score of 5, 2.96 for a score of 6, and so on (see Helmus, Hanson, & Thornton, 2009).

Even 1-point score differences on the Static-99 or Static-99R can have important implications for offenders. For example, Texas sexual offenders with two or more sex offense convictions who receive a score of 6 or higher on the Static-99 are automatically required to receive GPS monitoring (Texas Department of Criminal Justice Parole Division, 2011). For many years, Virginia statutes dictated that only offenders scoring 4 and above could be considered for postrelease civil commitment, while those scoring 3 and below could not (Joint Legislative Audit and Review Commission, 2012).

Most studies report relatively high levels of rater agreement for Static-99 total scores, with one review reporting a median rater-agreement coefficient of .90 across 12 studies (Hanson & Morton-Bourgon, 2009). These agreement coefficients do not, however, necessarily mean that evaluators usually agree about the exact value of the offender's score, especially in field settings. For example, researchers have reported good to excellent intraclass correlation coefficient (ICC) values for Static-99 scores assigned by field evaluators in New Jersey (ICC = .88), California (ICC = .87), and Texas (ICC = .79) but have found that pairs of evaluators assigned identical scores to anywhere between 45% and 55% of offenders (Boccaccini et al., 2012; Hanson, 2001).

One factor that could help explain the percentage-agreement findings in these field studies is that each examined scores for a relatively select group of presumably high-risk and relatively high-scoring offenders being considered for civil commitment as sexually violent predators (SVPs). Indeed, the mean Static-99 score was 4.9 in the California sample, which

---

¹Recidivism rate examples are based on Routine Correctional Services of Canada (CSC) sample norms (Static-99 Clearinghouse, 2008).

also had the lowest total score agreement rate (45.5%). Total score agreement was better in the Texas sample (55.0%), which had a mean Static-99 score of only 2.2 (see Boccaccini et al., 2012; Hanson, 2001). There are several reasons why rater agreement might be lowest for those with relatively high Static-99 scores. First, offenders who receive higher scores often have many charges and convictions, and, thus, more file data to review than lower risk offenders. Evaluators may be more likely to overlook information or miscount offenses when records are lengthy or when there are a large number of offenses that need to be reviewed. Files for these offenders may be more likely than those from other offenders to contain inconsistencies. Indeed, one group of researchers who reported relatively low rater agreement for Static-99 scores ($ICC = .63$) attributed much of the disagreement to the inconsistency of records (Ducro & Pham, 2006).

Another reason to expect more disagreement for higher scores is that scores tend to regress toward the population mean upon rescore, an effect that becomes more pronounced as scores get further away from the mean (Campbell & Kenny, 1999). It is for this reason that confidence intervals for very high and very low intelligence test scores are larger toward the mean than away from it. For example, the 95% confidence interval for a score of 130 on the Wechsler Adult Intelligence Scale–IV (Wechsler, 2008) full-scale IQ score is [125, 133]—5 points toward the mean but only 3 points away from the mean. For a score at the mean (100), the 95% confidence interval (CI) is symmetrical [96, 104]. The mean Static-99 score is about 3.1 points (Helmus, 2009). Because scores on the Static-99 range from 0 to 12, there are more score options above the mean (4–12) than below it (0–2). Thus, there are more and larger extreme score options above the mean than below, and the extreme scores can be more extreme above the mean than below. The lowest Static-99 score can be only 3 points below the mean, but the highest score can be nearly 9 points above the mean (although scores of 12 probably never occur).

The current study examined whether rater agreement for Static-99 scores decreased as scores increased among 21,983 incarcerated offenders scored on the Static-99 on multiple occasions by correctional staff. We expected that rater agreement would be stronger for offenders with lower Static-99 total scores than higher scores. We expected that this pattern of more disagreement for high scores would be most evident for Static-99 items that require evaluators to count convictions, charges, or sentencing occasions, rather than record the presence or absence of an event or characteristic.

The Static-99 scores we examine in this study were used to make real-life decisions about offenders, including referrals for sexually violent predator evaluations and postrelease monitoring requirements. A finding that Static-99 rater agreement decreases as scores increase would have important implications for evaluators, courts, and criminal justice agencies that use these scores to make decisions about offenders and suggest that they need to be especially mindful of measurement error for offenders with high scores. Because scoring procedures are nearly identical for the Static-99 and Static-99R, these findings would have similar implications for evaluators and systems that have adopted the Static-99R.

Method

Offenders and Static-99 Scores

We obtained Static-99 total scores from an electronic Texas Department of Criminal Justice (TDCJ) database that contained total scores for all male offenders ($N = 55,146$) scored on the Static-99 between 1999 and 2011, 23,684 (42.9%) of whom had been scored on at least two occasions during the same period of incarceration. Offenders had been scored on the Static-99 for purposes of risk-level determination, prerelease evaluation, parole evaluations, and civil commitment screening. The database included a name, score, evaluator name, and evaluation date for each Static-99 administration. Because an offender's Static-99 score can change after he is released (e.g., commits a new offense), it was important to focus on scores that were assigned during the same period of incarceration (i.e., offender had not been released). Because there can be a valid change in an offender's score if he passes his 25th birthday or was charged with a new offense while incarcerated, we excluded offenders who passed their 25th birthday ($n = 1,192$) or were charged with a new offense ($n = 509$) between their earliest and most recent Static-99 scores. We obtained information about new charges from the Texas Department of Public Safety. Although these data allowed us to identify offenders with new arrests, they did not provide the detailed information needed to score Static-99 items.

We focused our rater-agreement analyses on the remaining sample of 21,983 (39.1%) offenders who had been scored on at least two occasions during the same incarceration. This sample included offenders who had been scored on the Static-99 multiple times and had never been released ($n = 6,552$), and those who had been scored multiple times over their incarceration history but had multiple scores for only one incarceration ($n = 14,621$). For example, an offender may have been scored a total of five times on the Static-99 across two incarcerations but scored only once during his first incarceration and four times during his second. We also included data from offenders who had multiple Static-99 scores for multiple incarcerations ($n = 810$). For these offenders, we randomly selected data from only one incarceration so that each offender would be represented only once in our analyses. The average amount of time between the earliest and most recent evaluations was 37.70 months ($SD = 30.34$).

There is no reason to suspect that those who were scored multiple times were selected for rescoring due to concerns about score accuracy. Rather, offenders are rescored because they are being evaluated for different programs (e.g., parole, SVP screening) or are having their scores checked before release (prerelease evaluation). Nevertheless, we compared those included ($n = 21,983$) and excluded ($n = 33,163$) from the study to gauge the representativeness of the offenders in our agreement analyses. Offenders in the study sample had somewhat higher initial Static-99 total scores ($M = 2.67$, $SD = 1.74$) than those who were not included ($M = 2.48$, $SD = 1.63$). Although this difference was large enough to reach statistical significance ($p < .001$) in this sample of more than 55,000 offenders, the difference was small in terms of effect size (Cohen's $d = 0.11$). Offenders in the study sample were also somewhat older at the time of their initial evaluation ($M = 40.72$ years, $SD = 10.97$) than offenders not included ($M = 38.04$, $SD = 12.14$; $d = 0.23$, $p < .001$). This effect

is a product of excluding the offenders who passed their 25th birthday between their initial and most recent evaluations. The two samples were similar in terms of offender race or ethnicity: White (49.5% vs. 47.9%), African American (29.8% vs. 29.0%), and Hispanic (20.3% vs. 22.6%).

We obtained Static-99 item scores for 1,594 of the 21,983 (7.3%) offenders from a partially overlapping dataset of scores assigned to offenders who had been screened for SVP civil commitment between 1999 and 2008; of these, 308 were included in a prior Static-99 rater-agreement study (Boccaccini et al., 2012). This dataset included only the two most recent Static-99 scores for each offender. Mean age at the time of the first evaluation was 42.83 ($SD = 11.57$), and the mean Static-99 total score at the time of the first evaluation was 2.41 ($SD = 1.70$).

Evaluators

TDCJ staff, including mental health professionals, parole officers, and administrative staff, assigned Static-99 scores. There were more than 600 different evaluators who scored the Static-99 for at least one offender. These evaluators vary in education, training, and profession, with all evaluators having a bachelor's degree and some having a master's degree in psychology or criminal justice. Current TDCJ staff report that most evaluators are trained to score the Static-99 in-house by TDCJ staff (see Boccaccini et al., 2012), but we do not have any information about the training or background of individual evaluators. We do, however, know that only about 3% of the variance in Static-99 scores in this sample is attributable to differences among evaluators (Rice, Boccaccini, & Collier, 2012). In other words, there is little evidence that some evaluators assign markedly different Static-99 scores than others. Evaluators were not necessarily blind to the Static-99 scores assigned by prior evaluators. Although this might lead to inflated rater agreement, it is an unavoidable aspect of most field-reliability research.

Results

There were more than 20,000 offenders in most of our analyses. Because even very small effects are large enough to reach statistical significance in this large sample, we emphasize effect size over statistical significance when describing findings.

Static-99 Total Scores

TDCJ staff scored offenders on the Static-99 an average of 2.53 ($SD = 0.81$, range = 2–8) times during their incarceration period, with 8,298 (37.7%) offenders scored on three or more occasions. We used two different analytic approaches for examining rater agreement. First, we compared each offender's initial (i.e., earliest) Static-99 score to his most recent score for the same incarceration. The utility of this two-score approach is that it allowed us to classify each case as indicating agreement or disagreement while using the same number of evaluations (i.e., two) for each offender. This approach also provides information about the reliability of the first score assigned to an offender, which is the only score that many offenders receive. Indeed, 57.1% of the 55,146 offenders scored by the Texas Department of Criminal Justice during the timeframe of this study had been scored on only one occasion

(i.e., had only an initial score). Findings from the two-score approach provide information about the reliability of these initial scores.

The drawbacks to the two-score approach are that it does not consider all of the scores for each offender and does not consider the size of scoring disagreements. For example, the two-score approach treats a score difference of 1 point similarly to a score difference of 5 points (i.e., both coded as *disagree*). Thus, we also examined the relation between initial Static-99 score values and rater agreement by calculating a Static-99 score standard deviation separately for each offender. We used this standard deviation variable in a second set of agreement analyses. The standard deviation values are based on all scores assigned to the offender during the same incarceration. Because larger scoring disagreements will lead to larger standard deviation values, these analyses consider the extent to which the size of scoring disagreements may differ for those with different initial Static-99 scores.

Agreement between pairs of scores—The initial and most recent Static-99 scores were identical for about half (55.2%) of the 21,983 offenders. The two scores differed by 1 point for 6,815 (31.2%) offenders, 2 points for 2,161 (9.8%) offenders, and more than 2 points for 881 (4.0%) offenders. The absolute-agreement single evaluator ICC for these pairs of Static-99 total scores was strong ($ICC_{A,1} = .81$, 95% CI [.81, .82]). This ICC value provides information about the proportion of variance in Static-99 scores that is attributable to differences in the traits and behaviors measured by the Static-99, as opposed to other sources of variance (e.g., variance attributable to evaluators, random measurement error).

To examine whether rater agreement decreased as Static-99 total scores increased, we calculated a percentage-agreement value separately for each possible Static-99 initial score (see Table 1). For example, we calculated the percentage of offenders with an initial score of 1 who also were assigned a score of 1 during their most recent Static-99 assessment. The agreement values, which are listed in Table 1, indicate that agreement decreased as initial scores increased. For example, percentage agreement was 74.1%, 65.4%, and 57.8% for offenders with initial scores of 0, 1, or 2, respectively, but was 42.1%, 39.9%, 39.8%, and 36.0% for offenders with initial scores of 5, 6, 7, and 8, respectively. The one exception to the pattern was a 46.7% agreement value for the relatively small number of offenders ($n = 45$) with an initial score of 9.

Although the percentage agreement values in Table 1 show that rater agreement decreased as initial scores increased, we used logistic regression to provide information about the size and statistical significance of this effect. The dependent variable in this model was a dichotomous variable indicating whether the Static-99 initial score was identical to the most recent score (0 = identical, 1 = not identical). The predictor variable was the initial Static-99 score. A positive regression coefficient for the initial Static-99 score would indicate that disagreement became more common as the value of the initial Static-99 score increased.

There was a statistically significant positive association between initial Static-99 scores and likelihood of disagreement in the logistic regression model (see Table 2). The odds ratio (OR) of 1.26 indicates that there was a 1.26 increase in the odds of disagreement for every 1-point increase in the initial Static-99 score. The odds of disagreement are larger for

increases of more than 1 point, which one can obtain by multiplying the regression coefficient ($B = 0.230$) by the size of the point increase and then using the result as the exponent of e (base of the natural logarithm; see Wright, 1998). For example, the odds of two evaluators disagreeing about the Static-99 score for an offender with an initial score of 6 are about 3.16 times the odds of them disagreeing for an offender with an initial score of 1.

We performed a second logistic regression model that also included time between evaluations (years), initial Static-99 score, and the interaction of time between evaluations and the initial Static-99 score as predictors in the model (Static-99 scores centered prior to analysis). We used this model to examine whether the pattern of decreasing agreement with increasing scores might be attributable to the amount of time that elapsed between the two Static-99 evaluations. For example, it may be that the pattern is only evident when the two scores were assigned many months or years apart, as opposed to more closely together in time. Evidence for this type of effect would come from the interaction term. There was, however, no evidence of a statistically significant interaction (see Table 2). There was a small, but statistically significant effect for time ($B = 0.108, p < .001$). The odds ratio of 1.11 indicates that there was only a 1.11 increase in the odds of disagreement for every year that elapsed between evaluations. The effect for initial Static-99 score was still statistically significant in this model ($B = 0.214, OR = 1.24$), and only marginally smaller than in the original model ($B = 0.230, OR = 1.26$). Thus, there was still evidence of agreement decreasing as scores increased, even after controlling for the amount of time between evaluations.

We also considered whether agreement might have been affected by availability of a new Static-99 coding manual in 2003 (Harris, Phenix, & Thornton., 2003). We tested rater agreement and logistic regression analyses using only offenders with initial and most recent scores assigned between 2004 and 2011 ($n = 12,332$). Rater agreement was strong among this subsample offenders ($ICC_{A,1} = .88, 95\% CI [.87, .88]$). The initial and most recent scores were identical for 67.0% of these offenders, differed by 1 point for 24.9%, and differed by two or more points for 8.1%. Yet, even under these conditions of strong agreement, there was still evidence of decreasing agreement with increasing scores (Table 2). The odds ratio of 1.21 indicates that there was a 1.21 increase in the odds of disagreement for every 1-point increase in the initial Static-99 score.

Agreement using the score standard deviation—We obtained similar findings when we used the Static-99 score standard deviation for each offender as the outcome variable in a linear regression analysis, with the initial Static-99 score as the predictor variable (see Table 3). The mean standard deviation across offenders was 0.44 ($SD = 0.56$). Once again, there was a positive relation between Static-99 initial scores and disagreement ($\beta = 0.230, p < .001$), with the positive coefficient indicating that there was more disagreement (i.e., higher score standard deviation values) for offenders with higher as opposed to lower initial Static-99 scores.

We also examined time between evaluations (years) as a moderator of the association between initial Static-99 scores and standard deviation values (see Table 3). Once again, there was a small but statistically significant effect for time between evaluations ($\beta = 0.108$,

$p < .001$) indicating that standard deviation values increased as the amount of time between the initial and most recent score increased. The effect for the interaction term also reached statistical significance ($p < .01$), but the effect was small ($B = 0.003$, $\beta = .021$, $p = .001$). This effect indicates a tendency for the relation between initial score and standard deviation values to increase as the time between evaluations increases. The effect for the initial Static-99 score was still statistically significant in this model ($\beta = 0.216$) and only somewhat smaller than in the original model ($\beta = 0.231$). In other words, there was still evidence of agreement decreasing as scores increased, even after controlling for the amount of time between evaluations.

We also examined the relation between earliest Static-99 scores and Static-99 standard deviation values using only offenders scored between 2004 and 2011 (see Table 3). Although the effect was smaller than in the full-sample model, there was still evidence of increasing Static-99 standard deviation values with increasing Static-99 scores ($\beta = 0.160$, $p < .001$).

Differences of two or more points—One implication of our linear regression findings is that the size of score disagreements increased as the offenders' initial scores increased. This pattern is evident when we consider how common it was for offenders' initial and most recent scores to differ by two or more points (13.8%), which the Static-99 authors argue should be rare (Harris et al., 2003). The most recent Static-99 score differed from the initial score by two or more points for 7.5% of offenders with an initial score of 0 and 6.9% of offenders with an initial score of 1. The percentage of offenders with disagreements of two or more points tended to increase as initial scores increased: score of 2 (11.0%), 3 (15.1%), 4 (18.0%), 5 (22.7%), 6 (26.1%), 7 (28.2%), 8 (35.3%), 9 (28.9%), and 10 (50.0%).

Regression to the mean—Table 1 also provides information about the percentage of cases in which an offender's most recent score was higher than his initial score, and the percentage of cases in which his most recent score was lower than his initial score. We calculated these percentages to examine the extent to which the disagreement we observed might be attributable to scores increasing over time, perhaps due to new records about prior offenses becoming available. If scores were systematically increasing over time, we would expect to see that most disagreements were due to the most recent score being higher than the initial score.

The findings in Table 1 provide no evidence that scores were systematically increasing over time. Instead, they reflect a pattern consistent with regression to the mean. The mean Static-99 score in our sample was 2.67. Offenders with initial scores below the mean (0, 1, 2) tended to receive higher scores in later evaluations (i.e., toward the mean), whereas offenders with initial scores above the mean (3 and higher) tended to receive lower scores in later evaluations (i.e., toward the mean). For example, 64.5% of offenders with an initial score of 1 were assigned a higher score in their most recent evaluation, whereas only 34.9% with an initial score of 4 were assigned a higher score in their most recent evaluation.

Also consistent with a regression to the mean interpretation is the finding that the percentage of disagreement cases that regressed toward the mean was larger for cases further away from

the mean than for cases closer to the mean (see Campbell & Kenny, 1999). For example, approximately 80% of disagreements for offenders with scores of 6 were a product of recent scores being lower than initial scores, compared with only 65% of disagreements for offenders with scores of 4, which is much closer to the sample mean of 2.67.

Score Disagreement and Risk-Level Classification

In Texas, an offender who scores 6 or higher on the Static-99 qualifies for high-risk release status and is subject to more intensive management and monitoring in the community than those classified as lower risk. To examine the extent to which score disagreements could have impacted risk classification status, we coded each offender's initial and most recent scores as being low risk (scores of 0 to 5) and high risk (scores of 6 or higher). We then compared risk classification status changes for the high and lower risk groups.

Although risk classification status was unchanged for most offenders (94.9%), the pattern of effects was markedly different for those in different classification groups. Of the 1,359 offenders with an initial score of 6 or higher, 38.9% ($n = 529$) received a score of 5 or lower during their most recent evaluation. Of the 20,624 offenders with an initial score of 5 or lower, only 2.9% ($n = 594$) received a score of 6 or higher during their most recent evaluation. In other words, risk classification status may have changed for nearly 40% of those with initial scores in the high-risk category but only about 3% of those with initial scores in the low risk category. The odds ratio for this difference was 52.91 (95% CI [46.17, 60.63]), indicating a very large and statistically significant effect, $\chi^2(1, N = 21,983) = 7,127.33, p < .001$.

Static-99 Item Scores

Score agreement—Pairs of Static-99-item scores were available for 1,594 offenders. These were offenders who had been screened for civil commitment as sexually violent predators and had two or more scores on file. The mean score among these offenders was 2.41 ($SD = 1.70$), only slightly lower than the sample as a whole ($M = 2.67, SD = 1.74$). $ICC_{A,1}$ for total scores in this subsample was .83 (95% CI [.81, .85]), with identical scores assigned to 55.1% of offenders. As expected, agreement, defined as two evaluators assigning an identical score, was lowest for items that require the counting of incidents, including number of prior sex offenses (81.6%) and more than four sentencing occasions (85.9%). Agreement values for other items were as follows: age at release (98.7%), ever lived with partner (90.3%), index nonsexual violence (97.1%), prior nonsexual violence (91.9%), noncontact sex offense (96.2%), unrelated victim (89.6%), stranger victim (92.8%), and male victim (96.7%).²

Because scores on Item 5 (prior sex offenses) can range from 0 (*no prior sex offenses*) to 3 (*six or more sex offense chargers or four or more convictions*), we examined agreement for each initial Item 5 score value to better understand disagreements on this item. Of the 294 disagreements, 71.0% ($n = 209$) were 1-point differences, 26.2% ($n = 77$) were 2-point

²We used percentage agreement instead of kappa values as our index of agreement because percent agreement is more directly relevant to the coding for item variables in our regression analysis (agree vs. not). Kappa values were .83 (Item 1), .78 (Item 2), .67 (Item 3), .73 (Item 4), .67 (Item 5), .58 (Item 6), .58 (Item 7), .76 (Item 8), .77 (Item 9), and .89 (Item 10).

differences, and 2.7% ($n = 8$) were 3-point differences. As with Static-99 total scores, Item 5 agreement decreased as score values increased. The two evaluators assigned the same score in 92.5% of the cases in which the initial score was 0, 83.7% when the initial score was 1, 40.4% when the initial score was 2, and 40.5% when the initial score was 3.

For most items, it was less common for offenders to receive a score of 1 or higher (present) than 0 (absent): age (4.6%), lived with partner (32.6%), index nonsexual violence (5.2%), prior nonsexual violence (19.1%), prior sex offenses (37.9%), sentencing occasions (24.2%), noncontact sex offense (4.6%), stranger victim (17.3%), and male victim (17.4%). For these same items, agreement was weaker for those scoring 1 (or higher) than 0: age (72.6% vs. 99.9%), lived with partner (85.0% vs. 92.9%), index nonsexual violence (62.7% vs. 98.9%), prior nonsexual violence (68.7% vs. 89.4%), sentencing occasions (57.7% vs. 95.0%), noncontact sex offense (63.0% vs. 97.8%), stranger victim (88.4% vs. 93.7%), and male victim (98.2% vs. 96.4%). The pattern was reversed for the unrelated victim item, in which it was more common for offenders to receive a score of 0 (65.4%), and agreement was weaker for those scoring 0 than 1 (77.5% vs. 96.0%).

Relation between item disagreement and total scores—We used linear regression to examine whether particular item score disagreements were more likely for offenders with higher than lower Static-99 total scores. The dependent variable in this model was the initial Static-99 total score. We entered 10 dummy-coded variables into the regression model indicating whether the two evaluators agreed (coded 0) or disagreed (coded 1) in their scoring for a specific item.

The regression analysis (Table 4) revealed positive regression coefficients for the nine items in which agreement was stronger for an item score of 0 than 1 (or higher), and a negative coefficient for the one item in which agreement was stronger for a score of 0 (unrelated victim). These generally positive coefficients indicate that the likelihood of disagreement for the item increased as Static-99 total scores increased. This pattern was strongest for Item 6 (sentencing occasions; $\beta = .22, p < .001$) and Item 5 (prior sex offenses; $\beta = .18, p < .001$).

An unexpected finding was that disagreement about Item 1 (age at release) was one of the larger effects (age; $\beta = .17, p < .001$). There were 21 cases (1.3%) with scoring disagreements on Item 1. Most of these offenders (95.2%) had received a score of 1 during their initial assessment, and a score of 0 during their second assessment. There were seven offenders who were under the age of 25 for both assessments, suggesting that disagreements might be attributable to changes in anticipated release dates or age-calculation errors. There were 14 offenders who were over the age of 25 at both assessments (25 to 72 years), suggesting clear scoring errors.

Related scoring errors—One possible explanation for the pattern of decreasing agreement with increasing Static-99 total scores is that item disagreements are related to one another. We used odds ratios to examine whether disagreement on one item may increase the likelihood of a disagreement on another. For example, we compared whether there was an association between disagreeing about Item 4 (1 = disagree, 0 = agree) and Item 5 (1 = disagree, 0 = agree). These analyses were admittedly exploratory and resulted in 45 item

comparisons, but they did identify several instances in which disagreement about one item was associated with an increased likelihood of disagreement about another item.

The largest associations involved Items 3, 4, and 5. When evaluators disagreed in their scoring for Item 3, they disagreed in their scoring for Item 4 for 38.3% of offenders. When they agreed in their scoring for Item 3, they disagreed in their scoring for Item 4 for only 7.2% of offenders (OR = 8.03, 95% CI [4.43, 14.92]). But, these joint disagreements tended to cancel one another out with respect to their influence on the total score. There were 18 offenders with disagreements on Items 3 and 4. For 14 offenders, an increase in one item was associated with a decrease in the other. There was a similar pattern for Items 3 and 5. When evaluators disagreed in their scoring for Item 3, they disagreed in their scoring for Item 5 for 53.2% of offenders. When they agreed in their scoring for Item 3, they disagreed in their scoring for Item 5 for only 17.4% of offenders (OR = 5.40, 95% CI [3.00, 9.72]). But, once again, these joint disagreements tended to cancel out, with an increase in one item leading to a decrease in the other for 17 of the 25 offenders with disagreements on both items.

Discussion

In this study of Static-99 scores assigned and used as part of routine correctional practice for more than 20,000 sexual offenders, both the likelihood and the size of score disagreements increased as Static-99 scores increased. These findings are concerning in that it is offenders with the highest Static-99 scores who are considered for the most restrictive and expensive offender management programs. For example, Texas considered scores of 6 and higher to indicate high risk (during the timeframe of this study). We found that approximately 40% of offenders with initial scores in this high-risk range scored in a lower risk range upon reassessment. The odds of risk-level disagreement for an offender who initially scored in the high-risk range were more than 50 times higher than the odds of disagreement for an offender who initially scored in a lower risk range.

Although these findings may seem to suggest poor reliability, the overall level of rater agreement was strong. The ICC_{A,1} value for Static-99 total scores was .81, which falls between the .78 (Miller, Kimonis, Otto, Kline, Wasserman, & Otto, 2012) and .88 (Boccaccini et al., 2012) values reported in other U.S. Static-99 field studies. Thus, our findings appear to provide information about score disagreements in a typical U.S. field setting, where rater agreement is good but not necessarily excellent. Although research teams have been able to achieve better rater agreement when the Static-99 is scored for research purposes (see Hanson & Morton-Bourgon, 2009), it is important to study field scores if we want to understand the real-world strengths and limitations of Static-99 scores.

At the same time, there are important limitations associated with using field scores to study rater agreement. In a traditional rater-agreement study, two evaluators score the same person at the same point in time, using the same materials. In our field setting, scores were often assigned months or even years apart. We focused on initial scores because many offenders receive only one score, and we wanted our findings to provide information about the extent to which these initial scores may change upon rescoring. This feature of our study raises

questions about whether our data provide information about rater agreement or test–retest reliability, where one question of interest is the extent to which reliability is impacted by legitimate score changes.

Static-99 item scores should not change while an offender is incarcerated, unless he was charged with a new crime, passed his 25th birthday, or had a change in anticipated release date (that would have affected age-at-release). Because we removed these offenders from our analyses, we know of no systematic reason why scores should change over time. But it is unlikely that we accounted for every instance of a legitimate score change. When we included years between evaluations as a variable in our analyses, we found that agreement did tend to increase as the number of years between evaluations decreased, although this effect was small. But there was still clear evidence of decreasing agreement with increasing scores, even after controlling for the amount of time that elapsed between evaluations.

Ultimately, we describe our findings as providing information about rater agreement, which is consistent with contemporary field reliability studies (Miller et al., 2012; Sturup et al., in press). But any interpretation of our findings should focus more on the nature of the data used to obtain them as opposed to the label of the reliability coefficient. Our findings provide information about the extent to which the first Static-99 score an offender receives is likely to change if he is re-evaluated, and the likelihood of subsequent scores being lower or higher than his initial score. This is true regardless of whether we call the findings rater-agreement or test–retest reliability.

Regression to the Mean

Regression to the mean describes a pattern of findings in which “unusually large or small measurements tend to be followed by measurements that are closer to the mean” (Barnett, van der Pols, & Dobson, 2005, p. 215). Regression to the mean occurs “any time a group of people is assessed on two occasions and the correlation between the two tests is less than perfect” (Streiner, 2001, p. 73). Several of our findings are consistent with regression to the mean. First, offenders with initial scores below the mean tended to obtain higher scores upon reassessment, while those with initial scores above the mean tended to obtain lower scores upon reassessment (see Table 1). Second, the proportion of disagreement cases regressing toward the mean increased as scores became farther away from the mean. Third, our linear regression findings indicated that the size of disagreements increased as initial scores increased. These last two findings are consistent with regression to the mean because the size of the regression to the mean effect increases as scores become more extreme (Campbell & Kenny, 1999). Because there are more Static-99 scoring options above (3, 4, 5, 6, 7, 8, 9, 10) than below (0, 1, 2) the Static-99 mean, we see more disagreement and larger disagreements for high Static-99 scores than low scores. It is important to note that the mean Static-99 total score was not uncharacteristically low in our sample. The mean Static-99 score across samples is about 3.10, with only one sample reporting a mean of 6.00 (see Helmus, 2009). Thus, there will be more score options above the mean than below in most Static-99 samples.

If there were more score options below the mean, we might not see a pattern of decreasing agreement with increasing scores. Instead, we would see higher levels of disagreement for

both extremely high and extremely low scores. This does not, however, mean that we would expect a different pattern of findings among subsamples of especially high-scoring offenders (e.g., SVPs). Regression to the mean refers to a shift toward the population mean (Streiner, 2001). Indeed, methodologists encourage treatment researchers to avoid study designs in which participants are selected based on very high pretest scores because most of these high scoring participants will score lower on a re-test due to regression to the mean, making it difficult to distinguish true treatment effects from regression effects (e.g., Streiner et al., 2001).

Although our findings provide evidence of regression to the mean, regression cannot completely explain our findings. Because regression toward the mean impacts both high and low scores, we should expect similar findings for scores that are the same distance away from the mean. In our sample, a score of 1 and a score of 4 are both about 1.5 points away from the mean. Although disagreements for offenders with initial scores of 1 and 4 tended to regress toward the mean at the same rate (64.5% and 65.1%, respectively), there were more disagreements for offenders with an initial score of 4 (53.1%) than an initial score of 1 (34.6%).

There are several possible explanations for this finding of decreasing agreement with increasing scores, even for scores below the mean. First, it may be the result of there being few lower score options for those with already low scores (floor effect). Although we expect scores to regress to the mean, some scoring disagreements should be in the opposite direction, especially for scores close to the mean. There are no (score = 0) or few (score = 1) scoring options away from the mean for low Static-99 scores, which may have led to higher agreement for offenders with these scores. A second possible explanation is that this pattern is due to positive correlations between item disagreements, with a disagreement on one item being associated with an increase in the likelihood of a disagreement on another item (or items). Our exploratory comparisons of item score disagreements indicated several instances in which disagreeing about one item was associated with an increased likelihood of disagreeing on another, but these disagreements tended to cancel out, with an increase in one item being associated with a decrease in the other. Finally, it could be that the offenders who tend to get higher scores are simply more difficult to score. High-scoring offenders may have more records for evaluators to review and score, which may increase the likelihood of measurement error. Indeed, we found that score disagreements were more common for items that required counting offenses or sentencing dates. For Item 5 (prior sex offenses), disagreement was more common (about 60% of cases) for high scores than low scores (fewer than 20% of cases). In other words, agreement was lower when there were more offenses to count.

Item Scores

There was no single item that appeared to explain the overall pattern of decreasing agreement with increasing scores. For nine of the 10 Static-99 items, the likelihood of disagreement for the item increased as initial Static-99 total scores increased. The pattern was strongest for Item 6 (sentencing occasions) and Item 5 (prior sex offenses). There are several possible reasons why disagreements on these two items may have had the strongest

association with total scores. For example, these were among the items most likely to be scored 1 or higher (37.9% and 24.2%, respectively), meaning there was a higher likelihood (compared with other items) that they contributed to the initial Static-99 score, which tended to decrease upon rescoring. In addition, Item 5 is the only item for which a scoring disagreement can lead to a total score change of 2 or more points, which occurred for 85 (28.9%) offenders. Thus, a disagreement about Item 5 could lead to a larger total score difference than any other item. Finally, Items 5 and 6 both require evaluators to count incidents (e.g., offenses, sentencing dates), whereas other items are scored as *yes* or *no*. It may be that evaluators are more likely to miscount incidents than to completely miss the presence or absence of an offender or offense characteristic.

Implications for Practice

Although we cannot definitively explain why agreement for Static-99 scores decreases as scores increase, the inevitability of regression to the mean, the frequency of disagreements for high Static-99 scores, and the size of disagreements for high Static-99 scores all suggest that evaluators and systems should carefully consider adopting practices to account for measurement error in Static-99 scores, especially for high-scoring offenders. Because Static-99 and Static-99R item scoring rules are identical for nine of 10 items, our findings also have implications for Static-99R scoring.

Perhaps the most important implication is that evaluators and systems should do all they can to improve scoring accuracy and rater agreement. Indeed, we found that the pattern of decreasing agreement with increasing scores was smaller among the sub-sample of offenders scored between 2004 and 2011. Rater agreement was stronger in this subsample ($ICC_{A,1} = .88$), than in the sample as a whole ($ICC_{A,1} = .81$), which may be attributable to the release of an updated scoring manual in 2003 (Harris et al., 2003). Agencies that rely on the Static-99 or Static-99R may be able to improve accuracy and reliability by requiring staff to participate in ongoing training and supervision, and private practitioners may benefit from continued training as well. Indeed, when researchers checked the scoring accuracy of Static-99 scores assigned to more than 3,000 offenders in New Jersey, they found that the most common type of scoring error was failing to follow the scoring rules in the manual (Quesada, Calkins, & Jeglic, 2013). Scoring rules for Static-99 and Static-99R items are sometimes more complicated than they first appear. For example, even the scoring rules for the age-at-release item can become complicated when an offender with a history of sexual offending is scored after committing a new nonsexual offense (Cauley & Brownfield, 2013). Continued training may be the most useful way to ensure that evaluators follow appropriate scoring rules.

Another option is to have multiple evaluators score offenders who receive high scores. One limitation of using multiple evaluators is that there is no clear method for integrating scores. Scores averaged across evaluators are, by definition, more reliable than scores from a single evaluator, but averaged scores can result in Static-99 score values that do not have a clear interpretation (e.g., 3.3, 4.5). Another option is to consider the most recent score to be a reliability check and use it for decision making. For example, Texas identifies one score—usually the most recent score—as the “control score” and uses it for decision making. Our

findings that score changes show evidence of regression to the mean provide some support for this approach, at least with respect to identifying and correcting inaccurate scores at the extremes. Yet another option is to have multiple evaluators score the Static-99 and then have the evaluators discuss and resolve scoring disagreements. This type of consensus process may be especially useful for identifying clear scoring errors, such as miscalculation and overlooking offense information. However, there may also be drawbacks to the consensus approach, including the time and costs associated with having staff meet to discuss scores.

But none of these options can completely prevent random measurement error, which exists for all assessment measures. For this reason, all evaluators and systems that use the Static-99 and related measures should consider using confidence intervals to assist in their decision making processes. For those who use recidivism rates, the Static-99 authors have provided confidence intervals for Static-99 and Static-99R estimated recidivism rates (see the Static-99 Recidivism Tables; Static-99 Clearinghouse, 2008; Phenix, Helmus, & Hanson, 2012). If we assume that the amount of measurement error in the normative samples is similar to that in field settings, then these recidivism rate confidence intervals should adequately account for measurement error. Indeed, measurement error is one reason why observed scores do not fall exactly on the regression line (i.e., one reason why there is error in prediction). Although the Static-99 authors encourage evaluators to report confidence intervals for recidivism rates, findings from a recent study of more than 100 SVP evaluators found that only 48% of those who report recidivism rates also report confidence intervals for those rates (Chevalier, 2013). A clear implication of our findings is that evaluators who report recidivism rates should also report confidence intervals for those rates.

Systems that use cut scores to separate offenders into groups should consider using confidence intervals based on the standard error of measurement. These confidence intervals are not reported in the Static-99 or Static-99R manuals (Harris et al., 2003; Phenix et al., 2012) but have been estimated by researchers to be about ± 1 point (Boccaccini et al., 2012). But our findings of larger disagreements for larger scores suggest this interval may not adequately capture the likelihood of large disagreements among high-scoring offenders. Because score disagreements were often as large as 2 points for high-scoring offenders, perhaps the most defensible approach for systems that use high cut scores to separate offenders into groups is to rescore or use consensus judgment for all offenders within 2 points of the cut score.

Limitations

The extent to which our Static-99 findings generalize to other settings is unclear. Although our findings may be limited to Texas evaluators or records, the similarity of our rater-agreement coefficients to those from Florida and California (Hanson, 2001; Miller et al., 2012) and the clear pattern of regression to the mean argue against these findings being unique to Texas. As with other field studies (e.g., Boccaccini et al., 2012), we cannot guarantee that evaluators were blind to the Static-99 scores assigned by previous evaluators. Access to prior scores may have led to inflated rater-agreement coefficients, but this may be unavoidable in many field settings. Forensic evaluators are trained to review collateral information before evaluating an offender, and those who score the Static-99 and Static-99R

in field settings do so based on the information in the offender's file. Thus, field evaluators will inevitably see prior risk measure scores if they are in an offender's file. Although this may limit the extent to which our findings can be compared with those from studies in which raters are blind to prior scores, our findings provide important information about agreement in field settings for real-world scores that were used to make decisions about offenders.

Another limitation of this study is that our agreement analyses necessarily focused on subsets of offenders with at least two scores from the same period of incarceration. These offenders had marginally higher Static-99 total scores ($d = 0.11$) than those who were not rescored, and it could be that our findings apply to only these offenders. We were also unable to study evaluator characteristics that might help explain disagreement (e.g., evaluator training, position) or rescore the Static-99 to identify scoring errors. We encourage researchers to conduct similar studies in other jurisdictions to determine whether the pattern of decreasing agreement with increasing scores documented in this study applies to other contexts. It is especially important for researchers to examine whether this pattern applies to scores assigned by other types of evaluators (e.g., private practitioners retained for SVP hearings) and to identify variables that may explain it (e.g., file quality, record quantity). Finally, we were not able to examine the relationship between score changes and predictive validity. The Static-99 is, after all, designed to provide information about future offending.

Conclusion

Our findings add to the small but growing body of research documenting good to strong rater-agreement coefficients for Static-99 field scores (e.g., Boccaccini et al., 2012; Hanson, 2001). Indeed, Static-99 scores appear to be less vulnerable to the types of subjectivity and bias noted for other measures used in risk assessment (see Murrie, Boccaccini, Guarnera, & Rufino, 2013; Murrie et al., 2009). At the same time, our findings show that good agreement coefficients for Static-99 scores do not mean that evaluators necessarily assign the same score, especially for high-scoring offenders. This is especially important for the Static-99 and Static-99R, because each possible score on these measures has its own interpretation. A 1-point difference in a Static-99 score could influence whether an offender receives intensive community management or is evaluated for SVP civil commitment. Our findings suggest that evaluators and systems consider adopting practices to account for measurement error in Static-99 scores, including rescoring those scoring within 2 points of a cut score and using confidence intervals to assist in decision-making processes.

References

- Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: What is it and how to deal with it. *International Journal of Epidemiology*. 2005; 34:215–220.10.1093/ije/dyh299 [PubMed: 15333621]
- Boccaccini MT, Murrie DC, Mercado C, Quesada S, Hawes S, Rice AK, Jeglic E. Implications of Static-99 field reliability findings for score use and reporting. *Criminal Justice and Behavior*. 2012; 39:42–58.10.1177/0093854811427131
- Campbell, DT.; Kenny, DA. A primer on regression artifacts. New York, NY: Guilford Press; 1999.
- Cauley, DR.; Brownfield, MD. Static-99R: Item #1—What is the offender's age? A lack of consensus leads to a defective actuarial. 2013. Retrieved from Social Science Research Network at <http://ssrn.com/abstract=2237968>

- Chevalier, CS. Unpublished master's thesis. Sam Houston State University; Huntsville, TX: 2013. Evaluators' use of PCL-R and Static-99R scores in forensic reports.
- Ducro C, Pham T. Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research and Treatment*. 2006; 18:15–26.10.1177/107906320601800102 [PubMed: 16598662]
- Hanson, RK. Note on the reliability of Static-99 as used by California DMH evaluators. Sacramento: California Department of Mental Health; 2001. Unpublished report
- Hanson RK, Morton-Bourgon KE. The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*. 2009; 21:1–21.10.1037/a0014421 [PubMed: 19290762]
- Hanson RK, Thornton D. Improving risk assessment for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*. 2000; 24:119–136.10.1023/A:1005482921333 [PubMed: 10693322]
- Harris, A.; Phenix, A.; Hanson, RK.; Thornton, D. Static-99 coding rules. 2003. Revised –2003 (Corrections Research User Report 2003–03). Retrieved from http://www.static99.org/pdfdocs/static-99-coding-rules_e.pdf
- Helmus, L. Unpublished master's thesis. Department of Psychology, Carleton University; Ottawa, ON, Canada: 2009. Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples. Retrieved from <http://www.static99.org/pdfdocs/helmus2009-09static-99normsmathesis.pdf>
- Helmus L, Hanson RK, Thornton D. Reporting Static-99 in light of new research on recidivism norms. *The Forum*. 2009; 21:38–45.
- Helmus L, Thornton D, Hanson RK, Babchishin KM. Improving the predictive accuracy of the Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment*. 2012; 24:64–101.10.1177/1079063211409951 [PubMed: 21844404]
- Joint Legislative Audit and Review Commission. Review of the civil commitment of sexually violent predators: Report to the governor and general assembly of Virginia—January, 2012. 2011. House Document No. 5 Retrieved from <http://jlarc.virginia.gov/reports.shtml>
- Miller CS, Kimonis ER, Otto RK, Kline SM, Wasserman AL. Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment*. 2012; 24:944–953.10.1037/a0028411 [PubMed: 22563984]
- Murrie DC, Boccaccini MT, Guarnera L, Rufino KA. Are forensic experts biased by the side that retained them? *Psychological Science*. 2013; 24:1889–1897.10.1177/0956797613481812 [PubMed: 23969777]
- Murrie DC, Boccaccini MT, Turner D, Meeks M, Woods C, Tussey C. Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law*. 2009; 15:19–53.10.1037/a0014897
- Phenix, A.; Helmus, L.; Hanson, RK. Static-99R and Static-2002R evaluators' workbook. 2012. Retrieved from http://www.static99.org/pdfdocs/Static-99RandStatic-2002R_EvaluatorsWorkbook2012-07-26.pdf
- Quesada SP, Calkins C, Jeglic EL. An examination of the interrater reliability between practitioners and researchers on the Static-99. *International Journal of Offender Therapy and Comparative Criminology*. 2013 Advance online publication. 10.1177/0306624X13495504
- Rice, AR.; Boccaccini, MT.; Collier, TL. Evaluator differences in assigning Static-99 scores. Poster presented at the meeting of the American Psychology Law Society; San Juan, PR. 2012 Mar.
- Static-99 Clearinghouse. Static-99 recidivism tables: October 2008. 2008. Retrieved from <http://www.static99.org/pdfdocs/detailedrecidivismtablesOctober2008.pdf>
- Streiner DL. Regression toward the mean: Its etiology, diagnosis, and treatment. *Canadian Journal of Psychiatry/Revue canadienne de psychiatrie*. 2001; 46:72–76.
- Sturup J, Edens JF, Sorman K, Karlberg D, Fredriksson B, Kristiansson M. Field reliability of the Psychopathy Checklist–Revised among life-sentenced prisoners in Sweden. *Law and Human Behavior*. in press. 10.1037/lhb0000063

- Texas Department of Criminal Justice Parole Division. Public concern and offender release/supervision. 2011. Retrieved from http://www.tdcj.state.tx.us/documents/parole/01.01.08_parole_policy.pdf
- Wechsler, D. WAIS–IV: Administration and scoring manual. San Antonio, TX: Pearson; 2008.
- Wright, RE. Logistic regression. In: Grimm, LG.; Yarnold, PR., editors. Reading and understanding multivariate statistics. Washington, DC: American Psychological Association; 1998. p. 217-244.

Table 1

Rater Agreement and Direction of Change in Score as Determined by Initial Static-99 Score

Initial score	Frequency	% same score ^a	When scores differed	
			% most recent lower ^b	% most recent higher ^c
0	2163	74.1		100.0
1	3907	65.4	35.5	64.5
2	4892	57.8	46.8	53.2
3	4472	50.5	53.5	46.5
4	3353	46.9	65.1	34.9
5	1837	42.1	73.8	26.2
6	839	39.9	80.2	19.8
7	337	39.8	78.8	21.2
8	136	36.0	90.8	9.2
9	45	46.7	87.5	12.5
10	2	0.0	100.0	

Note. Bold = value shows that disagreements tended to be toward the overall sample mean of 2.67.

^a Percentage of cases in which the initial score is the same as the most recent score for the same offender.

^b Percentage of cases in which the initial score and most recent score differ, and the most recent score decreased from the initial score.

^c Percentage of cases in which the initial score and most recent score differ, and the most recent score increased from the initial score.

Logistic Regression Models Examining the Relation Between Offenders' Initial Static-99 Score and Subsequent Scoring Disagreement

Table 2

Model/predictors	B	SE B	OR	95% CI		Model χ^2
				Lower	Upper	
Initial score ($N = 21,983$)						828.20***
Constant	-0.825***	0.026				
Initial score	0.230***	0.008	1.26	1.24	1.28	
Initial score and years between evals ($N = 21,983$)						1,210.39***
Constant	-0.215***	0.014				
Initial score	0.214***	0.008	1.24	1.22	1.26	
Years between scores	0.108***	0.006	1.11	1.10	1.13	
Score $\times$ Years interaction	0.004	0.003	1.00	1.00	1.00	
Scores between 2004 and 2011 ($n = 12,332$)						288.95***
Constant	-1.205***	0.036				
Initial score	0.193***	0.011	1.21	1.19	1.24	

Note. OR = odds ratio; CI = confidence interval; evals = evaluations.

***
 $p < .001$.

Linear Regression Models Examining the Relation Between Offenders' Initial Static-99 Score and the Standard Deviation of Static-99 Scores Across Evaluations

Table 3

Model/predictors	β	<i>SE B</i>	95% CI <i>B</i>	Model <i>R</i> ²
Initial score (<i>N</i> = 21,983)				.053***
Constant	0.239	.007	[.226, .252]	
Initial Static-99 score	0.231***	.002	[.070, .079]	
Initial score and years between evals (<i>N</i> = 21,983)				.066***
Constant	0.436	.004	[.429, .443]	
Initial score	0.216***	.002	[.065, .074]	
Years between scores	0.108***	.001	[.021, .027]	
Score × Years interaction	0.021**	.001	[.001, .004]	
Scores between 2004 and 2011 (<i>n</i> = 12,332)				.035***
Constant	0.184	.008	[.169, .199]	
Initial score	0.160***	.003	[.041, .051]	

Note. CI = confidence interval; evals = evaluations

**
p < .01.

***
p < .001.

Table 4
Linear Regression Model Examining Relation Between Static-99 Item Score Agreement and Static-99 Total Score

Predictor	β	<i>B</i>	<i>SE B</i>	95% CI <i>B</i>	<i>R</i> ²
Item 1 (age)	.17***	2.56	0.33	[1.91, 3.22]	.029
Item 2 (intimate partner)	.08***	0.48	0.13	[0.23, 0.73]	.007
Item 3 (index: assaultive)	.07**	0.74	0.23	[0.29, 1.19]	.005
Item 4 (prior: assaultive)	.13***	0.80	0.14	[0.52, 1.08]	.015
Item 5 (prior sex offenses)	.18***	0.78	0.10	[0.58, 0.97]	.030
Item 6 (sentencing)	.22***	1.08	0.11	[0.86, 1.29]	.047
Item 7 (noncontact)	.07**	0.62	0.20	[0.23, 1.01]	.005
Item 8 (unrelated victim)	-.14***	-0.75	0.12	[-0.99, -0.51]	.018
Item 9 (stranger victim)	.13***	0.83	0.15	[0.54, 1.11]	.016
Item 10 (male victim)	.05*	0.43	0.21	[0.01, 0.84]	.002

Note. Each item coded as 0 = evaluators assigned identical score, 1 = evaluators assigned different scores. *N* = 1,594. Model *R*² = .22. CI = confidence interval.

*
p < .05.
**
p < .01.
***
p < .001.