Running head: RISK ASSESSMENT ACCURACY

Does Staff See What Experts See?

Accuracy of Front Line Staff in Scoring Juveniles' Risk Factors

Patrick Kennealy

Travis County Community Justice Services

Jennifer L. Skeem

University of California, Berkeley

jenskeem@berkeley.edu

Isaias Hernandez, Parajo Valley Unified School Services

Abstract

Although increasingly complex risk assessment tools are being marketed, little is known about "real world" practitioners' capacity to score them accurately. In this study, we assess the extent to which 78 staff members' scoring of juveniles on the California-Youth Assessment and Screening Instrument (CA-YASI; Orbis Partners, 2008) agree with experts' criterion scores for those cases. There are three key findings. First, at the total score level, practitioners manifest limited agreement (*Mean ICC* = .63) with the criterion: Only 59.0% of staff scores the tool with "good" accuracy. Second, at the subscale level, practitioners' accuracy is particularly weak for treatment-relevant factors that require substantial judgment—like procriminal attitudes (*M ICC* = .52)—but good for such straightforward factors as legal history (*M ICC* = .72). Third, practitioners' accuracy depended on their experience—relatively new staff's scores were more consistent with the criterion than those with greater years of experience. Results suggest that attention to parsimony (for tools) and meaningful training and monitoring (for staff) are necessary to realize the promise of risk assessment for informing risk reduction.

Keywords: juvenile justice, inter-rater reliability, risk assessment, risk-needs

Does Staff See What Experts See?

Accuracy of Front Line Staff in Scoring Juvenile's Risk Factors

Risk assessment instruments have become widely adopted in state juvenile justice agencies. In fact, 34 states have implemented a juvenile risk assessment tool for use in juvenile probation departments (Wachter, 2015). A variety of instruments are being marketed to juvenile justice agencies, including Multi-Health Systems' Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002), Northpointe's COMPAS Youth (CRN; Brennan, 2003), and Orbis Partners' Youth Assessment and Screening Instrument (YASI; Orbis Partners, Inc., 2007). Orbis Partners Inc.'s risk assessment instruments alone are administered in over 70 criminal justice systems throughout North America (http://www.orbispartners.com). Depending on the size of the agency, the cost of implementing risk assessment tools can reach as much as $1.2 million (Baird, Healy, Johnson, Bogie, Dankety, & Scharenbroch, 2013). Although the tools are viewed as an essential component of "evidence-based practice" (Vincent, Guy, & Grisso, 2012), relatively little is known about their reliability and validity in real world settings.

Given recent reinvigoration of the juvenile justice system's focus on rehabilitation, juvenile risk assessment instruments often measure "dynamic" or variable risk factors to target in treatment (Skeem, Scott, & Mulvey, 2014). Variable risk factors can be individual (e.g., personality, attitudes, and substance use) and/or contextual (e.g., family, peers, and community). Juvenile justice agencies can use risk scores to inform disposition, placement, and treatment decisions. In fact, research indicates that recidivism is reduced when programs match the intensity of supervision to adolescents' level of risk, and use ratings on treatment-relevant (i.e., variable) risk factors to allocate services and treatment (Hoge & Andrews, 2006; Vitopoulos, Perterson-Badali, & Skilling, 2012).

The utility of risk assessment tools for informing decisions about juveniles depends on how well the tool is implemented (Bonta, Bogue, Crowley, & Mottuk, 2001; Latessa & Lovins, 2010; Vincent, Paliva-Salisbury, Cooke, Guy, & Perrault, 2012). At the most basic level, there must be demonstrations that staff can score the tool accurately. When staff have inadequate accuracy, a juvenile's global level of risk could be mis-classified (e.g., as "low" when it is really "moderate"), and his or her most pronounced risk factors for recidivism could be mis-represented (e.g., as "antisocial cognition" when "substance abuse" is more pronounced). Subscale scores are most relevant to allocating services designed to reduce risk. For example, if a subscale assessing antisocial attitudes is scored inaccurately, a youth who needs cognitive-behavioral therapy to address these attitudes may not receive such services. Misallocation of resources is wasteful at best—and may even lead to iatrogenic effects (Lipsey, 2009; Gatti, Tremblay, & Vitaro, 2009).

The utility of a juvenile risk assessment instrument may also be complicated by the tool's own characteristics. For example, the simple length of an instrument can vary tremendously (e.g., the SAVRY is 30 items, whereas the YASI is 117 items). As tools increase in length and/or complexity, this places greater burden on staff to complete the assessment and can introduce noise that compromises the instrument's reliability (Baird, 2009).

In the risk assessment literature, there has been a greater focus on *inter-rater reliability* rather than *rater accuracy.* Although these phrases are sometimes used interchangeably, they have different meanings. Inter-rater reliability—the degree to which different practitioners produce the same rating for the same case—does not necessarily translate into rater accuracy, or a correct rating for that case (Wu, Whiteside, & Neighbors, 2007). Practitioners can be reliable, but wrong. Rater accuracy is the degree that a practitioner's rating is concordant with the correct rating or criterion score. In this study, we assess the extent to which practitioners' scores for juveniles agree with expert criterion scores for the same cases. Our estimates are not stringent, in that practitioners' and experts' scores are based on identical information (i.e., video recorded interviews and file information), which eliminates inaccuracy that can arise from discrepant information obtained in independent interviews.

Global Level Reliability and/or Accuracy

At the total score level, it may be fairly easy to attain adequate inter-rater accuracy. Why? There are countless ways to obtain a given total score on a risk assessment tool by combining different scores across items and subscales. Nonetheless, if risk assessment instrument total scores are used to designate low, medium, and high levels of risk for re-offending, inter-rater accuracy at this level is important for informing decisions about placement and release.

A recent multi-site study indicated that staff's inter-rater reliability (not accuracy) in scoring risk assessment tools varies by site and tool (Baird et al., 2013). This study featured evaluations of inter-rater reliability of the YLS/CMI in Arkansas, the CRN (a variant of the COMPAS Youth) in Georgia, and the YASI in Virginia. At each site, researchers developed 10 videotaped reliability cases, administered those cases to staff who had been trained to score the tool, and then computed inter-rater reliability (i.e., the consistency with which practitioners scored cases—*not* the extent to which those scores were correct). They found good inter-rater reliability for total scores on the YASI (*ICC* = .89; *n* = 69) and CRN (*ICC* = .93; *n* = 50), but not the YLS/CMI (*ICC* = .67; *n* = 15). These findings indicate that staff can manifest good agreement amongst each other, but reliability is not guaranteed (Skeem, Barnoski, Latessa, Robinson, & Tjaden, 2013). Because practitioner ratings were not compared to expert-developed criterion ratings, these findings do not indicate whether staff is scoring the instruments correctly.

In perhaps the strongest of the few field studies of total score inter-rater accuracy, Vincent, Guy, Fusco, and Gershenson (2012) evaluated the Structured Assessment of Violence Risk for Youth (SAVRY; Borum, Bartel, & Forth, 2006) as implemented in Louisiana. The SAVRY has 30 items and 4 scales: Historical, Social/Contextual, Individual/Clinical, and Protective. Prior to field reliability assessments, all juvenile probation officers underwent 2-day SAVRY training workshops, which included standardized vignette-based accuracy assessments. Next, 3 trained research assistants observed 36 juvenile probation officers interview 80 young offenders. Afterward, both the officer and research assistant rated the young offender on the SAVRY. The authors found good levels of inter-rater accuracy (ICC > .75; Parkerson, Broadhead, and Tse, 1993) between staff and a researcher for SAVRY total scores (*ICC* = .86). These findings indicate that a well-validated tool can be accurately scored at the global level, when implemented in a setting where staff was rigorously trained to reliability *before* using the tool.

Subscale Level Reliability and/or Accuracy

Evidence for the inter-rater reliability and accuracy of juvenile risk assessment tools is less apparent at the subscale level than the total score level. One of the few studies to investigate field accuracy at the subscale level amongst juvenile offenders was the study of Vincent, Guy, Fusco, and Gershenson (2012), described above. Here, good levels of field accuracy were observed for the SAVRY's Historical factor (*ICC* = .81; e.g., history of offending, violence, and childhood abuse) and Individual/Clinical factor (*ICC* = .86: e.g., substance use, anger, and attentional issues). Less promising findings were found for Social/Contextual scores (*ICC* = .67: e.g., coping ability, parental supervision, and personal/social support). Notably, each of the treatment-relevant risk factors (e.g., substance use, anger, and attentional issues) of the SAVRY's

Individual/Clinical factor is assessed by one item. Not only may this result in a less nuanced assessment of a risk factor such as substance use, but it could also reduce the likelihood of inter-rater reliability or accuracy issues emerging because fewer items provide less of an opportunity for staff to rate items differently.

Present Study

In the present study, we investigate the inter-rater accuracy of California's state juvenile corrections staff (*n* = 78) in scoring a juvenile risk assessment tool. Vincent, Guy, Fusco, and Gershenson (2012) nicely characterized inter-rater accuracy in a progressive or "model" implementation context. The context studied here may be more representative of standard practice in juvenile justice (see Baird et al., 2013). Compared to the model implementation context, this standard implementation context involves an instrument that is less well-validated (i.e., the California Youth Assessment and Screening Instrument; CA-YASI; Orbis Partners, 2008), a 1.5-day training that did not require staff to achieve certain levels of inter-rater reliability or accuracy, and a time window of one year after "roll-out" of the instrument.

Within this context, we address three aims. First, we assess the CA-YASI's inter-rater accuracy at the total score level. Unlike previous research that has either (a) employed a single expert rating as a criterion or (b) simply assessed inter-rater reliability, we assess inter-rater accuracy by using expert consensus scores as criterion (described below). We expect to find higher levels of inter-rater accuracy at this level because practitioners can score the CA-YASI's 107 items quite differently, but still end up with a total risk score that is fairly similar to the expert consensus score.

Second, we assess the CA-YASI's inter-rater accuracy at the subscale score level. We hypothesize that scales requiring less judgment (e.g., straight-forward coding of criminal history indices from records) will have higher levels of inter-rater accuracy than those that are dependent on the staff to interpret information from both interview and file review (e.g., criminal thinking). This is primarily because record coding-based items reduce the potential for individual differences in rating styles (Murrie, Boccaccini, Turner, Meeks, Woods, & Tussey, 2009).

Third, we explore the possibility that inter-rater accuracy is moderated by setting differences (i.e., site) and individual characteristics of each rater (i.e., gender and years of professional experience with the agency). Given previous research on agreement in adversarial settings (Vincent, Paiva, Cook, Guy, & Perrault, 2012; Boccaccini, Turner, & Murrie, 2008; Murrie et al., 2009; Miller, Rufino, Boccaccini, Jackson, & Murrie, 2011), we tentatively expect the characteristics of staff members (i.e., site, years of experience) to impact inter-rater accuracy. Findings have implications for the practical use of structured risk assessment tools to prevent and manage risk in routine practice settings.

Method

Sample

Participants were 78 staff members (45 women and 33 men) from all four sites of a California correctional agency. A staff member was eligible if he or she (a) completed the Orbis Partners Inc. training (described above) and (b) administered the CA-YASI on a regular basis (i.e., at least once every 90 days). Typically, each staff member completed approximately 3 CA-YASI ratings every 30 days (*M* = 3.23, *SD* = 1.69). The sample was comprised of 92% of all eligible staff employed by the agency; the other 8% (*n* = 7) did not participate due to prior job commitments, transfers within the agency, or medical emergencies. The participants were between the ages of 31 and 63 (*M* = 43.40, *SD* = 7.59). Nearly half of the staff had a master-level degree (47.8%), whereas approximately one-third had a bachelor-level degree (30.4%). The majority of staff possessed a background in mental health (60.9%; e.g., social work) or law enforcement (71.1%; e.g., criminal justice). On average, staff members worked in the agency approximately 11 years (*M* = 10.84, *SD* = 8.09).

Prior to participation in this study, all staff members completed 1.5 days of training on the CA-YASI. This training was conducted by the company that developed this risk assessment tool (Orbis Partners Inc.). Training included overview of the measure's items and scoring procedures, correctional intervention principles, motivational interviewing techniques, and completion of a practice case using video and collateral file information. Further, the training involved an overview of utilizing the tool for case planning, software support guidance, and a time for open discussion of questions.

CA-YASI

The California Youth Assessment and Screening Instrument (CA-YASI; Orbis Partners Inc., 2008) is a juvenile risk management tool that was designed to help (a) structure placement and release decisions, (b) identify supervision and intervention targets, and (c) capture change in risk of violence, infraction, and recidivism over time. The CA-YASI was customized by Orbis Partners Inc. to meet the specific requirements of the state agency where the tool was being used. This tool features 107 items and is scored based on a semi-structured interview and file review. The CA-YASI renders a total score and 12 subscale scores (see Table 1 for scale and item-content descriptions). Each CA-YASI subscale includes a mixture of dichotomous items (e.g., "present" or "not present"), 5-point Likert-type scale items ("High Risk" to "High Protection"), and a variety of count- (i.e., "previous sustained petitions) and age-based (i.e., "age at first arrest") criminal history variables. Although no published studies have investigated the psychometric properties of the CA-YASI, research has found that other versions of the YASI can be administered with inter-scorer reliability (for a description, see the introduction; Baird et al., 2013) and demonstrate predictive utility for violence and other criminal justice outcomes (Bostwick, 2010; Orbis Partners, Inc., 2007).

Procedure

Training cases. To assess the inter-rater accuracy of the CA-YASI, practitioners rated four training cases. Training cases included a video-recorded interview conducted with a youth aged 15 to 18 and file information on the youth's criminal, educational, social, physical and mental health background.

For each training case, the co-authors of this manuscript developed criterion scores and rationales by consensus as a research team based on the CA-YASI scoring guide. Prior to completing these ratings, we went through the same 1.5 days of training on using the CA-YASI (i.e., measure overview, scoring procedures, practice cases) as agency staff (described above in greater detail). To supplement this training, we relied on our research team's training (i.e., master's and doctoral level) and expertise in conducting risk assessments in research and clinical settings. Based on our training and the CA-YASI scoring guide, we rated each item by conducting a comprehensive review of the interview and collateral information and integrated information from both sources. To compliment these criterion scores, we culled information from the interview and file to ensure consistent explanations in why or why not a score was appropriate for a given item. As an additional check for quality assurance, our criterion scores and rationales were reviewed by Orbis Partners Inc. to ensure consistency with the official CA-YASI scoring guidelines.

Administration. We administered these training cases to practitioners at each of four state facilities. Each case required approximately 4 hours to complete (view the interview, review collateral information, and score the CA-YASI). After practitioners scored each case, we reviewed the correct ratings for each item and discussed any scoring discrepancies (clarifying the criterion scoring rationale as needed) in a group setting. For the majority of staff (60%), the reliability estimates reported below are based on CA-YASI ratings of at least 3 (19%) or 4 (41%) cases. The reliability estimates of a minority of staff (40%) reflect 2 cases due to scheduling conflicts and illness.

After administering the first two training cases, we found that some CA-YASI items were poorly defined. This included items from the Aggression/Violence (e.g., Anger/Frustration Tolerance; Identification of Anger Triggers, Instrumental Violence), Attitudes (e.g., Commitment to Criminal Lifestyle; Attitude Towards the Criminal Justice System), and Social/Cognitive Skills (e.g., Impulsivity, Consequential Thinking) sections. To help increase reliability levels the second two training cases, we developed additional definitions with the approval of Orbis Partners Inc. These definitions were adopted from existing established measures (e.g., Psychopathy Checklist: Youth Version; Forth, Kosson & Hare, 2003; Level of Service/Case Management Inventory; Andrews, Bonta, & Wormith, 2004). Staff members received these additional item definitions prior to administering the last two training cases. Because these supplemental definitions did not improve accuracy from the first pair of training cases (*M ICC* = .75) to the second pair of training cases (*M ICC* = .54), this is not discussed in further detail and reliability analyses are presented based on all four cases.

Analyses

The primary index of each staff member's level of agreement with expert ratings was the intra-class correlation coefficient (ICC). ICCs estimate the amount of variance in a group of ratings that is attributable to the cases being rated, rather than measurement error. In other words, ICCs index whether practitioners can consistently rate a given youth in a similar manner to the criterion, when accounting for chance agreement.

As can be gleamed from the recent debate over how to define reliability standards in the DSM-5 field trial reliability study (Kraemer, Kupfer, Clarke, Narrow, & Regier, 2012; Spitzer, Williams, & Endicott, 2012), a lack of consensus on reliability standards has existed in the fields of psychology and psychiatry for decades. In fact, several different reliability standards have been proposed. For example, Landis and Koch (1977) asserted that a coefficient greater than .61 is “substantial” and a coefficient greater than .81 is “almost perfect.” Similarly, Cicchetti and Sparrow (1981) asserted that a value of less than .40 is “poor,” .40–.59 is “fair,” .60–.74 is “good,” and .75 and above is “excellent.” To maintain consistency with recent research on the field reliability of risk assessment instruments (Vincent, Guy, Fusco, & Gershenson, 2012), we consider an ICC of .60–.74 as “good” and .75 or greater as “excellent” (Cicchetti & Sparrow, 1981). That said, we stress the importance of focusing on ICC values for full information rather than simply relying on interpretive labels.

Because ICCs can be calculated in a number of ways, three aspects of our analyses are important to note. First, we used absolute agreement estimation because it accounts not only for the degree of convergence between raters (where staff’s/expert’s *pattern* of ratings across youth is similar), but also the absolute value of ratings (whether staff’s/expert’s *actual scores* across youth is similar). Second, we report single-rater ICCs because CA-YASI scores from multiple raters are not averaged together for decision-making; instead, decisions about youth are based on a single staff member’s CA-YASI rating. Third, we computed ICCs for summed subscale and summed total scores, rather than basing ICCs on the individual items that comprise each summed score. Although this results in a higher ICC estimate, it represents practice in this state agency, where total scale and subscale scores (rather than items) are used to inform decision-making.

For a minority of staff (*n* = 9, or 12%), we could not validly calculate ICCs. In such cases, incomplete ratings (i.e., missing items) and/or deviation from the criterion (i.e., scoring mistakes) led to restricted variance in ratings. For these staff we used Kappa, an alternative measure of inter-rater agreement, to provide an estimate of reliability at the total score level. Kappa is a suitable alternative in these cases because it is calculated at the item level (where variance issues can be addressed), whereas ICCs are typically more suitable for summed scale scores. Similar to ICCs, Kappa also accounts for agreement between raters that is due to chance. Kappa was computed based on item scores at the total and subscale score level. Consistent with Cicchetti and Sparrow (1981), we adopted the same reliability standards for Kappa as ICCs.

Results

Total Score Agreement

First, we assessed inter-rater accuracy at the total score level to provide a glimpse of global performance (see Table 2). As a group, state agency staff obtained an average ICC of .63 (*SD* = .36), which is just barely in the "good" range (*ICCs* = .60–.75; Cicchetti & Sparrow, 1981). This average ICC was obtained by (a) calculating each practitioner's ICC across their completed training cases and (b) calculating the mean ICC of all practitioners. Further, an average Kappa of .44 (*SD* = .13) was found among staff participants for those whom we could not validly calculate ICCs. This Kappa of .44 indicates "fair" (.40–.59) rather than "good" (.60–.74) accuracy (Cicchetti & Sparrow, 1981). Across ICC and Kappa coefficients, these findings indicate that, on average, practitioners demonstrate "fair" to "good" levels of accuracy at the total score level.

Although average scores are useful for summarizing performance at the group level, they can mask individual differences. To paint a clearer picture at the individual level, we calculated the percent of staff whose ICC or Kappa values achieved “good” levels of agreement (*ICCs and Kappas* ≥ .60; Cicchetti & Sparrow, 1981). 59.0% of staff had good accuracy. These results indicate that 41.0% of practitioners are not rating this tool at “good” levels of accuracy. This is troubling, given how easily reliability at this level could be attained (via myriad combinations of the CA-YASI’s 107 items and 12 subscales).

Subscale Score Agreement

Next, we evaluated inter-rater accuracy at the subscale level to provide a clearer picture of agreement on measures used to inform decisions about intervention and risk reduction. We hypothesized that reliability would be higher for subscales that require less judgment relative to subscales that are dependent on the staff to interpret information from both interview and file review to make a judgment (i.e., more objective). To more clearly address this issue, we have identified CA-YASI prototypical scales that require less judgment (i.e., Legal History, Education/Employment, Family, Correctional Response) and more judgment (i.e., Aggression, Attitudes, Social Cognition, and Social Influence) based on a review of each scale’s content.

Table 2 shows the average ICCs that practitioners at each facility attained for CA-YASI subscale scores. As a group, practitioners subscale inter-rater accuracy varied substantially (ICC Range: .50 to .79), with almost half of subscales (5 of 11) falling below “good” inter-rater accuracy levels (*ICCs* ≥ .60; Cicchetti & Sparrow, 1981). Further, the mean ICC of the prototypical scales that require less judgment (*M ICC* = .72) was higher than those that require more judgment (*M ICC* = .52). Specifically, weak subscale inter-rater accuracy was particularly true of judgment-heavy scales like Attitudes (*ICC* = .53) and Social-Cognition (*ICC* = .50)—

whereas the average ICC of staff for scales that require less judgment like Legal History (*ICC* = .74) and Educational/Employment History (*ICC* = .77) reached good levels of accuracy (Parkerson et al., 1993). Together, these findings suggest that scale-level accuracy may be a function of the degree that a scale requires judgment.

Similar findings were observed at the individual level (see Table 3); where we calculated the percent of staff whose ICC or Kappa values achieved "good" levels of agreement (*ICCs and Kappas* > .60; Cicchetti & Sparrow, 1981). Overall, less than half of the CA-YASI subscales (4 out of 11) reached "good" levels of agreement by over 70% of the practitioners. Further, the mean percent of practitioners who achieved "good" reliability on the prototypical scales that require less judgment (*M* = 79.9.0%) was substantially higher than those that require more judgment (*M* = 50.3%). These results indicate that a substantial proportion of practitioners are not accurately rating this tool, particularly for the subscales most likely to be used when making treatment and services provision decisions.

Moderators of Inter-Rater Accuracy

Finally, we investigated whether inter-rater accuracy at the total score level was moderated by setting differences (i.e., site) and individual characteristics of each rater (i.e., gender and years of professional experience with the agency). For each potential moderator, we conducted a series of multilevel linear models using the "xtmixed" command using STATA 10.1 to test for the presence of significant nesting within mean ICCs. In other words, we tested whether these moderators explained variation within mean ICCs. Inter-rater accuracy of CA-YASI total score was not significantly impacted by site or gender of staff. For years of experience with the agency, we found the model that added nesting by experience significantly improved the model fit compared to the model that included no nesting (see Table 4). These analyses indicate that 6% of the variance in CA-YASI total score ratings is explained by the staff member's years of experience with the agency (*ICC* = .06), with relatively new staff members (0 to 5 years of experience) performing better than staff with more experience.

Discussion

This study is one of the first to investigate the inter-rater accuracy of a risk assessment tool (CA-YASI) completed by juvenile corrections staff in a standard practice setting. Our main findings may be organized into three points. First, over 40% of practitioners are not accurately rating the CA-YASI at even the total score level. Second, practitioners' often did not reach "good" accuracy levels at the subscale level, especially for factors that require much judgment to score like pro-criminal attitudes and criminal thinking. Third, accuracy was somewhat dependent on the rater's years of experience with the agency: Staff with less experience tended to demonstrate greater accuracy than staff with more experience.

Risk assessment tools are routinely used in juvenile justice settings to inform decision-making (e.g., release, placement, and treatment/service eligibility)—with the laudable goals of improving public safety and promoting youth development. If these tools are not rated accurately, decisions based on these tools will be misinformed. For example, inaccurate ratings may lead to the allocation of expensive resources like cognitive behavioral therapy for criminal thinking to a juvenile who does not need it…at the expense of another who would benefit greatly from such treatment. As discussed later, agencies must implement risk assessment tools more judiciously, to avoid wasting valuable resources and to realize good-faith efforts to improve youths' outcomes.

Limitations

Prior to discussing these findings in greater detail, we note important limitations that should be considered when interpreting this study's results. First, we could not disentangle the effects of two potential contributors to poor rater accuracy: the CA-YASI itself (e.g., item definitions, content, etc.) and poor implementation of the CA-YASI (e.g., insufficient training; no coaching; poor ongoing oversight). Accuracy levels, however, actually decreased after we developed more concrete definitions for problematic items (which may be regarded as a form of instrument improvement) and reviewed item-level scoring rationales for each case (which may be regarded as a form of training). Anecdotally, the length and complexity of the CA-YASI seemed to play some role in the limited accuracy observed here.

Second, this evaluation of inter-rater accuracy occurred approximately one year after the staff was initially trained to use the CA-YASI. It is possible that accuracy decreased over this roll-out period—with staff drifting away from initial training and knowledge on how to rate specific items. Nonetheless, our results are representative of accuracy after an initial roll-out period for an instrument. Whether the results generalize to immediately after training is unclear.

Limited Global Accuracy

Our first primary finding suggested that a little over half of practitioners (59.0%) obtained "good" levels of accuracy (*ICCs and Kappas* ≥ .60; Cicchetti & Sparrow, 1981) for CA-YASI total scores. As noted earlier, total score accuracy should be relatively easy to attain, because there are a variety of ways to combine different scores across items and subscales that result in the same total score. For staff members who obtained "good" accuracy, CA-YASI total score ratings could be accurately used to designate youths as having low, medium, and high levels of risk for re-offending. That said, almost half of practitioners (41.0%, *ICCs and Kappas* ≤ .60) fell below the minimum threshold for a "good" level of agreement with the criterion -- and one-quarter (29.5%, *ICCs and Kappas* ≤.40) fell below the minimum threshold for “adequate” performance (Cicchetti & Sparrow, 1981). Taken together, these findings suggest that a substantial proportion of youth in this state system are being rated inaccurately – even at the global level -- on the CA-YASI.

There are likely two sources of poor accuracy at the total score level. First, the CA-YASI is a lengthy and somewhat complex measure composed of 107 items and 12 scales. As a point of comparison, the juvenile risk assessment instrument from the study of Vincent, Guy, Fusco, and Gershenson (2012), the SAVRY, only features 30 items and 4 scales. The sheer size of CA-YASI likely makes rating it a more demanding endeavor for staff. Moreover, several CA-YASI items are poorly defined, particularly in the Aggression/Violence (e.g., Anger/Frustration Tolerance) and Attitudes (e.g., Commitment to Criminal Lifestyle) domains. The ambiguity in how several items were worded even led our research team to develop item definitions after the administration of the first two of four training cases (described in Method section), but even providing somewhat clearer definitions did not improve staff performance when rating these items.

Second, and perhaps the least commonly considered cause of poor accuracy (Gendreau, Goggin, & Smith, 1999), Orbis Partners Inc. and the California Division of Juvenile Justice may have employed an insufficient implementation program. This implementation was centered on a 1.5-day training workshop that provided an overview on the principles of risk assessment, presentation of each respective measure’s items and scoring procedures, and completion of practice ratings. In contrast with many studies conducted in research settings and the procedures used by Vincent, Guy, Fusco, and Gershenson (2012), the CA-YASI training workshop did not require staff members to achieve accuracy on the tool before they were allowed to administer it in routine practice. Further, the CA-YASI implementation program did not include booster training cases, which are intended to prevent the impact of rater drift from standard scoring criteria. As such, it is not possible to identify whether staff members have been rating the CA-YASI inaccurately from the beginning of implementation, or became unreliable at some point after the initial training workshop. The inclusion of a strong implementation plan is imperative if a risk assessment tool is going to be used to inform decisions about placement and release.

These findings are troubling because it is impossible for a risk assessment tool to inform accurate decisions about treatment, placement and release if the risk assessment tool is not reliably and accurately scored. Inter-rater reliability is a necessary (but not sufficient) characteristic of evidence-based risk assessment: “An instrument that is not reliable cannot be valid…” (Latessa & Lovins, 2010; p. 212). Staff must be capable of scoring a tool reliably and accurately if they wish to then use these scores to inform decision-making about offender placement and supervision intensity, which is the primary purpose of conducting such an assessment. However, our findings indicate that CA-YASI scores are unlikely to accurately index youths’ risk for re-offending and, as a result, misinform decisions about placement and release.

Accuracy Weak for Treatment-Relevant Subscales

Our second primary finding was that inter-rater accuracy levels were often lower at the subscale level than the total score level. In fact, ratings fell below “good” inter-rater accuracy levels (*ICCs and Kappas* ≥ .60; Cicchetti & Sparrow, 1981) for almost half of the CA-YASI subscales (5 of 11). This was particularly true of scales that required more judgment (i.e., Aggression, Attitudes, Social Cognition, and Social Influence) in comparison to scales that required less judgment (i.e., Legal History, Education/Employment, Family, and Correctional Response) on the part of the rater.

These findings are consistent with those observed by Girard (1999) with the LS/CMI in adults. Here, Girard (1999) found relatively static scales that require less clinical judgment (e.g., Criminal History = .88; Alcohol and Drug Problems = .75) could be rated with relatively good levels of reliability. In contrast, potentially dynamic scales that depend more on clinical judgment (e.g., Family/Marital = .38; Leisure/ Recreation = .26; Procriminal Attitudes = .16) were rated with relatively poor levels of reliability.

Taken together, this lack of reliability and accuracy for scales more reliant on clinical judgment is troubling, but perhaps not surprising because these scales require substantial judgment, and thus more training to score accurately. Similar to Girard (1999), the one subscale that consistently reaches adequate levels of accuracy across the present study assesses criminal justice history. Although the utility of this scale for risk reduction is limited (as one cannot reduce a criminal history), its utility for predicting recidivism may exceed the unique contributions of dynamic risk factors (McGrath & Thompson, 2012; Vincent, Chapman, & Cook, 2011). The success of criminal history may simply reflect that the best predictor of future behavior is similar past behavior (Kroner, Mills, & Reddon, 2005). Thus, it may be possible that staff ratings on criminal justice history may have predictive utility for criminal justice outcomes, but needs to be explored in future predictive utility-focused research.

These data paint a dimmer picture for the use of youth risk assessment tools in the allocation of treatment and corrections programming resources that attempt to prevent and manage recidivism risk. Quite simply, findings suggest that many DJJ staff rate youth differently than CA-YASI experts do. Other research indicates that even when dynamic risk factors are accurately assessed, youth may simply not receive the interventions that address their needs (Peterson-Badali, Skilling, & Haqanee, 2015; Vitopoulos et al., 2012). If agencies are interested in using subscales to inform decisions about which youth should be referred to substance abuse treatment, or family counseling, or cognitive-behavioral therapy for criminogenic thinking, these findings suggest that more work must be done to attain better subscale accuracy and correct allocation of resources. If not, use of such risk assessment tools will contribute to the misallocation of resources yielding unnecessary program spending and potentially to iatrogenic effects for those who were not in need of such resources (Lipsey, 2009; Gatti, Tremblay, & Vitaro, 2009).

Rater Experience Impacts Accuracy

A third key finding of this study was that CA-YASI scoring accuracy was moderated by practitioners' years of experience: staff with fewer than six years of experience performed better than staff with more experience. This finding needs to be replicated, but (speculatively) is consistent with the possibility that newer staff more readily embrace new tools than older staff, and/or are more adept at applying the scoring criteria than experience-based preconceptions. Our finding that practitioners' experience moderates their scoring accuracy is generally consistent with past findings that practitioners' with more experience are less likely to use risk assessment information in their case management decisions (Vincent, Paiva, Cook, Guy, & Perrault, 2012; Viglione, Rudes, & Taxman, 2014). These findings may also be consistent with research that found personality traits and other factors help explain scoring discrepancies (Murrie et al., 2009; Boccaccini et al., 2008; Miller et al., 2011). This could also be related to the possibility of burnout in staff that have been with the agency longer (Lambert, Hogan, Griffin, & Kelley, 2015). Future research should focus on identifying training-relevant mechanisms that explain the link between experience and scoring accuracy.

Conclusions

The findings of the present study and other research indicate that "Simply selecting and adopting a risk assessment tool will not accomplish the desired objective unless it is implemented properly" (Vincent, Guy, & Grisso, 2012, p. 7). Although clear, empirically-supported steps to achieving inter-rater accuracy do not exist, several key considerations should be made when selecting and implementing a risk assessment tool. First, a tool should not only be validated, but feasible to use in the intended setting. Selection of an instrument that is unnecessarily dense and lengthy will limit the ability of staff to effectively use the instrument with any degree of regularity (Baird, 2009). Second, training staff to inter-rater accuracy and reliability is a necessary step, and may only be possible under the necessary circumstances (i.e., sound risk assessment tool and thorough implementation plan). Further, continuous training such as booster training cases may be integral in preventing staff ratings from drifting apart across time. Third, going beyond basic classroom training efforts at the onset of implementation and extending training into practice via coaching offers promise. Such approaches have been successful at improving the implementation of core corrections practices in programs like Staff Training Aimed at Reducing Re-arrest (STARR; Robinson, Lowenkamp, Holsinger, VanBenschoten, Alexander, & Olsen, 2012). Fourth, agencies must ensure that officers do in fact do use the information obtained with the risk assessment tools to inform the supervision of their clients. Emerging research has found that officers often administer risk assessment tools, but fail to consider scores when making important supervision decisions (Viglione et al., 2014). If these issues are not carefully considered, the benefits of employing a risk assessment tool to manage an offender's risk of future recidivism may be negated. The importance of such work is only emphasized by the fact that 86% of state juvenile justice agencies had adopted risk assessment instruments as of 2003 (Griffin & Bozynski, 2003; as in Vincent, Guy, Fusco, & Gershenson, 2012), with costs sometimes reaching over one million dollars to implement and sustain such juvenile risk assessment tools in a given agency (Baird et al., 2013).

References

Andrews, D. A., Bonta, J., & Wormith, J. (2004). Manual for the Level of Service/Case Management Inventory (LS/CMI). Toronto, Ontario, Canada: Multi-Health Systems.

Baird C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime & Delinquency. Retrieved from http://nccd-crc.issuelab.org/research

Baird, C., Healy, T., Bogie, A., Wicke Dankert, E., & C. Scharenbrock. (*2013*).Risk and Needs Assessments in Juvenile Justice: A Comparison of Widely Available Risk and Needs Assessment Systems. National Council on Crime and Delinquency.

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law*, *14*, 262-283. doi: 10.1037/a0014523

Bonta, J., Bogue, B., Crowley, M., & Mottuk, L. (2001). Implementing offender classification systems: Lessons learned. In G. A. Bernfeld, D. P. Farrington, & A. Leschied (Eds.), Offender rehabilitation in practice (pp. 227-245). Chichester, UK: Wiley.

Borum, R., Bartel, P., & Forth, A. (2006). *SAVRY: Structured Assessment of Violence Risk in Youth: Professional manual*. PAR.

Bostwick, L. (2010). Mental health screening and assessment in the Illinois Juvenile Justice System. Prepared for the Illinois *Juvenile Justice Commission*. Retrieved from http://www.icjia.state.il.us/public/pdf/ResearchReports/Mental%20health%20screening%20and%20the%20juvenile%20justice%20system.pdf

Brennan, T. (2003). Youth-COMPAS psychometrics report. *Traverse City, MI: Northpointe*

Institute for Public Management.

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

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version manual.* Toronto: Multi-Health Systems.

Gatti, U., Tremblay, R. E., & Vitaro, F. (2009). Iatrogenic effect of juvenile justice. *Journal of Child Psychology and Psychiatry, 50*, 991-998. doi: 10.1111/j.1469-7610.2008.02057

Gendreau, P., Goggin, C., & Smith, P. (1999). The forgotten issue in effective correctional treatment: Program implementation. *International Journal of Offender Therapy and Comparative Criminology*, *43*, 180-187. doi: 10.1177/0306624X99432005

Girard, L. (1999). *The Level of Service Inventory-Ontario Revision: Risk/need assessment and recidivism*. University of Ottawa (Canada).

Hoge, R. D., & Andrews, D. A. (2002). The Youth Level of Service/Case Management Inventory manual and scoring key. Toronto, Canada: Multi-Health Systems

Hoge, R. D., & Andrews, D. A. (2006). Youth Level of Service/Case Management Inventory: User's manual. North Tonawanda, NY: Multi-Health Systems.

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). DSM-5: how reliable is reliable enough?. *American Journal of Psychiatry*, *69*, 13-15. doi: 10.1176/appi.ajp.2011.11010050

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*, 360-374. doi: 10.1016/j.ijlp.2004.01.011

Lambert, E. G., Hogan, N. L., Griffin, M. L., & Kelley, T. (2015). The correctional staff burnout literature. *Criminal Justice Studies*. Advance online publication. doi: 10.1080/1478601X.2015.1065830

Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims & Offenders, 5*, 203-219. doi:10.1080/15564886.2010.485900

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders, 4*, 124-147. doi:10.1080/15564880802612573

McGrath, A. & Thompson, A. P. (2012). The relative predictive validity of the static and dynamic domain scores in risk-need assessment of juvenile offenders. *Criminal Justice and Behavior, 39*, 250-263. doi: 10.1177/0093854811431917

Miller, A. K., Rufino, K. A., Boccaccini, M. T., Jackson, R. L., & Murrie, D. C. (2011). On individual differences in person perception: Raters' personality traits relate to their psychopathy checklist-revised scoring tendencies. *Assessment*, *18*, 253-260. doi: 10.1177/1073191111402460

Murrie, D. C., Boccaccini, M. T., Turner, D. B., 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

Orbis Partners. (2007). *Long−term validation of the Youth Assessment and Screening Instrument (YASI) in New York State Juvenile Probation*. Ottawa, Otario: http://criminaljustice.state.ny.us/opca/pdfs/nyltyasifullreport20feb08.pdf

Orbis Partners. (2008). California Youth Assessment and Screening Instrument (CA-YASI).

Ottawa, Ontario.

Parkerson, G., Broadhead, E., & Tse, C.-K. (1993). The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and co-morbidity. *Journal of Clinical Epidemiology, 46*, 379-393. doi: 10.1016/0895-4356(93)90153-R

Peterson-Badali, M., Skilling, T., & Haqanee, Z. (2015). Examining implementation of risk assessment in case management for youth in the justice system. *Criminal Justice and Behavior*, *42*, 304-320. doi:10.1177/0093854814549595

Robinson, C. R., Lowenkamp, C. T., Holsinger, A. M., VanBenschoten, S., Alexander, M., & Oleson, J. C. (2012). A random study of Staff Training Aimed at Reducing Re-arrest (STARR): using core correctional practices in probation interactions. *Journal of Crime and Justice*, *35*, 167-188. doi: 10.1080/0735648X.2012.674823

Skeem, J. L., Scott, E. S., & Mulvey, E. P. (2014). Justice Policy Reform for High-Risk Juveniles: Using Science to Achieve Large-Scale Crime Reduction. *Annual Review of Clinical Psychology, 10,* 709-739. doi: 10.1146/annurev-clinpsy-032813-153707

Skeem, J. L., Barnoski, R., Latessa, E., Robinson, D., & Tjaden, C. (2013). Youth risk assessment approaches: Lessons learned and question raised by Baird et al.'s study (2013). Rebuttal prepared for the National Council on Crime & Delinquency (NCCD) study funded by the Office of Juvenile Justice and Delinquency Prevention (OJJDP). Available online at: http://risk-resilience.berkeley.edu/sites/default/files/wp-content/gallery/publications/BairdRebuttal2013_FINALc1.pdf

Spitzer, R. L., Williams, J. B., & Endicott, J. (2012). Standards for DSM-5 reliability. *American Journal of Psychiatry*, *169*, 537.

Viglione, J., Rudes, D. S., & Taxman, F. S. (2014). Misalignment in Supervision: Implementing

Risk/Needs Assessment Instruments in Probation. *Criminal Justice and Behavior, 42*, 263-285. doi: 10.1037/0093854814548447.

Vincent, G. M., Chapman, J., & Cook, N. E. (2011). Risk/Needs assessment in juvenile justice: Predictive validity of the SAVRY, racial differences, and contribution of needs factors. *Criminal Justice and Behavior, 38*, 42-62. doi: 10.1177/0093854810386000

Vincent, G. M., Guy, L. S., Fusco, S. L., & Gershenson, B. G. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior*, *36*, 225-236. doi: 10.1037/h0093974

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). Risk assessment in juvenile justice: A guidebook for implementation. Retrieved from http://escholarship.umassmed.edu/psych_cmhsr/573/

Vincent, G. M., Paiva-Salisbury, M. L., Cook, N. E., Guy, L. S., & Perrault, R. T. (2012). Impact of risk/needs assessment on juvenile probation officers' decision making: Importance of implementation. *Psychology, Public Policy, and Law*, *18*, 549-576. doi: 10.1037/a0027186

Vitopoulos, N. A., Peterson-Badali, M., & Skilling, T. A. (2012). The relationship between matching service to criminogenic need and recidivism in male and female youth examining the RNR principles in practice. *Criminal Justice and Behavior*, *39*, 1025-1041. doi: 10.1177/0093854812442895

Wachter, A. (May, 2015). Statewide risk assessment in juvenile probation. JJGPS StateScan. Pittsburgh, PA: National Center for Juvenile Justice.

Wu, S. M., Whiteside, U., & Neighbors, C. (2007). Differences in Inter-Rater Reliability and Accuracy for a Treatment Adherence Scale. *Cognitive Behaviour Therapy*, *36*, 230-239. doi: 10.1080/16506070701584367

Table 1.

CA-YASI Item Description

Subscale	**Number of Items**	**Alpha**	**Definition**
Legal History	7	.62	Early onset, frequent, varied criminal behavior
Correctional Response	11	.65	Noncompliance with rules of institutional or community placement, including misconduct, technical violations, and new offenses
Aggression-Violence	24	.86	Past violent behavior (static risk); anger/hostility, callousness, attitudes supportive of aggression

Social Influences	10	.89	Attachment to antisocial peers, absence of constructive adult role models
Substance Use	8	.52	Frequent alcohol and drug use that can impair functioning
Attitudes	8	.87	Antisocial attitudes, including minimization of responsibility, denial of harm, poor attitudes toward the justice system/authority
Social-Cognitive Skills	6	.93	

Poor decision--making skills (consequential thinking, goal setting, problem-solving) and/or interpersonal skills (perspective taking) relevant to antisocial behavior

Family

10

.75

Poor family relationships or role modeling

Education-Employment

8

.80

Poor educational achievement, employment potential, or motivation

Health

7

.49

Mental health problems

Community Linkages	4	.65	Lack of relevant services to address criminogenic needs in the community
Community Stability	6	.62	Poor finances, accommodation, or transportation

Table 2.

YASI Total- and Subscale-Level Inter-rater Accuracy as Indexed by Intra-class Correlation Coefficients.

	Total Score	A. Legal History	B. Corr. Response	C. Agg./ Vio.	D. Social Influences	E. Substance Use	F. Attitudes	G. Soc./ Cog. Skills	H. Family	I. Edu./ Employ.	K. Comm. Linkages	L. Comm. Stability
Total	.63	.74	.65	.53	.50	.57	.53	.50	.73	.77	.79	.66
Sample	(.36)	(.34)	(.48)	(.41)	(.43)	(.37)	(.49)	(.31)	(.38)	(.27)	(.20)	(.35)
(*n* = 69)												
Site												
Site A	.72	.83	.60	.67	.57	.65	.60	.63	.72	.91	---	.53
(*n* = 13)	(.27)	(.13)	(.49)	(.50)	(.38)	(.23)	(.50)	(.25)	(.34)	(.10)		(.51)
Site B	.51	.71	.77	.33	.41	.52	.45	.33	.75	.61	.78	.78
(*n* = 15)	(.45)	(.40)	(.36)	(.42)	(.42)	(.39)	(.53)	(.34)	(.36)	(.44)	(.19)	(.22)
Site C	.64	.70	.69	.44	.41	.48	.58	.35	.85	.74	.80	.73
(*n* = 25)	(.36)	(.43)	(.40)	(.36)	(.50)	(.42)	(.45)	(.16)	(.14)	(.21)	(.20)	(.30)
Site D	.64	.78	.48	.75	.68	.76	.45	.77	.50	.87	---	.55
(*n* = 16)	(.35)	(.20)	(.65)	(.25)	(.30)	(.32)	(.56)	(.25)	(.59)	(.11)		(.31)
Gender												
Female	.60	.78	.64	.53	.53	.56	.50	.53	.68	.73	.73	.68
(*n* = 39)	(.39)	(.27)	(.47)	(.42)	(.38)	(.42)	(.52)	(.27)	(.47)	(.32)	(.20)	(.39)
Male	.66	.70	.65	.54	.46	.58	.57	.47	.79	.82	.88	.65
(*n* = 30)	(.32)	(.42)	(.50)	(.41)	(.48)	(.30)	(.45)	(.36)	(.20)	(.20)	(.16)	(.30)
Experience												
0-5 years	.75	.77	.66	.61	.36	.46	.72	.60	.70	.82	.72	.69
(*n* = 20)	(.27)	(.34)	(.42)	(.42)	(.44)	(.48)	(.23)	(.25)	(.39)	(.24)	(.24)	(.36)
6-15 years	.50	.71	.57	.54	.46	.52	.51	.55	.76	.72	.93	.63
(*n* = 23)	(.46)	(.39)	(.50)	(.43)	(.49)	(.40)	(.50)	(.31)	(.42)	(.37)	(.09)	(.40)
16-plus	.63	.73	.67	.52	.61	.67	.47	.40	.68	.72	.74	.72
years	(.31)	(.37)	(.52)	(.29)	(.32)	(.23)	(.50)	(.36)	(.38)	(.25)	(.21)	(.28)

(*n* = 21)

Notes: *Mean ICC (SD)*. Values presented outside out the parentheses are mean intra-class correlation coefficients. Values presented within the parentheses are standard deviations of the mean intra-class correlation coefficients.

Table 3.

Percent of Staff with “Good” Reliability for CA-YASI Total and Subscale Scores

Subscale	**Percent of Staff with “Good” Reliability**
Total Score	59.0%
Legal History	82.7%
Correctional Response	74.7%
Aggression-Violence	53.8%
Social Influences	49.3%
Substance Use	61.4%
Attitudes	59.1%
Social-Cognitive Skills	38.9%
Family	77.5%
Education-Employment	84.8%
Community Linkages	62.5%
Community Stability	66.7%

Table 4.

Summary of Moderator Analyses for CA-YASI Total Score Inter-rater Accuracy.

	Baseline	**Model with**	**-2LL Model**	**ICC**
	Model LL	**Nesting LL**	**Difference**	
Site	-29.62	-29.62	0.00	0.00
Gender	-29.62	-29.62	0.00	0.00
Experience	-29.62	-26.62	6.00*	0.06

Note: * *p* < .05. LL = log restricted likelihood. ICC = intra-class correlation coefficient.