WILEY

Taking "the boss" into the real world: Field interrater reliability of the Short-Term Assessment of Risk and Treatability: Adolescent Version

Tamara L. F. De Beuf[1,2] | **Corine de Ruiter**[2] | **John F. Edens**[3] | **Vivienne de Vogel**[4]

[1]Research Department, Ottho Gerhard Heldring Institution, Zetten, Netherlands

[2]Department of Clinical Psychological Science, Maastricht University, Maastricht, Netherlands

[3]Department of Psychological and Brain Sciences, Texas A&M University, College Station, Texas, USA

[4]Research Department, De Forensische Zorgspecialisten, Utrecht, Netherlands

Correspondence
Tamara L. F. De Beuf, P. O. Box 1, 6670 AA Zetten, the Netherlands.
Email: tamara.debeuf@maastrichtuniversity.nl

Abstract

There is emerging evidence that the performance of risk assessment instruments is weaker when used for clinical decision-making than for research purposes. For instance, research has found lower agreement between evaluators when the risk assessments are conducted during routine practice. We examined the field interrater reliability of the Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV). Clinicians in a Dutch secure youth care facility completed START:AV assessments as part of the treatment routine. Consistent with previous literature, interrater reliability of the items and total scores was lower than previously reported in non-field studies. Nevertheless, moderate to good interrater reliability was found for final risk judgments on most adverse outcomes. Field studies provide insights into the actual performance of structured risk assessment in real-world settings, exposing factors that affect reliability. This information is relevant for those who wish to implement structured risk assessment with a level of reliability that is defensible considering the high stakes.

1 | INTRODUCTION

One of the first steps in implementing structured risk assessment is the selection of an appropriate risk assessment instrument. When deciding on the most suitable instrument for a particular setting or jurisdiction, multiple factors should be considered, such as feasibility, possibility of training, and costs (Vincent et al., 2012). The psychometric properties of potential instruments should also guide the decision, and peer-reviewed empirical research on the reliability and validity of an instrument should be evaluated. In regard to risk assessment instruments, reliability generally refers to the consistency of results when the assessment is repeated, and validity addresses whether the instrument's scores accurately predict the occurrence of the outcome of interest. Contrary to what the term "properties" suggests, reliability and validity are not intrinsic features of any risk assessment instrument; they are not static characteristics that automatically translate to other settings, populations, and evaluators [DeMatteo et al., 2020; Edens et al., 2013; more generally, see the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 2014)].

Performance may differ not only by location or setting (e.g., prison vs. community service), but also by the evaluator's assessment goal: data collection for research purposes versus data generated for clinical/legal practices (Edens & Boccaccini, 2017). There is emerging evidence that instruments perform better in more controlled studies (hereafter referred to as "non-field studies") than when administered by practitioners in everyday practice (Neal et al., 2015). In recent years, this awareness has contributed to a growing effort to assess the performance of risk assessment instruments in real-world settings, referred to as field studies (Edens & Boccaccini, 2017).

1.1 | Field studies of risk assessment

Field studies are distinct from non-field studies in that data are generated as part of clinical, forensic, or correctional practice. Thus, field studies are based on an assessment outcome "used as part of a decision-making process that would have real-world implications for the person being evaluated" (Edens & Boccaccini, 2017, p. 600). Assessments conducted by professionals in the field with the sole purpose of research are not considered field studies by this definition. Field studies on reliability may be particularly important, as there is emerging evidence that lower interrater reliability may limit the predictive validity of risk assessment instruments in real-world settings (Edens & Kelley, 2017). For example, in a meta-analysis, Hanson and Morton-Bourgon (2009) found that risk assessment studies with lower interrater reliability had significantly lower effect sizes for recidivism than studies with higher reliability. Low reliability indicates measurement error that is either nonsystematic (i.e., random) or systematic (e.g., rater bias); the more error, the more difficult it is to make valid predictions (Boccaccini et al., 2009; Neal et al., 2015). For this reason, Edens and Kelley (2017) labeled reliability, and particularly interrater reliability, as "the boss" of psychometric qualities in the field.

1.2 | Field interrater reliability

Field reliability is often assessed in terms of interrater reliability, that is, the consistency of findings when different evaluators assess the same subject. Establishing interrater reliability is relevant to ensure the fairness of decisions that result from the outcome of the risk assessment. It can be understood as a quality check for those who decide to implement the instrument within their practice. Until now, the majority of studies on field interrater reliability in forensic assessment involved actuarial risk assessment instruments for (sexual) violence (Boccaccini et al., 2009; Edens et al., 2016; Miller et al., 2012; Rettenberger et al., 2017; Schmidt et al., 2005) and the Psychopathy Checklist-Revised (PCL-R; Hare, 2003), an instrument to assess psychopathy, also commonly used to aid in risk assessment (Edens et al., 2019). These studies have typically been conducted on test results that were collected

specifically to inform case decision-making as part of some adversarial judicial proceeding (e.g., criminal trial or sexually violent predator civil commitment hearing). However, there is emerging literature on the field interrater reliability of structured professional judgment (SPJ) risk assessment instruments that are frequently implemented in post-adjudication forensic psychiatric settings (de Vogel & de Ruiter, 2004; Jeandarme et al., 2017; Nicholls et al., 2006; Troquete et al., 2015). Findings suggest that risk assessments completed by practitioners as part of routine decision-making (e.g., leave status, discharge) – whether using actuarial instruments or SPJ approaches – obtain weaker interrater reliability than assessments in controlled non-field studies (e.g., Jeandarme et al., 2017). However, some measures seem to be less affected (e.g., Vincent et al., 2012).

Authors have presented various arguments for this decrement in interrater reliability. They argue that studies in the field are conducted in less-than-ideal conditions, whereas in non-field studies, conditions can be better controlled. For example, evaluators in field studies are more likely to show wider variability in terms of experience, training, and/ or professional background, compared with research assistants in non-field studies (Boccaccini et al., 2008; de Vogel & de Ruiter, 2004). They are also less likely than research assistants to have received consistent and extensive training, and their risk assessment performance is usually not as closely monitored or supervised (DeMatteo et al., 2020; Guarnera & Murrie, 2017). Even when the evaluators in a non-field study are practitioners from the field, awareness that they are participating in an empirical study or that researchers are involved, could improve the evaluators' usual performance (Penney et al., 2014; Vincent et al., 2012). When individuals know they are being evaluated and their errors are being monitored, they will perform differently (Dror, 2020). Additionally, risk assessments conducted for clinical purposes might be influenced by contextual pressures, such as political or organizational demands (e.g., to facilitate a transfer of the patient; Jeandarme et al., 2017) and emotional biases (e.g., positive vs. negative emotional involvement with the patient; de Vogel & de Ruiter, 2004). Furthermore, in contrast to non-field studies in which evaluators usually receive the same information package or case vignette, clinicians in field studies are responsible for gathering their own information (Boccaccini et al., 2008). In practice, information might not only be qualitatively and quantitatively limited compared with more controlled studies, but the information used for the assessment might also differ per evaluator (DeMatteo et al., 2020; Penney et al., 2014). Similarly, it is possible that interviews conducted for research purposes versus clinical-legal purposes will elicit a different degree of honesty, disclosure, and impression management from the person evaluated, beyond the effect of an evaluator's interview style (Edens & Kelley, 2017; Guarnera & Murrie, 2017).

In addition to these evaluator and contextual characteristics, features of the risk assessment instrument itself can also impact reliability. For example, item subjectivity and a lack of clear instructions potentially increase noise in the ratings and may limit agreement between evaluators (Edens & Kelley, 2017). As such, actuarial risk assessment instruments that mainly consist of static factors targeting quantifiable behaviors (e.g., the number of prior charges) are more likely to be completed reliably than instruments that include interpersonal and affective features that require a certain degree of clinical judgment (e.g., self-esteem, empathy; Edens et al., 2019). Indeed et al. (2011) found moderate to large inverse correlations between the perceived subjectivity of risk factors and interrater agreement on these factors, suggesting that at least some variation in reliability could be explained by item subjectivity. Similarly, Kennealy et al. (2016) found lower interrater agreement for features that required more clinical judgment (e.g., attitudes) compared with more straightforward features (e.g., criminal history). Given the more central role of evaluator judgment in the use of SPJ risk assessment instruments, these instruments might be particularly prone to the above-mentioned influences when applied in the field, resulting in lower field interrater reliability compared with actuarial instruments (Edens & Kelley, 2017).

1.3 | Field reliability of the START:AV

A fairly recently developed SPJ instrument for adolescents is the Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV; Viljoen et al., 2014). This instrument assesses the short-term risk of eight adverse

outcomes, such as violence to others and victimization (see the Method section). Evaluators reach a final risk judgment (low, moderate, or high) by weighing historical information against the adolescent's current presentation on items related to characteristics of the adolescent (e.g., impulsivity), their relationships (e.g., parenting), and their response to treatment (e.g., insight). The START:AV items are combined into overall ratings of "strengths" and "vulnerability," and total scores for both can be calculated for research purposes. With half a dozen publications, research on the psychometric properties and utility of the START:AV is still in its infancy (Desmarais et al., 2012; Klimukienė et al., 2018; Sellers et al., 2017; Sher et al., 2017; Singh et al., 2014; Viljoen et al., 2012). Moreover, most of these studies have used the pilot version of the instrument. To our knowledge, there are no START:AV field studies in the literature as yet. In the following, we describe the results of two non-field studies that have examined the interrater reliability of the START:AV.

The first interrater reliability examination was conducted by the instrument's authors and involved 12 START: AV assessments of adolescent offenders on probation completed by graduate students (Viljoen et al., 2012). The students received training on the interview process and the administration of the START:AV and completed multiple practice cases, including an in-person practice assessment alongside an experienced rater. Before conducting START:AV assessments for data collection, the students had to demonstrate interrater reliability consistent with a consensus or gold standard rating, that is, the ratings agreed by a team of START:AV experts. The student evaluators based their assessments on the youth's justice records and an interview with the adolescent. Intraclass correlation coefficients (ICCs; two-way random-effects model, absolute agreement, single measure) for the total scores were 86 for vulnerabilities and 92 for strengths. ICCs for six of the START:AV's final risk judgments were as follows: 52 for suicide, 60 for violence to others and substance abuse, 75 for general reoffending, 86 for victimization and 88 for self-injury.

More recently, the START:AV was evaluated in a Lithuanian juvenile probation sample (Klimukienė et al., 2018). Similar to the study by Viljoen et al. (2012), this was a non-field study in which the START:AV assessments were conducted by a research team (not further specified). The research team practiced with interviews gathered during a pilot study until they reached agreement on the pilot START:AV assessments. For the actual interrater reliability study, the team randomly selected 30 interviews that involved a team member questioning a juvenile probation officer about an adolescent under their supervision. These audio-taped interviews were subsequently used as the only source of information for the risk assessments. ICCs (two-way random-effects model, absolute agreement, single measure) for the total scores were 91 for vulnerabilities and 82 for strengths. Additionally, ICCs for four of START:AV's final risk judgments were calculated: 49 for substance abuse, 51 for violence to others, 66 for general reoffending, and 89 for unauthorized absences. Klimukienė et al. (2018) also analyzed the interrater reliability of the items. The ICCs for vulnerabilities ranged from 29 to 86 (average ICC = 67), with lowest agreement for external triggers (ICC = 29), parental functioning (ICC = 37), and social support from peers (ICC = 39). For the strengths items, ICCs ranged from 36 to 80 (average ICC = 58), with lowest agreement for coping (ICC = 36), social support from adults (ICC = 39), and community (ICC = 39). The authors argued that the low agreement for these items was likely due to the fact that the interview with the juvenile probation officer was the only source of information.

1.4 | Present study

Beyond these two non-field studies with juvenile probation samples, there have been no studies investigating the interrater reliability of the START:AV, nor have there been field studies on this issue. As underlined by Webster et al. (2002) almost two decades ago and recently reiterated by Edens and Boccaccini (2017), there is a need for risk assessment studies using data from clinicians during their everyday work. The present study addresses the need for field studies in relation to the START:AV. This instrument was implemented in a Dutch secure residential youth care facility where the risk assessments were used for decision-making regarding level of supervision and treatment objectives. In this study, we examine "the boss" of psychometric qualities (i.e., interrater reliability) and in line with

the emerging field research, we expect to find lower interrater reliability for the START:AV features than previously reported in the non-field studies.

2 | METHOD

2.1 | Setting

The study was conducted in one of 14 secure residential youth care facilities in the Netherlands. Secure residential services provide the most intensive type of youth care and are often considered a "last resort" for boys and girls with severe behavioral and/or mental health problems (Ten Brummelaar et al., 2017). Adolescents can only be admitted to residential youth care by a court order. This setting is distinct from juvenile detention centers in that youth are admitted under civil law rather than criminal law. Mandated youth care occurs under this program when deemed necessary by a judge to guarantee the adolescents' safety (e.g., from self-harm, neglect) and/or the safety of their environment (e.g., violence to others, criminal activities). At the present location, 137 adolescents were admitted in 2016 and 225 adolescents (52% girls) were in treatment over the course of that year. The average treatment duration was 262 days (8.6 months, range: 4–717 days). At the time of the study, the service had three high secure units (including two observation units) and six medium secure units.

2.2 | Youth sample

The START:AV forms ($n = 100$) were randomly selected from all START:AVs completed between 1 March 2016 and 2 November 2017 that had fewer than six missing strength or vulnerability ratings ($N = 237$). The START:AV assessments concerned 82 unique adolescents; several (20%) had multiple START:AV forms.

The sample consisted of 41 male and 41 female adolescents (in line with the setting's boy:girl ratio) between the ages of 12.4 and 17.9 years at the time of the START:AV assessment ($M = 16.3$, SD = 1.3). They spent, on average, 300 days (9.9 months) in medium and/or high secure care (SD = 131.17; range = 66–664). Forty-five percent resided in a high secure ward (24% observation, 21% treatment), while the others resided in a medium secure treatment unit. The sample's average total IQ score, measured with the Wechsler Intelligence Scale for Children-V-NL (WISC-V-NL; Wechsler, 2014/2017), was 87.7 (SD = 14.38) and ranged from 54 to 119. For 28% of the adolescents, no total IQ score could be calculated because of a disharmonic intelligence profile (i.e., 1 SD or more discrepancy between verbal and performance IQ score). All youths had at least one diagnosis according to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; American Psychiatric Association, 2013). The most common diagnoses were oppositional defiant disorder ($n = 22$, 33%), conduct disorder ($n = 26$, 32%), attention deficit hyperactivity disorder ($n = 25$, 31%), as well as post-traumatic stress disorder ($n = 19$, 23%), substance-related disorders ($n = 17$, 21%), intellectual disability ($n = 16$, 20%), and autism spectrum disorder ($n = 11$, 13.4%). A majority of the adolescents had three ($n = 24$, 30%) or four diagnoses ($n = 26$, 32%), and 17% ($n = 14$) had more than four diagnoses. Non-selected adolescents did not differ significantly from the included sample concerning these characteristics.

2.3 | Evaluators

2.3.1 | Clinical evaluators

The START:AV forms were completed as part of routine practice by 12 evaluators who were employed as treatment coordinators within the service. These professionals, with at least a master's degree in psychology or special

needs education, were responsible for the adolescents' treatment process. In consultation with the team on the unit, they made decisions about treatment (e.g., treatment goals, types of therapy) and evaluated treatment progress. All evaluators were women, aged 26–40 years ($M = 32.5$, SD = 4.4), with on average 6 years of service as a treatment coordinator within the organization (SD = 4.8, range: 0–14). Prior to the START:AV training, 50% had previously used risk assessment tools in practice and 60% had received some sort of risk assessment training. All evaluators were trained by the first author who, in turn, had received training and supervision from the instrument's developers. Nine evaluators participated in a 2-day group training event and discussed additional practice cases during another 2-hour training session led by the first and third authors. Three evaluators, hired after the initial training, received one-on-one training shortly after starting employment. In the present sample, the evaluators completed, on average, eight START:AVs (SD = 6), ranging from one to 21 START:AVs per evaluator.

2.3.2 | Research evaluators

Four female graduate-level students in psychology conducted independent ratings on the START:AV as research assistants. They all had completed coursework on forensic populations and adolescent development and had experience with young people in residential youth care. One research evaluator (first author) was trained by the original authors of the START:AV, two evaluators participated in a 1-day group training event provided by the first author followed by one-on-one feedback sessions discussing practice cases, and the fourth research rater was trained one-on-one by the first author, including one-on-one feedback on multiple practice cases.

2.4 | Measure

The START:AV is an SPJ risk assessment instrument assessing the short-term risk of multiple adverse outcomes, including violence to others, nonviolent offending, substance abuse, unauthorized absences, suicide, self-injury, victimization, and health neglect. Prior to formulating a final risk judgment (low, moderate, or high), the evaluator rates 24 items (low, moderate, or high) based on information of the previous months and assesses the occurrence of adverse outcomes in the recent and prior history (present/absent). All START:AV items are dynamic items that can change over time due to normal development or in response to life events and interventions (Viljoen et al., 2012). The items involve behavioral, cognitive, and interpersonal characteristics of the adolescent and how they respond to interventions, as well as features of their relationships and environment (e.g., "social support," "peers"). Items are assessed in two directions: as strengths (i.e., positive features that may reduce risk) and as vulnerabilities (i.e., difficulties that may increase risk). The START:AV user guide (Viljoen et al., 2014, 2016) provides a definition for each item as well as indicators (i.e., item anchors) for both strengths and vulnerabilities. In addition to the item anchors, the user guide provides examples of moderate versus high strengths/vulnerabilities. After rating the items, the evaluator identifies which strengths and vulnerabilities are particularly relevant to future risks. Key strengths may be important in protecting against adverse outcomes, whereas critical vulnerabilities could play a crucial role in contributing to adverse outcomes. The key and critical items thus identified are building blocks for the risk formulation and the intervention plan.

2.5 | Implementation

In February 2016, the service introduced the START:AV as the risk assessment instrument within its practice (see also De Beuf et al., 2020). The START:AV was embedded within the treatment cycle: every 4 months, the START:AV was completed again as part of the periodic treatment plan evaluation. At first, comprehensive rating forms were

completed and coded by treatment coordinators. Starting in mid-2017, group care workers and other care professionals (e.g., therapists, family social workers) also started reporting their observations in the START:AV form, but the treatment coordinators remained responsible for rating the items and adverse outcomes. Various sources of information were used to complete the risk assessments: file review of internal (e.g., daily progress notes, previous START:AVs) and external (e.g., supervision order, treatment reports from previous service agencies) documentation, as well as interviews with the adolescent, their caregivers and/or other collateral informants. In addition, treatment coordinators had the opportunity to observe the adolescent on the unit.

2.6 | Procedure

The clinician-rated START:AV forms were gathered from the patient files with permission from the service's general director and approval from the Ethics Review Committee Psychology and Neuroscience (ERCPN) of Maastricht University (ERCPN number 174_05_12_2016). As described earlier, the clinician-rated forms were completed as part of routine clinical practice. The second rating of each START:AV was completed by a research evaluator; it was never the case that two clinicians rated the same case for clinical purposes. Therefore, research assistants (see section on Research evaluators) conducted the second assessments for 100 randomly selected START:AVs. Although the research evaluators completed the assessment at another time point compared with the clinical evaluators, they relied on the same file information (i.e., internal and external documentation) that was available to the clinical evaluators at the time of their assessment.

2.7 | Statistical analysis

Before assessing interrater reliability, we explored several features of the sample, including missing ratings and the distribution of scores per evaluator group. When there were missing items, total scores were prorated by dividing the raw total score by the maximum total score, and multiplying this by 52, because the maximum total score for 26 items (items 13 and 14 each have two separate ratings), each with a maximum score of 2, is 52. When item 23 (medication adherence) was rated as "not applicable," the item was handled as a missing item. The prorating formula differs from previous studies with the START:AV (Desmarais et al., 2012; Klimukienė et al., 2018), but is consistent with other risk assessment instruments (Boer et al., 1997; Webster et al., 1997). We tested the above formula and the formula applied in previous START:AV studies on a subsample of complete START:AVs in which we artificially created missing ratings. We found that the prorating formula of our choice best approximated the true values in the sample.

We calculated interrater reliability of prior and recent history of adverse outcomes using Gwet's agreement coefficient (AC1; Gwet, 2002). This statistic is preferred to a kappa coefficient because Gwet's AC is not affected by prevalence rates, whereas kappa tends to underestimate agreement in situations with high or low prevalence (e.g., prevalence of suicide), a phenomenon referred to as the "kappa paradox" or the "paradox of high agreement, low reliability" (Feinstein & Cicchetti, 1990). Gwet's AC1 was calculated using the AgreeStat360 Excel program (Gwet, 2020).

For the interrater reliability of the individual items, the strength and vulnerability total scores, and the eight final risk judgments, the ICC was calculated. More specifically, a single-measures, two-way random model was selected, using both the consistency and absolute definition (Koo & Li, 2016). A two-way random ICC was deemed appropriate because the forms were completed by evaluators from the same set of evaluators who represented a larger group of evaluators of interest. A single-measures type was most relevant to the current risk assessment practice as every assessment was conducted by a single evaluator ("single measures" type) rather than using the average score of multiple evaluators ("average measures" type). For the total scores, we were interested in the

"consistency" definition rather than absolute agreement, because total scores of the START:AV are not used in an absolute manner (i.e., as a cutoff); in fact, in clinical practice they are not used at all. For the item ratings and final risk judgments (both rated as low, moderate, or high), we calculated absolute agreement ICCs. For all ICCs values, the 95% confidence intervals (CIs) were determined. Koo and Li (2016) argue that it is more appropriate to interpret the CIs rather than the ICC point estimates. CIs give an indication of the range in which the true ICC value lands, whereas the ICC point estimate is only an expected value of the true ICC.

We interpreted the CIs of the Gwet's AC1 and the single-rater two-way random ICCs according to Koo and Li's (2016) guidelines: ICC < 0.50, poor; 0.50–0.75, moderate; 0.75–0.90, good; > 0.90, excellent. These interpretations are stricter than the commonly used guidelines by Landis and Koch (1977), Cicchetti and Sparrow (1981), or Fleiss (1986). It has been repeatedly argued that these classic guidelines might be too lenient for applied settings, where important decisions are made based on risk assessments (Edens & Boccaccini, 2017; Levenson, 2004). For example, the threshold of 0.80 presented by Heilbrun (1992) has been put forward as a favorable, although difficult to achieve, benchmark for field reliability in settings where the stakes are high (Edens et al., 2019; Guarnera & Murrie, 2017).

3 | RESULTS

Forty-six of the 100 START:AVs were completed upon admission to the service, on average, after 68 days (SD = 37, range: 0–216). The remaining 54 START:AVs in the sample were follow-up START:AVs, completed for the purpose of treatment evaluation.

3.1 | Descriptive statistics

As presented in Table 1, the evaluator groups (researchers vs. clinicians) differed significantly on the average total score for strengths [$t(99)$ = 5.50, $p < 0.001$, d = 0.58] and vulnerabilities [$t(99)$ = −7.34, $p < 0.001$, d = −0.71]. Research evaluators assigned lower scores for strengths (M = 14.8, SD = 5.9; clinician evaluators: M = 19.2, SD = 8.8) and higher scores for vulnerabilities (M = 36.1, SD = 7.1; clinician evaluators: M = 31.4, SD = 6.3). Across the entire sample of juveniles, the researcher-rated total scores for strengths ranged from 3.69 to 33.58, and those for vulnerabilities ranged from 21.67 to 48.45, while the clinician-rated total scores ranged from 5.00 to 45.76 for strengths and 16.22 to 47.00 for vulnerabilities. The average absolute difference in total scores between the research and clinical evaluators was 7.5 (SD = 5.4) for strengths, with the largest absolute difference being 27 points and three cases reaching perfect agreement. With regard to vulnerabilities, the average absolute difference was 6.6 (SD = 4.5), with the largest absolute difference being 18 points and four cases reaching perfect agreement. Lastly, Table 2 presents the prevalence of adverse outcomes in the prior and recent history of the adolescents as well as the distribution of risk estimates per outcome, as assessed by both evaluator groups.

3.2 | Missing ratings

Table 1 also presents the percentage of missing ratings per item for both evaluator groups. Because the research evaluators relied solely on file information, they had significantly more missing items than the clinical evaluators, both for strengths [$t(99)$ = −5.77, $p < 0.001$, d = −0.76] and for vulnerabilities [$t(99)$ = −7.13, $p < 0.001$, d = −0.96]. The clinical evaluators had a high percentage of missing ratings on item 20, "external triggers," whereas the research evaluators had multiple items with a relatively large number of missing ratings, especially items related to the adolescent's social context. For example, there were more than 10% missing ratings on both strengths and vulnerabilities for items 14a and 14b (social support from adults and peers), item 16 (parental functioning), and item

TABLE 1 Distribution of item ratings and missing ratings for strengths and vulnerabilities for both evaluator groups, presented in percentages

		Clinical evaluators				Research evaluators			
		Strength		Vulnerability		Strength		Vulnerability	
	START:AV item	Low/medium/high (L/M/H)	Missing	L/M/H	Missing	L/M/H	Missing	L/M/H	Missing
1.	School and work	26/51/23	0	11/67/22	0	32/45/16	7	9/43/41	7
2.	Recreation	29/57/13	1	12/67/20	1	26/61/9	4	24/43/28	5
3.	Substance use	49/39/12	0	20/46/34	0	45/25/28	2	40/27/33	0
4.	Rule adherence	27/58/13	2	13/43/43	1	38/51/11	0	9/35/56	0
5.	Conduct	24/71/5	0	4/48/47	1	39/60/1	0	2/29/69	0
6.	Self-care	22/62/15	1	10/67/21	2	46/33/6	15	10/49/29	12
7.	Coping	62/37/0	1	7/21/78	0	75/24/0	1	1/13/85	1
8.	Impulse control	56/36/5	3	13/56/31	0	76/17/1	6	9/30/56	5
9.	Mental state	45/46/8	1	12/54/34	0	62/30/3	5	8/34/55	3
10.	Emotional state	47/49/1	3	2/29/68	1	82/15/0	3	3/17/78	2
11.	Attitudes	35/52/7	6	9/47/41	3	68/17/0	15	2/31/55	12
12.	Social skills	18/65/17	0	17/63/19	1	33/54/8	5	14/38/42	6
13.	Relationships adults	27/51/20	2	12/45/41	2	23/60/12	5	9/38/49	4
13b.	Relationships peers	43/47/6	4	12/53/35	0	51/41/3	5	9/30/55	6
14.	Social support adults	41/43/15	1	21/48/29	2	34/41/9	16	13/22/49	16
14b.	Social support peers	60/26/6	8	3/49/44	4	61/17/3	19	5/12/62	21
15.	Parenting	26/55/15	4	10/41/47	2	52/31/8	9	5/29/59	7
16.	Parental functioning	34/49/15	2	15/49/29	7	39/38/8	15	9/30/41	20
17.	Peers	74/18/3	5	9/45/42	4	83/5/1	11	2/23/63	12
18.	Material resources	7/79/13	1	24/72/3	1	7/82/7	4	8/84/4	4
19.	Community	7/84/5	4	8/85/3	4	1/96/1	2	6/89/2	3
20.	External triggers	40/28/4	28	25/37/20	18	52/33/2	13	16/44/33	7

(Continues)

T A B L E 1 (Continued)

	Clinical evaluators				Research evaluators			
	Strength		Vulnerability		Strength		Vulnerability	
START:AV item	Low/medium/high (L/M/H)	Missing	L/M/H	Missing	L/M/H	Missing	L/M/H	Missing
21. Insight	43/39/13	5	11/42/44	3	63/30/5	2	3/33/62	2
22. Plans	30/49/16	5	20/57/18	5	42/33/12	13	18/37/29	16
23. Medication adherence[a]	11/20/7/56	6	12/25/3/56	4	16/16/10/54	4	21/14/6/54	5
24. Treatability	28/51/16	5	15/59/22	4	41/44/10	5	16/29/50	5
Total scores[b]	19.2 (8.8)	–	31.4 (6.3)	–	14.8 (5.9)	–	36.1 (7.1)	–

[a]Medication adherence as an additional fourth category for "not applicable" ratings.

[b]Descriptives for total scores are presented as mean (standard deviation).

17 (peers). In addition, items such as self-care (item 6), attitudes (item 11), external triggers (item 20) and plans (item 22) were among the items with the greatest number of missing ratings. With regard to the adverse outcomes, the clinical evaluators had especially high missing ratings for "health neglect" (see Table 2).

3.3 | Intraclass correlation coefficients[1]

Two-way random absolute ICCs for the items ranged from −0.01 (external triggers) to 0.67 (parental functioning) for the strength items, and from 0.19 (peers; treatability) to 0.70 (parental functioning) for the vulnerability items (Table 3). As discussed in the Method section, we relied on the 95% CIs for interpretation because these present a more accurate picture of the true reliability scores. Almost half of the strength items ($n = 12$) fell within the poor reliability range and the other half ($n = 13$) fell within the poor to moderate range. Only for the item "parental functioning" did the evaluators reach poor to good reliability. The item-level mean ICC for strengths was 0.31 (range: −0.01–0.67). For the vulnerability items, a similar pattern was observed: 11 items demonstrated poor reliability, 14 items poor to moderate, and again reliability of parental functioning ranged from poor to good. The item-level mean ICC for vulnerabilities was 0.38 (range: 0.19–0.70). For the total scores, the two-way random consistency ICC was 0.42 [95% CI 0.24–0.57] for strengths and 0.54 [0.38–0.66] for vulnerabilities, both demonstrating poor to moderate reliability (Table 3).

As shown in Table 4, the two-way random absolute ICCs for the final risk judgments ranged from 0.57 (health neglect) to 0.84 (suicide). For half of the risk judgments, including nonviolent offenses, substance abuse, self-injury, and victimization, the evaluators reached moderate to good agreement. For unauthorized absences and health neglect, poor to moderate interrater reliability was found, for violence moderate reliability, and for suicide good reliability. We found category errors for all adverse outcomes, except suicide. A category error occurs when one evaluator rates the case as high risk, whereas the other evaluator rates the same case as low risk. More specifically, there were six category errors for violence, three for nonviolent offending, three for substance abuse, three for unauthorized absences, one for self-injury, four for victimization, and one for health neglect. In half (10) of the major disagreements, the clinical evaluators assigned low risk, whereas the research evaluators estimated high risk; in the other half (11), it was the opposite. Exploratory analysis revealed no significant differences between evaluators in the prevalence of category errors.

TABLE 2 Percentage of adverse outcomes rated as present in prior and recent history as well as distribution of risk ratings for both evaluator groups

Adverse outcomes	Clinical evaluators						Research evaluators					
	Prior history		Recent history		Risk judgment		Prior history		Recent history		Risk judgment	
	Present	Missing	Present	Missing	Low/medium/high (L/M/H)	Missing	Present	Missing	Present	Missing	L/M/H	Missing
1. Violence	76	0	37	0	39/31/29	1	74	0	54	0	34/30/35	1
2. Nonviolent offending	62	1	34	1	32/34/34	0	66	0	36	0	37/40/23	0
3. Substance abuse	74	1	58	2	25/28/45	2	71	0	54	0	28/26/46	0
4. Unauthorized absences	88	2	61	3	9/43/47	1	89	0	71	0	8/40/52	0
5. Suicide	24	1	11	1	84/11/4	1	22	0	7	0	83/13/4	0
6. NS self-injury	39	1	28	1	66/18/14	2	42	0	34	0	68/18/14	0
7. Victimization	82	1	30	1	30/35/34	1	80	0	45	0	24/26/50	0
8. Health neglect	67	6	63	7	26/52/18	4	71	0	77	0	21/57/21	1

Note: NS, non-suicidal.

TABLE 3 Interrater reliability of Short-Term Assessment of Risk and Treatability: Adolescent (START:AV) items and total scores for strengths and vulnerabilities, including interpretation of confidence intervals (CIs)

		Strength				Vulnerability			
START:AV feature		ICC	95% CI	CI interpretation	n	ICC	95% CI	CI interpretation	n
1.	School and work	0.45	[0.28–0.60]	Poor to moderate	93	0.32	[0.13–0.49]	Poor	93
2.	Recreation	0.19	[−0.02–0.38]	Poor	95	0.31	[0.11–0.48]	Poor	94
3.	Substance use	0.38	[0.20–0.53]	Poor to moderate	98	0.53	[0.37–0.66]	Poor to moderate	100
4.	Rule adherence	0.44	[0.27–0.59]	Poor to moderate	98	0.37	[0.19–0.53]	Poor to moderate	99
5.	Conduct	0.14	[−0.05–0.32]	Poor	100	0.33	[0.14–0.49]	Poor	99
6.	Self-care	0.29	[0.08–0.47]	Poor	84	0.44	[0.25–0.59]	Poor to moderate	87
7.	Coping	0.33	[0.14–0.49]	Poor	98	0.43	[0.26–0.58]	Poor to moderate	99
8.	Impulse control	0.08	[−0.10–0.27]	Poor	92	0.44	[0.24–0.60]	Poor to moderate	95
9.	Mental/Cognitive state	0.42	[0.22–0.58]	Poor to moderate	94	0.41	[0.22–0.56]	Poor to moderate	97
10.	Emotional state	0.19	[−0.01–0.38]	Poor	94	0.30	[0.11–0.47]	Poor	97
11.	Attitudes	0.12	[−0.06–0.30]	Poor	80	0.28	[0.08–0.46]	Poor	85
12.	Social skills	0.37	[0.18–0.53]	Poor to moderate	95	0.44	[0.24–0.60]	Poor to moderate	93
13	Relationships adults	0.46	[0.29–0.61]	Poor to moderate	93	0.47	[0.30–0.62]	Poor to moderate	94
13b.	Relationships peers	0.42	[0.24–0.58]	Poor to moderate	91	0.28	[0.09–0.45]	Poor	94
14.	Social support adults	0.34	[0.14–0.52]	Poor to moderate	83	0.34	[0.14–0.52]	Poor to moderate	82
14b.	Social support peers	0.44	[0.24–0.60]	Poor to moderate	74	0.31	[0.09–0.50]	Poor to moderate	75
15.	Parenting	0.39	[0.16–0.57]	Poor to moderate	87	0.52	[0.35–0.66]	Poor to moderate	91
16.	Parental functioning	0.67	[0.52–0.78]	Poor–good	83	0.70	[0.55–0.81]	Poor to good	76
17.	Peers	0.46	[0.27–0.62]	Poor to moderate	84	0.19	[0.00–0.38]	Poor	84
18.	Material resources	0.12	[−0.08–0.31]	Poor	96	0.29	[0.10–0.46]	Poor	96
19.	Community	0.00	[−0.20–0.20]	Poor	95	0.30	[0.11–0.48]	Poor	94
20.	External triggers	−0.01	[−0.26–0.25]	Poor	62	0.27	[0.05–0.46]	Poor	76
21.	Insight	0.41	[0.22–0.57]	Poor to moderate	94	0.33	[0.14–0.50]	Poor to moderate	95
22.	Plans	0.30	[0.09–0.48]	Poor	83	50	[0.31–0.65]	Poor to moderate	79
23.	Medication adherence[a]	0.23	[−0.13–0.53]	Poor to moderate	32	0.45	[0.13–0.68]	Poor to moderate	31
24.	Treatability	0.31	[0.12–0.48]	Poor	90	0.19	[0.00–0.37]	Poor	91
Total score		0.42	[0.24–0.57]	Poor to moderate	100	0.54	[0.38–0.66]	Poor to moderate	100

Note: ICC, intraclass coefficient.

[a]Only assessed for low, moderate or high medication adherence, excluding the "not applicable" category.

3.4 | Gwet's AC1

Gwet's AC1 values were calculated for prior and recent history of adverse outcomes (see Table 4). Values ranged from 0.55 (health neglect) to 0.89 (suicide) for prior history, with wide CIs covering all levels of interrater reliability,

TABLE 4 Interrater reliability of history ratings and risk judgments of the adverse outcomes with an interpretation of confidence intervals (CIs)

Adverse outcome	Prior history		Recent history				Risk judgment		
	Gwet's AC1 (95% CI)	CI interpretation	n	Gwet's AC1 (95% CI)	CI interpretation	n	ICC (95% CI)	CI interpretation	n
1. Violence	0.58 [0.42–0.75]	Poor–good	100	0.50 [0.33–0.68]	Poor to moderate	100	0.64 [0.51–0.74]	Moderate	98
2. Nonviolent offending	0.57 [0.40–0.74]	Poor to moderate	99	0.68 [0.53–0.83]	Moderate to good	99	0.66 [0.52–0.76]	Moderate to good	100
3. Substance abuse	0.77 [0.64–0.89]	Moderate to good	99	0.64 [0.48–0.67]	Poor to moderate	98	0.75 [0.65–0.83]	Moderate to good	98
4. Unauthorized absences	0.88 [0.79–0.96]	Good to Excellent	98	0.72 [0.59–0.86]	Moderate to good	97	0.39 [0.21–0.55]	Poor to moderate	99
5. Suicide	0.89 [0.81–0.97]	Good to Excellent	99	0.90 [0.83–0.97]	Good to Excellent	99	0.84 [0.76–0.89]	Good	99
6. Nonsuicidal self-injury	0.78 [0.66–0.91]	Moderate to excellent	99	0.72 [0.58–0.86]	Moderate to good	99	0.79 [0.70–0.86]	Moderate to good	98
7. Victimization	0.72 [0.59–0.86]	Moderate to good	99	0.37 [0.18–0.56]	Poor to moderate	99	0.65 [0.50–0.76]	Moderate to good	99
8. Health neglect	0.55 [0.38–0.73]	Poor to moderate	94	0.67 [0.51–0.82]	Moderate to good	93	0.57 [0.41–0.69]	Poor to moderate	95

Note: AC, agreement coefficient.

from poor to excellent. For recent history, Gwet's AC1 ranged from 0.37 (victimization) to 0.90 (suicide), again with wide CIs. The level of interrater reliability was good to excellent for suicide, and moderate to good for most other adverse outcomes (nonviolent offending, unauthorized absences, self-injury, and health neglect). Recent history of violence, substance abuse, and victimization demonstrated poor to moderate reliability.

We also explored the number of disagreements between the evaluators for both prior (i.e., more than 4 months prior to the risk assessment) and recent history ratings, that is, how often one evaluator identified an adverse outcome in the past, while the other evaluator did not. For prior history, this occurred equally frequently. For example, in 14 cases, clinical evaluators identified violent incidents in prior history, whereas the researchers did not. Likewise, in 12 cases, research evaluators identified violent incidents in prior history, whereas the clinicians did not. For recent history, the research evaluators identified more incidents than clinical evaluators, especially for the recent history of violence (i.e., 21 identified by researchers but not identified by clinicians vs. four identified by clinicians but not identified by researchers), unauthorized absences (15 vs. 5), victimization (24 vs. 9), and health neglect (15 vs. 4).

4 | DISCUSSION

To our knowledge, this study was the first to assess the interrater reliability of the START:AV completed by professionals in a secure youth forensic mental health setting as part of their day-to-day responsibilities. Consistent with emerging research on field reliability, we found lower interrater reliability than studies that examined the START:AV for research purposes only (Klimukienė et al., 2018; Viljoen et al., 2012). We first discuss our findings in light of previous research, and then describe factors that could have contributed to these lower interrater reliability findings.

4.1 | Field interrater reliability of the START:AV

4.1.1 | START:AV total scores

In line with the literature, we assessed the interrater reliability of the total scores for vulnerabilities and strengths. However, by doing so we treated an SPJ instrument actuarially, which is inconsistent with its intended use in practice (Neal et al., 2019). Therefore, our findings related to SPJ total scores are not generalizable or relevant to typical clinical practice. Nevertheless, we have included the total scores in our reliability analyses to allow for comparison with previous studies. The ICCs for both total scores were weak, with the ICC for the vulnerability total score falling barely within the moderate range and the strength total score demonstrating poor interrater reliability, according to the more stringent interpretation guidelines of Koo and Li (2016). By comparison, the total score ICCs found in the START:AV non-field studies (Klimukienė et al., 2018; Viljoen et al., 2012) fell within the good and excellent range according to the Koo and Li guidelines.

The findings in the present study are more comparable to those reported in a field study (Troquete et al., 2015) of the adult version of the risk assessment instrument, the START (Webster et al., 2009). Like us, Troquete and colleagues found a moderate interrater reliability for the vulnerability total score (ICC = 0.64) and poor interrater reliability for the strength total score (ICC = 0.49)[2]. Other studies also corroborate our finding that reliability is lower for strengths than for vulnerabilities. For example, in a Norwegian field study with the adult START, Nonstad et al. (2010) reported that the interrater reliabilities for risks were higher than for strengths. The authors speculated that practitioners in secure settings might be more familiar with the vulnerabilities of patients than with their strengths.

4.1.2 | START:AV items

When considering the mean item ICCs, the trend described above was also evident for the individual strength and vulnerability items in our study, despite some strengths reaching higher ICCs than their vulnerability counterparts. This is in line with Klimukienė et al. (2018), who found lower interrater reliability for strengths than for vulnerabilities for the majority of the items. Nevertheless, the ICCs for the items in the present study were considerably lower than reported in the latter study.

4.1.3 | START:AV adverse outcomes

Central to SPJ instruments are the final risk judgments – that is, judgments about whether a person is low, moderate or high risk with respect to violence to others, nonviolent offending, substance abuse, unauthorized absences, suicide, self-injury, victimization, and health neglect. These decisions are based on empirically derived risk factors and are to be distinguished from the total scores on strengths and vulnerabilities. Overall, evaluators in our study reached better agreement for the final risk judgments than for total scores.

Also in contrast to other studies examining the START:AV (Klimukienė et al., 2018; Viljoen et al., 2012), we found that the majority of the adverse outcomes demonstrated similar or higher interrater reliability. Only the reliability of the risk of unauthorized absences was considerably lower than reported by Klimukienė et al. (2018). Of all risk estimates in their study, reliability was highest for this adverse outcome, whereas in our study, it was the lowest of all risk estimates. Specifically, the two evaluator groups in our study disagreed about whether the risk of this outcome was moderate versus high (in 28% of the cases). Further, this disagreement was not always in the same direction. In 15 cases, the research evaluator assigned high risk, while in the other 13 cases the clinical evaluator assigned high risk. We could not find a plausible explanation for this finding, except perhaps that the scoring instructions for this adverse outcome do not differentiate sufficiently between moderate and high risk.

Likewise, the interrater reliability of the history ratings – which ranged from poor to excellent, with the majority of history ratings falling within the moderate range – were lower than what is typically found for static factors (Edens et al., 2019) and for historical factors in SPJ instruments in particular. For example, field studies with the Historical Clinical Risk Management-20 (HCR-20; Webster et al., 1997) have found good interrater reliability for the historical scale, ranging from 0.78 to 0.86 (de Vogel & de Ruiter, 2004; Jeandarme et al., 2017). However, de Vogel and de Ruiter (2004) also reported low interrater reliability for the item "previous violence." They found that this item was misunderstood by some clinical evaluators (i.e., to not include the index offense, when it fact it does), resulting in disagreements between clinical and research evaluators.

When we further explored the disagreements on history ratings, we found that research evaluators identified considerably more incidents in the recent history of several different adverse outcomes than did clinical evaluators. One hypothesis is that research evaluators were more thorough in their rating of incidents. However, we did not observe a similar pattern for prior history. An alternative explanation is that research evaluators identified more incidents because they read the daily reports and immediately documented their findings in the START:AV form, whereas clinical evaluators, at the time of the risk assessment, tended to rely more on their memory of incidents. Personal communication with treatment coordinators confirmed this hypothesis: they assigned ratings based on their recollection from observations and conversations with the youth and the team because they lacked the time to sift through the progress notes.

Also noteworthy is the very low interrater reliability on the final risk judgment of health neglect, an issue not addressed in any other study. The ICC for this scale was one of the lowest we found. Health neglect behaviors may be more challenging to rate than, for example, violent or suicidal behaviors. Evaluators may have differing understandings of the extent to which behaviors such as insufficient sleep, unhealthy diet, or smoking should be considered as health neglect. In addition, discrepancies in the information available to the evaluator groups may

have contributed to a poor interrater reliability. Research evaluators had more missing ratings on the item involving self-care, which directly informs the adverse outcome of health neglect. Information required to rate self-care and the associated risk was likely not sufficiently documented in the progress notes, whereas clinicians could rely on their own or their team's observations.

4.2 | Factors influencing interrater reliability in the field

Confounds are an integral part of field studies and represent the realities of conducting risk assessment in practice settings. The goal is not to eliminate them, but to explore their influence on reliability (Guarnera & Murrie, 2017). We discuss factors that may have affected the present field study, resulting in lower interrater reliability than reported in the START:AV non-field studies, especially for the items and total scores.

4.2.1 | Factors related to the study design

The START:AV non-field studies (Klimukienė et al., 2018; Viljoen et al., 2012) likely benefited from a research design that required evaluators to reach sufficient reliability prior to starting the actual data collection. Such a procedure promotes consistency across the evaluators from the outset. In the present study, we did not establish interrater reliability for either the clinical evaluators or the research evaluators prior to data collection. This likely reflects most real-world services, which have to deal with time and resource constraints. Nevertheless, service providers and agencies are advised to conduct their own reliability and validity check of the risk assessment instrument prior to full implementation and to continue to assess its performance periodically (Vincent et al., 2012).

The present study also differed from the START:AV non-field studies with respect to the time between the training and the data collection. In our study, the last training moment (i.e., feedback on a practice case) occurred 3 months prior to the first START:AV assessments. In fact, 17 months had passed since the initial workshop on the administration of the START:AV. Clinicians have discussed this gap between training and the actual use of the START:AV as a barrier to the risk assessment implementation (De Beuf et al., 2020). In most non-field studies, this gap is smaller, because data collection usually starts shortly after training. This is relevant because over time, more errors may occur due to rater drift, hence weakening reliability (Vincent et al., 2012). In an early field study with the HCR-20, Belfrage (1998) actually excluded an evaluator from the reliability analyses, because this person scored significantly differently from the other clinicians, arguably due to a gap of 6 months between the training and the risk assessments. That said, in the present setting, some evaluators participated in a refresher workshop that took place halfway through the data collection period (De Beuf et al., 2020). This may have reduced rater drift, although the sample was too small to explore this further.

Another notable real-world factor that was at play in our field study was the variance in information sources between the evaluators. To complete the risk assessments, both evaluator groups had access to the same file information. However, the clinical evaluators additionally had direct contact with the adolescent and their treatment team. This may have resulted in qualitatively and quantitatively different input from both evaluators (Penney et al., 2014). For example, in the study conducted by de Vogel and de Ruiter (2004), treatment supervisors reported that they did not feel the need to review file information as they felt that this information was already known to them. They stated that they mainly relied on personal interactions with the patient. The treatment coordinators in the present setting confirmed this: for the sake of efficiency, they relied on their own observations and personal interactions. This approach to risk assessment by clinical staff may also explain why we found lower reliability for history ratings. Clinicians can have various reasons for not relying on file information, such as time constraints, negligence, overestimation of one's knowledge of the file, or simply because they did not adhere to the user manual (Storey et al., 2012).

4.2.2 | Factors related to the instrument

Rufino et al. (2011) found inverse correlations between interrater agreement and subjectivity of PCL-R and HCR-20 items. Similarly, other authors have found lower reliability for items that require more clinical judgment (Kennealy et al., 2016; Quesada et al., 2014). This conclusion might also be applicable to the START:AV; except for the history of adverse outcomes, the instrument consists of dynamic items that require clinical judgment. Ratings of low, moderate, or high are assigned based on the presence of "item anchors" (descriptors that represent the item) such as "acts without thinking" (item 7: impulsivity) or "caregivers provide a structured and stable environment" (item 15: parenting). However, these item anchors do not have explicit cutoff points to guide an evaluator in differentiating between, for example, a moderate and a high rating. The evaluator has discretion to rate the items as low, moderate, or high on the basis of the presence and severity of the item anchors observed in the assessed adolescent. The absence of clear boundaries between the ratings creates an opportunity for subjectivity. Thus, establishing behaviorally anchored cutoffs may help reduce subjectivity.

Dynamic items such as those found in the START:AV might be challenging to complete based on file information only (Penney et al., 2014). Indeed, the research evaluators in our study had considerably more missing item ratings than the clinical evaluators. Items that involved insight into the adolescent's situation outside the treatment setting (i.e., social support, peers, parental functioning, external triggers) had between 13% and 21% missing ratings. Similarly, items such as attitudes, plans, and self-care had considerably more missing ratings when completed by researchers, probably because this information was less often included in the daily progress notes.

4.3 | Practical implications

4.3.1 | Aiming for consensus

Our study indicates that, although they share some agreement, the two evaluator groups held different pieces of the risk assessment puzzle. Research evaluators extracted more incidents that occurred in the facility (i.e., recent history) from daily reports, whereas clinical evaluators relied on their interactions with the adolescent and team when rating the dynamic items. The evaluator groups not only brought different information to the table, they also differed in their relationship with the adolescent. The research evaluators had a more distant, objective position, whereas the clinicians, although serving in a supervisory and decision-making role rather than a therapeutic role, were more closely involved with the youth. Clinicians' feelings, positive and negative, that come with this relationship may result in subjectivity and impact the risk assessment (de Vogel & de Ruiter, 2004).

Nevertheless, de Vogel and de Ruiter (2006) demonstrated that risk assessments of both evaluator groups did not differ in terms of predictive performance. To optimally benefit from both information sources, the consensus approach is recommended (de Vogel & de Ruiter, 2004). In this approach, multiple staff members (e.g., researcher, treatment coordinator) complete the START:AV independently of each other and discuss their findings during a meeting, resulting in a consensus START:AV. Although the effect of the consensus approach on interrater reliability has, to our knowledge, not yet been evaluated, a study with the HCR-20 has demonstrated that a consensus assessment produced the highest predictive accuracy for future recidivism (de Vogel & de Ruiter, 2006). Agencies and services that prepare to introduce SPJ risk assessment instruments into their routine practice should consider integrating the consensus method in the implementation plan. Settings that do not have researchers or other objective evaluators available could create a rotation system of fellow clinicians to serve as second independent raters who complete risk assessments based on daily reports and other file information.

4.3.2 | Other suggestions for practice

In the event that consensus meetings may be challenging to organize due to limited resources or other constraints, service providers could consider alternatives, such as regular case discussions and peer review. A one-time training in the use of the risk assessment instrument is not sufficient to secure reliability because rater drift may occur. Vincent et al., (2012) suggested periodical monitoring to identify challenges in reliability and maintain a satisfactory level of rating consistency. Reliability may be considered a result of implementation fidelity, which is the degree to which a practice is being delivered as intended by its developers (Proctor et al., 2011). In the intervention literature, researchers have developed checklists to assess implementation fidelity of a practice of interest. Likewise, we developed a scale to assess the level of completion of START:AV rating forms: the START:AV Adherence Rating Scale (STARS; De Beuf, de Vogel, & de Ruiter, 2020). This scale codes how complete the rating forms are, but does not investigate the quality of the ratings as measured by the descriptors listed in the user guide. While the latter measure would obviously be most relevant to the concept of reliability, the extent to which the adherence scale could serve as a proxy for reliability should be explored. If the STARS score indeed correlates with reliability, it would present a quick and accessible method to periodically assess this psychometric quality.

In addition to training and monitoring efforts, well-communicated arrangements about which information is to be used for the risk assessment, taking feasibility into account, may benefit the reliability of the ratings. Moreover, service providers that are unfamiliar with documenting strengths may want to invest additional effort in coaching staff on the observation and documentation of strengths and talents of clients. This advice also applies to items that are routinely found to be absent in files (e.g., attitudes and self-care in the present setting).

Risk assessment tool developers and translators could help to improve reliability by presenting concrete examples and unambiguous scoring instructions in the user guide. As mentioned earlier, several START:AV rating instructions allow room for subjective opinion, which could be limited by adding more examples (e.g., examples of mild health neglect, yet serious enough to rate as present in history). It is recommended that developers of risk assessment instruments continue to collect feedback from practitioners and trainees, because it may provide helpful suggestions for fine-tuning of the instrument's instructions.

4.4 | Limitations and future directions

Field interrater reliability ideally involves two field evaluators who are both familiar with the examinee and who conduct the assessment for clinical-forensic decision-making purposes, based on the same information. However, in clinical practice, it is not that common to have two or more independent evaluators conduct the same risk assessment. Hence, in our study, we relied on research assistants as the second evaluator of the START:AV. This limits the interpretation of the study as a prototypical field study, where both assessments are completed for non-research purposes. Our study is also limited in that the evaluator groups had different information at their disposal. The clinical evaluators could rely on direct and indirect observations of the adolescent, in addition to file information, while the researchers had access only to the latter source. Nevertheless, even in practice, evaluators rarely will collect and use the exact same information. Therefore, the present study provided an opportunity to identify strengths in the approach of each evaluator group. Further, as discussed earlier, the consensus methodology is a way to turn these information differences into a strength by bringing together multiple perspectives in consensus meetings. Moreover, it would be informative to investigate in which direction the consensus ratings are trending across and within evaluators, compared with the individual ratings.

Multiple evaluator characteristics (e.g., training, profession, experience, personality traits) can affect interrater reliability (Boccaccini et al., 2008). Not assessing these variables can be considered a limitation of this study. Future research should aim to measure these and other possibly confounding variables to be better able to explain whether the variance in interrater reliability is true variance or error due to evaluator and context effects. By

increasing the number of available field studies and aggregating their findings, we may eventually be more confident in gauging the performance of the risk assessment instruments in real-world settings.

Furthermore, future (field) research should invest in assessing whether evaluators identify the same items or item clusters as key or critical. Ultimately, these strengths and vulnerabilities are most likely to inform risk management goals. If the determination of key and critical items demonstrates low reliability, their relevance to risk management may be questionable. Likewise, the reliability of risk formulations, scenario planning, and—most critically—risk management plans is a relevant topic for future research.

Another important point for future research is that researchers specify which interrater reliability statistic was used, because this affects the interpretation and comparisons that can be made. For example, there are 10 forms of ICCs and the reader should be able to verify whether the appropriate ICC was used. Likewise, it is recommended to report the CIs of reliability statistics because they present a more complete picture of the estimate, resulting in more appropriate conclusions about the level of reliability (Koo & Li, 2016). In the present study, the CIs are relatively wide, which is in line with Edens and Boccaccini's (2017) comment that with lower reliability come wider CIs. In four instances, the CI covered three qualitative labels (e.g., the reliability of prior history of Violence ranged from poor to good). Edens and Boccaccini (2017) anticipated that CIs would be wider in field studies than non-field studies. However, we could not confirm this for the START:AV, as the previous studies have not reported CIs for interrater reliability.

The field should also continue the conversation on what is an acceptable level of reliability in legal and forensic settings. Currently, there is no agreement on a benchmark for sufficient reliability of forensic risk assessment instruments in the field, nor for other forensic evaluations (Edens & Kelley, 2017). Even in forensic sciences, there is currently no standard of acceptable error rates (Dror, 2020). Establishing what is "acceptable" is a complex matter, because it is specific to a forensic domain, setting, evaluator, case, and so on (Dror, 2020). Conducting more field studies might give direction as to what is reasonable to expect, what is possible to achieve, and what is minimally required considering the high stakes of risk assessment decisions in the forensic and legal arena.

5 | CONCLUSIONS

Compared with the sheer volume of studies on the validity of risk assessment instruments, reliability studies are underrepresented in the peer-reviewed literature. This is especially true for field studies on reliability. The present field study addressed this gap by evaluating the START:AV in a secure youth care facility. We found moderate to good interrater reliability for most final risk judgments, which, based on how they are intended to be used clinically, are the most essential decisions made during an SPJ risk assessment before proceeding to risk management. Still, the interrater reliability for the individual items and total scores was lower than previously found in START:AV non-field studies. This is in line with the growing literature on risk assessment instruments in real-world settings.

Understanding an instrument's performance in a particular setting is relevant to agencies and services that wish to implement it. Implementation plans should take into account real-life factors that may affect interrater reliability and take steps to minimize their influence. For example, the consensus approach to risk assessment may strengthen field reliability, which in turn may enhance field validity. That is, the "boss" showing who's boss.

ORCID

Tamara L. F. De Beuf https://orcid.org/0000-0001-5273-8523

Corine de Ruiter https://orcid.org/0000-0002-0135-9790

John F. Edens [ORCID] https://orcid.org/0000-0002-2218-0650

Vivienne de Vogel [ORCID] https://orcid.org/0000-0001-7671-1675

ENDNOTES

[1] Due to potential concern about including multiple assessments of the same adolescent, the intraclass correlation coefficients (ICCs) were examined using the sample with unique adolescents only ($n = 82$). Results were highly similar to the full sample.

[2] The ICCs in Troquete et al. (2015) were calculated according to the more stringent requirement of absolute agreement rather than our consistency definition.

REFERENCES

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.) American Psychiatric Association.

Belfrage, H. (1998). Implementing the HCR-20 scheme for risk assessment in a forensic psychiatric hospital: Integrating research and clinical practice. *Journal of Forensic Psychiatry*, *9*(2), 328–338. https://doi.org/10.1080/09585189808402200

Boccaccini, M. T., Murrie, D. C., Caperton, J., & Hawes, S. (2009). Field validity of the STATIC-99 and MnSOST-R among sex offenders evaluated for commitment as sexually violent predators. *Psychology, Public Policy, and Law*, *15*(4), 278–314. https://doi.org/10.1037/a0017232

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*(4), 262–283. https://doi.org/10.1037/a0014523

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the sexual violence risk-20: Professional guidelines for assessing risk of sexual violence*. Simon Fraser University.

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

De Beuf, T. L. F., de Ruiter, C., & de Vogel, V. (2020). Staff perceptions on the implementation of Structured Risk Assessment with the START:AV: Identifying barriers and facilitators in a residential youth care setting. *International Journal of Forensic Mental Health*, *19*(3), 297–314. https://doi.org/10.1080/14999013.2020.1756994

De Beuf, T. L. F., de Vogel, V., & de Ruiter, C. (2020). Adherence to structured risk assessment guidelines: Development and preliminary evaluation of an adherence scale for the START:AV. *Journal of Forensic Psychology Research and Practice*, *20*(5), 413–435. https://doi.org/10.1080/24732850.2020.1756676

DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S., … Reidy, T. J. (2020). Statement of concerned experts on the use of the Hare Psychopathy Checklist-Revised in capital sentencing to assess risk for institutional violence. *Psychology, Public Policy, and Law*, *26*(2), 133–144. https://doi.org/10.1037/law0000223

Desmarais, S. L., Sellers, B. G., Viljoen, J. L., Cruise, K. R., Nicholls, T. L., & Dvoskin, J. A. (2012). Pilot implementation and preliminary evaluation of START:AV assessments in secure juvenile correctional facilities. *International Journal of Forensic Mental Health*, *11*(3), 150–164. https://doi.org/10.1080/14999013.2012.737405

de Vogel, V., & de Ruiter, C. (2004). Differences between clinicians and researchers in assessing risk of violence in forensic psychiatric patients. *Journal of Forensic Psychiatry and Psychology*, *15*(1), 145–164. https://doi.org/10.1080/14788940410001655916

de Vogel, V., & de Ruiter, C. (2006). Structured professional judgment of violence risk in forensic clinical practice: A prospective study into the predictive validity of the Dutch HCR-20. *Psychology, Crime and Law*, *12*(3), 321–333. https://doi.org/10.1080/10683160600569029

Dror, I. E. (2020). The error in "error rate": Why error rates are so needed, yet so elusive [commentary]. *Journal of Forensic Sciences*, *65*(4), 1–6. https://doi.org/10.1111/1556-4029.14435

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment "out of the lab" and into "the real world:" Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment*, *29*(6), 599–610. https://doi.org/10.1037/pas0000475

Edens, J. F., & Kelley, S. E. (2017). "Meet the new boss. Same as the old boss": A commentary on Williams, Wormith, Bonta, and sitarenios (2017). *International Journal of Forensic Mental Health*, *16*, 23–27. http://dx.doi.org/10.1080/14999013.2016.1268221

Edens, J. F., Magyar, M. S., & Cox, J. (2013). Taking psychopathy measures "out of the lab" and into the legal system: Some practical concerns. In K. Kiehl, & W. Sinnott-Armstrong (Eds.), *Handbook on psychopathy and law* (pp. 250–272). Oxford University Press.

Edens, J. F., Penson, B. N., Ruchensky, J. R., Cox, J., & Smith, S. T. (2016). Interrater reliability of violence risk appraisal guide scores provided in Canadian criminal proceedings. *Psychological Assessment*, *28*, 1543–1549. https://doi.org/10.1037/pas0000278

Edens, J. F., Petrila, J., & Kelley, S. E. (2019). Legal and ethical issues in the assessment and treatment of psychopathy. In C. J. Patrick (Ed.), *Handbook of psychopathy* (pp. 732–751). The Guilford Press.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. Wiley.

Guarnera, L. A., & Murrie, D. C. (2017). Field reliability of competency and sanity opinions: A systematic review and meta-analysis. *Psychological Assessment*, *29*, 795–818. http://dx.doi.org/10.1037/pas0000388

Gwet, K. L. (2002). Interrater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Interrater Reliability Assessment Series*, *2*, 1–9. Retrieved from http://agreestat.com/papers/papers/inter_rater_reliability_dependency.pdf

Gwet, K. L. (2020). *AgreeStat360 [Computer software]*. AgreeStat Analytics. Retrieved from http://agreestat.com/software/default.html

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*(1), 1–21. https://doi.org/10.1037/a0014421

Hare R. D. (2003). *The Hare psychopathy Checklist-revised* (2nd ed.) MultiHealth Systems, Inc.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior*, *16*(3), 257–272. http://dx.doi.org/10.1007/BF01044769

Jeandarme, I., Pouls, C., Laender, J. D., Oei, T. I., & Bogaerts, S. (2017). Field validity of the HCR-20 in forensic medium security units in Flanders. *Psychology, Crime and Law*, *23*(4), 305–322. https://doi.org/10.1080/1068316X.2016.1258467

Kennealy, P. J., Skeem, J. L., & Hernandez, I. R. (2016). Does staff see what experts see? Accuracy of front line staff in scoring juveniles' risk factors. *Psychological Assessment*, *29*(1), 26–34. https://doi.org/10.1037/pas0000316

Klimukienė, V., Laurinavičius, A., Laurinaitytė, I., Ustinavičiūtė, L., & Baltrūnas, M. (2018). Examination of convergent validity of START:AV ratings among male juveniles on probation. *International Journal of Psychology: Biopsychosocial Approach*, *22*, 31–54. https://doi.org/10.7220/2345-024x.22.2

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Levenson, J. S. (2004). Sexual predator civil commitment: A comparison of selected and released offenders. *International Journal of Offender Therapy and Comparative Criminology*, *48*(6), 638–648. https://doi.org/10.1177/0306624X04265089

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment*, *24*(4), 944–953. https://doi.org/10.1037/a0028411

Neal, T. M. S., Miller, S. L., & Shealy, R. C. (2015). A field study of a comprehensive violence risk assessment battery. *Criminal Justice and Behavior*, *42*(9), 952–968. https://doi.org/10.1177/0093854815572252

Neal, T. M. S., Slobogin, C., Saks, M. J., Faigman, D. L., & Geisinger, K. F. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom? *Psychological Science in the Public Interest*, *20*(3), 135–164. https://doi.org/10.1177/1529100619888860

Nicholls, T. L., Brink, J., Desmarais, S. L., Webster, C. D., & Martin, M. (2006). The short-term assessment of risk and treatability (START): A prospective validation study in a forensic psychiatric sample. *Assessment*, *13*(3), 313–327. https://doi.org/10.1177/1073191106290559

Nonstad, K., Nesset, M. B., Kroppan, E., Pedersen, T. W., Nøttestad, J. A., Almvik, R., & Palmstierna, T. (2010). Predictive validity and other psychometric properties of the Short-Term Assessment of Risk and Treatability (START) in a Norwegian high secure hospital. *International Journal of Forensic Mental Health*, *9*(4), 294–299. https://doi.org/10.1080/14999013.2010.534958

Penney, S. R., McMaster, R., & Wilkie, T. (2014). Multirater reliability of the historical, clinical, and risk management-20. *Assessment*, *21*(1), 15–27. https://doi.org/10.1177/1073191113514107

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., … Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 65–76. https://doi.org/10.1007/s10488-010-0319-7

Quesada, S. P., Calkins, C., & Jeglic, E. L. (2014). An examination of the interrater reliability between practitioners and researchers on the Static-99. *International Journal of Offender Therapy and Comparative Criminology*, *58*(11), 1364–1375. https://doi.org/10.1177/0306624x13495504

Rettenberger, M., Rice, M., Harris, G., & Eher, R. (2017). Actuarial risk assessment of sexual offenders: The psychometric properties of the Sex Offender Risk Appraisal Guide (SORAG). *Psychological Assessment*, *29*(6), 624–638. http://dx.doi.org/10.1037/pas0000390

Rufino, K. A., Boccaccini, M. T., & Guy, L. S. (2011). Scoring subjectivity and item performance on measures used to assess violence risk: The PCL-R and HCR-20 as exemplars. *Assessment*, *18*(4), 453–463. https://doi.org/10.1177/1073191110378482

Schmidt, F., Hoge, R., & Robertson, L. (2005). Reliability and validity analyses of the youth level of services/case management inventory. *Criminal Justice and Behavior*, *32*(3), 329–344. https://doi.org/10.1177/0093854804274373

Sellers, B. G., Desmarais, S. L., & Hanger, M. W. (2017). Measurement of change in dynamic factors using the START: AV. *Journal of Forensic Psychology Research and Practice*, *17*(3), 198–215. https://doi.org/10.1080/24732850.2017.1317560

Sher, M. A., Warner, L., McLean, A., Rowe, K., & Gralton, E. (2017). A prospective validation study of the START:AV. *Journal of Forensic Practice*, *19*(2), 115–129. https://www.emerald.com/insight/content/doi/10.1108/JFP-10-2015-0049/full/html

Singh, J. P., Desmarais, S. L., Sellers, B. G., Hylton, T., Tirotti, M., & Van Dorn, R. A. (2014). From risk assessment to risk management: Matching interventions to adolescent offenders' strengths and vulnerabilities. *Children and Youth Services Review*, *47*(1), 1–9. http://dx.doi.org/10.1016/j.childyouth.2013.09.015

Storey, J. E., Watt, K. A., Jackson, K. J., & Hart, S. D. (2012). Utilization and implications of the Static-99 in practice. *Sexual Abuse: A Journal of Research and Treatment*, *24*(3), 289–302. https://doi.org/10.1177/1079063211423943

Ten Brummelaar, M. D. C., Harder, A. T., Kalverboer, M. E., Post, W. J., & Knorth, E. J. (2017). Residential child and youth care in The Netherlands: Current practices and future perspectives. In T. Islam, & L. Fulcher (Eds.), *Residential child and youth care in a developing world. Volume 2: European Perspectives* (1st ed., pp. 339–355). CYC-Net Press.

Troquete, N. A. C., van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2015). Predictive validity of the Short-Term Assessment of Risk and Treatability for violent behavior in outpatient forensic psychiatric patients. *Psychological Assessment*, *27*(2), 377–391. http://dx.doi.org/10.1037/a0038270

Viljoen, J. L., Beneteau, J. L., Gulbransen, E., Brodersen, E., Desmarais, S. L., Nicholls, T. L., & Cruise, K. R. (2012). Assessment of multiple risk outcomes, strengths, and change with the START:AV: A short-term prospective study with adolescent offenders. *International Journal of Forensic Mental Health*, *11*(3), 165–180. https://doi.org/10.1080/14999013.2012.737407

Viljoen, J. L., Nicholls, T. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (2014). *Short-term assessment of risk and treatability: Adolescent version (START:AV) – user guide*. Mental Health, Law, and Policy Institute.

Viljoen, J. L., Nicholls, T. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (2016). *Short-term assessment of risk and treatability: Adolescent version (START:AV) – user guide (T. L. F. De Beuf, C., de Ruiter, & V. de Vogel, trans.)*. Mental Health, Law, and Policy Institute, Simon Fraser University (Original work published 2014).

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). *Risk assessment in juvenile justice: A guidebook for implementation*. John D. & Catherine T. MacArthur Foundation. Retrieved from http://modelsforchange.net/publications/346

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*(3), 225–236. https://doi.org/10.1037/h0093974

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence (Version 2)*. Simon Fraser University.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Desmarais, S. (2009). *Manual for the short-term assessment of risk and treatability (START)*. Forensic Psychiatric Services Commission and St. Joseph's Healthcare.

Webster, C. D., Müller-Isberner, R., & Fransson, G. (2002). Violence risk assessment: Using structured clinical guides professionally. *International Journal of Forensic Mental Health*, *1*(2), 185–193. https://doi.org/10.1080/14999013.2002.10471173

Wechsler, D. (2017). *Wechsler intelligence scale for Children (5th ed.): WISC-V*. translated by M. P. H. Hendriks & S. Ruiter. Pearson (Original work published 2014).