Journal of Forensic Psychology Research and Practice

ISSN: 2473-2850 (Print) 2473-2842 (Online) Journal homepage: https://www.tandfonline.com/loi/wfpp21

Adherence to Structured Risk Assessment Guidelines: Development and Preliminary Evaluation of an Adherence Scale for the START:AV

Tamara L. F. De Beuf, Vivienne de Vogel & Corine de Ruiter

To cite this article: Tamara L. F. De Beuf, Vivienne de Vogel & Corine de Ruiter (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

To link to this article: https://doi.org/10.1080/24732850.2020.1756676

Published online: 27 Apr 2020.

JOURNAL OF FORENSIC PSYCHOLOGY RESEARCH AND PRACTICE
https://doi.org/10.1080/24732850.2020.1756676

Routledge
Taylor & Francis Group

Adherence to Structured Risk Assessment Guidelines: Development and Preliminary Evaluation of an Adherence Scale for the START:AV

Tamara L. F. De Beuf [a,b], Vivienne de Vogel [c], and Corine de Ruiter [b]

[a]Research department, Ottho Gerhard Heldring Institution, Zetten, The Netherlands; [b]Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands; [c]Research department, The Forensic Care Specialists, Utrecht, The Netherlands

ABSTRACT

Risk assessment instruments are widely used to predict risk of adverse outcomes, such as violence or victimization, and to allocate resources for managing these risks among individuals involved in criminal justice and forensic mental health services. For risk assessment instruments to reach their full potential, they must be implemented with fidelity. A lack of information on administration fidelity hinders transparency about the implementation quality, as well as the interpretation of negative or inconclusive findings from predictive validity studies. The present study focuses on adherence, a dimension of fidelity. Adherence denotes the extent to which the risk assessment is completed according to the instrument's guidelines. We developed an adherence measure, tailored to the Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV), an evidence-based risk assessment instrument for adolescents. With the START:AV Adherence Rating Scale, we explored the degree to which 11 key features of the instrument were adhered to in 306 START:AVs forms, completed by 17 different evaluators in a Dutch residential youth care facility over a two-year period. Good to excellent interrater reliability was found for all adherence items. We identified differences in adherence scores on the various START:AV features, as well as significant improvement in adherence for those who attended a START:AV refresher workshop. Outcomes of risk assessment instruments potentially impact decision-making, for example, whether a youth's secure placement should be extended. Therefore, we recommend fidelity monitoring to ensure the risk assessment practice was delivered as intended.

KEYWORDS

Implementation; adherence; fidelity; structured risk assessment; START:AV

Evidence-based practice is an approach to clinical decision-making that integrates the best available research evidence with professional judgment and expertise of the service provider, and the values and preferences of the individual service user (American Psychological Association, 2006). It promotes the use of assessment methods and interventions that have

CONTACT Tamara L. F. De Beuf t.debeuf@ogheldring.nl Ottho Gerhard Heldring Institution, P. O. Box 1, Zetten 6670 AA, The Netherlands

demonstrated the best outcomes for a particular (mental) health issue in empirical research. It also implies an evidence-based practice should be delivered as designed, i.e., delivered with fidelity.

Fidelity

Fidelity is defined as the degree to which a method or intervention is implemented as it was intended by its developers (Proctor et al., 2011). It is also referred to as "integrity", "adherence", or "quality of delivery" (Breitenstein et al., 2010; Proctor et al., 2011). Without an evaluation of fidelity, it is unclear to stakeholders, such as patients, managers, and society, to what extent the quality of implementation is assured (Mowbray et al., 2003). Moreover, evaluating fidelity and providing practitioners with feedback on their performance can be a successful strategy to improve implementation quality (Brown et al., 2019). The effectiveness of performance feedback on fidelity has been demonstrated in medical and educational contexts (Fallon et al., 2018; Loy et al., 2016). A simple feedback intervention, such as distributing charts with fidelity outcomes supplemented with recommendations for compliance increases fidelity with long-term effects (Loy et al., 2016). In addition, monitoring and measuring fidelity helps avoid type III errors, which occur when implementation failures lead to poor outcomes of potentially effective interventions (Breitenstein et al., 2010). Results from effectiveness and predictive validity studies are difficult to interpret when there is no information on fidelity. This is especially true for negative or inconclusive findings, because it is impossible to determine whether the results reflect an inadequacy of the method itself or a failure to implement the method as intended (Carroll et al., 2007; Mowbray et al., 2003). Similarly, for secondary analysis purposes, such as systematic reviews and meta-analyses, information on implementation fidelity is essential for allowing legitimate comparisons (Carroll et al., 2007).

Fidelity is a complex and multidimensional construct. Five dimensions have been identified in the literature: adherence, quality of delivery (or provider competence), exposure (or dosage), program differentiation, and participant responsiveness (Carroll et al., 2007; Dusenbury et al., 2003; Proctor et al., 2011). *Adherence* refers to whether a practice is completed as it was designed and described in the guidelines. *Quality of delivery* refers to the practitioner's competence to deliver the practice. It concerns how *well* the practice is delivered, whereas adherence denotes how *compliant* the delivered practice is with the specified guidelines. *Exposure or dosage* refers to the number and frequency of the practice received by the intended population; in other words, is the practice delivered at the agreed times and as often as required? *Program differentiation* refers to the formulation of essential core features that are necessary for the practice to reach its goals. Lastly,

participant responsiveness indicates how responsive or involved participants are in the delivered practice (Dusenbury et al., 2003).

Structured risk assessment

The current paper focuses on the relevance of fidelity for the evidence-based practice of structured risk assessment in forensic mental health and criminal justice settings. Forensic risk assessment includes the judgment of an individual's potential for violence or other adverse outcomes, such as victimization or self-harm. The ultimate goal is to prevent adverse outcomes (Hart et al., 2017). *Structured* refers to the procedure: the risk assessment follows a systematic approach to information gathering and to the evaluation of empirically derived risk and protective factors. The structured approach has generated multiple instruments that have been widely adopted to make risk judgments and allocate resources for risk management (e.g., surveillance, interventions; Viljoen et al., 2018). These instruments are considered evidence-based if they are developed on the basis of the most accurate scientific and professional knowledge available on risk and protective factors, their potential interaction, and their relationship with adverse outcomes (Hart & Logan, 2011). However, fidelity to the instruments' guidelines is rarely addressed in structured risk assessment research and practice.

Fidelity in structured risk assessment

We have observed that fidelity dimensions are used interchangeably or operationalized inconsistently in violence risk assessment research. For example, Viljoen et al. (2018) used the term *"adherence to the tool"* as part of their systematic review on the utility of risk assessment instruments for risk management and recidivism reduction. However, "adherence to the tool" was operationalized in various ways across the studies included in the review. For instance, it was operationalized as whether professionals received training in rating the instrument, whether the instrument was completed as mandated by the service, as well as whether those who were scheduled to be assessed within a program were actually assessed. Overall, the authors found that many studies did not include information about adherence to the risk assessment instrument, and for those that did, adherence was often inadequate. From this review, it can be concluded that in structured risk assessment research, adherence is not univocally defined nor examined. Other studies in this field have addressed *"adherence to administration policy"*, operationalized as the number of youths that received a risk assessment (Vincent et al., 2012, 2016; Young et al., 2006). Although labeled adherence, according to the definitions in implementation science, this is rather a measure of the exposure/dosage dimension of fidelity. To promote

Table 1. Definitions of the fidelity dimensions applied to structured risk assessment practices.

Fidelity dimension	Definition in Implementation Research[a]	Applied to structured risk assessment
Adherence	The extent to which the delivery of the intervention content is consistent with how it was designed or written.	Whether the evaluators completed the risk assessment instrument as it was designed and presented via the guidelines in a user manual (e.g., Did the evaluator rate all the items?).
Quality of Delivery	The extent to which a provider approaches a theoretical ideal in terms of delivering the intervention content (i.e., provider effectiveness).	The evaluator's competence to conduct the assessment (e.g., Is the evaluator attending adequately to the rating criteria?).
Exposure/Dosage	The amount of intervention content received by participants in terms of number of sessions, duration, or intensity.	The number and frequency of risk assessments received by the intended population (e.g., Did individuals receive a risk assessment at the agreed times and occasions?).
Program Differentiation	Identification of an intervention's unique features, without which it would not have the intended effect.	Identification of indispensable core features of the risk assessment instrument that are necessary for the instrument to reach its goals.
Participant Responsiveness	Participants' engagement and involvement in the intervention.	How involved examinees are in the risk assessment (e.g., Is their own risk appraisal included?).

[a]Definitions are adapted from Dusenbury et al. (2003).

uniformity when discussing fidelity issues in this area of research, we apply the fidelity dimensions to the subject of structured risk assessment and suggest analogous definitions (seeTable 1).

In the present paper, we focus on the adherence dimension of fidelity, that is, whether the evaluators completed the risk assessment instrument as designed and described in a protocol or a user manual. One of the first attempts to assess adherence to risk assessment practice, to our knowledge, is the study of McNiel et al. (2008) on "documentation quality". Using structured content analysis, the authors evaluated the presence of key risk assessment features in progress notes from psychology interns and residents in psychiatry. Their main goal was to assess whether the participants' documentation of the risk assessment improved after attending a workshop on evidence-based risk assessment. Although not assessing adherence to a specific risk assessment instrument, the study examined the presence of 20 characteristics that were, according to the authors, recommended as best practices in the violence risk assessment literature at the time (e.g., Gutheil & Appelbaum, 2000; Monahan, 1993; Simon, 2003). They rated the presence of these items in the trainees' documentation as *item absent* (0), *item possibly present* (1), and *item present* (2). Findings showed that those who attended the workshop, compared to those who did not, improved significantly more on reporting specific risk and protective factors, and on articulating

a rationale for the level of risk and for the advised risk management strategies.

A few years later, McNiel et al. (2011) developed a structured tool to measure "clinical competency" in violence risk assessment and management: the Competency Assessment Instrument for Violence Risk (CAI-V). The checklist was based on an adaptation of the criteria used in the previous study on "documentation quality" (McNiel et al., 2008), on the literature on violence risk assessment and management, as well as the literature on competencies for psychiatric residents. The CAI-V included 31 components that assessed: 1) used sources, 2) identified risk and protective factors, 3) risk judgment and risk communication, 4) treatment planning, and 5) written documentation. In this study, trainees interviewed a mock patient (a senior trainee), wrote a progress note, and gave an oral summary of the risk assessment and risk management plan to independent observers (faculty members). Based on this procedure, the observers rated the CAI-V items as *task not done* (1), *working toward competency* (2), *competent* (3), and *advanced* (4). The study found that second-year trainees, who had more risk assessment experience, received higher ratings on the CAI-V than first-year trainees, and thus, adhered more adequately to the risk assessment practice. However, neither of the studies of McNiel and colleagues examined adherence to a specific risk assessment instrument.

To our knowledge, Reynolds and Miles (2009) were the first to measure adherence to a particular instrument, the Historical, Clinical, and Risk management-20 (HCR-20; Webster et al., 1997). On a 3-point scale (1 = *not present*, 2 = *poor*, 3 = *good*), they measured how complete the HCR-20 assessments were in terms of four components: historical factors, clinical factors, risk management factors, and the risk management plan. The components were rated as "poor" when essential information was missing and "good" when all key information was covered. Although the authors labeled this an HCR-20 "quality" evaluation, they were actually assessing adherence to the instrument's features. The aim of the study was to compare the quality of the HCR-20 assessments before and after staff had received HCR-20 training. Reynolds and Miles found an effect of training: the HCR-20 forms and risk management plans were significantly more complete after the training.

More recently, Sen et al. (2015) adjusted the CAI-V (McNiel et al., 2011) to assess adherence of completed HCR-20 forms to the instrument's instructions. With the resulting "HCR-20 quality assessment guide", the authors identified shortcomings in the HCR-20s completed within their service. For example, collateral information from significant others was not used, the psychopathy item was frequently omitted, only 40% of the forms included a rationale for the summary risk judgment, etc. Based on their audit, the authors formulated recommendations for improvement of the service's risk

assessment practice, such as encouraging clinicians to over-ride the patient's non-consent to access collateral information sources, when public safety was at issue. Although referred to as a quality assessment, this study assessed how complete the HCR-20 risk assessments were, a component of adherence.

In sum, only two studies have assessed adherence to a specific structured risk assessment instrument (Reynolds & Miles, 2009; Sen et al., 2015). Both studies concerned the HCR-20 version 2, a violence risk assessment instrument for adults, last updated in 2013 (HCR-20^{V3}; Douglas et al., 2013).

The present study

Conducting adherence checks followed by feedback is an important strategy for quality improvement of evidence-based practices (Loy et al., 2016). Thus, to ensure that structured risk assessment is delivered as intended, both in practice and in research, risk assessment evaluators might benefit from a similar feedback loop based on adherence measures tailored to the instrument of preference. We developed and tested an adherence rating scale for the Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV; Viljoen et al., 2014/2016), a structured risk assessment instrument for adolescents. The current paper reports on the interrater reliability of this scale, as well as the adherence results for 306 START:AV forms completed over a two-year period. In addition, adherence scores before and after a START:AV refresher workshop were evaluated.

Hypotheses

Based on previous literature, we formulate the following hypotheses:

(1) Adherence will vary per START:AV feature. In line with Sen et al.'s (2015) finding that rationales were only provided in 40% of the HCR-20s, it is hypothesized that adherence scores for features that require reflection and written arguments will be lower than features that only require checking off boxes.
(2) Reynolds and Miles (2009) found that HCR-20s were significantly more complete after training. Therefore, it is predicted that START:AV forms will have significantly higher adherence item and total scores after the refresher workshop compared to before. This effect is only expected for START:AVs completed by evaluators who participated in the workshop (i.e., refresher group) and not for the START:AVs completed by evaluators who did not participate (i.e., comparison group). Furthermore, we hypothesized that START:AV forms completed by the refresher group compared to forms completed by the comparison group will reach higher post-workshop adherence total scores. This is in line with McNiel et al. (2008) who found significantly

more improvement in the overall completeness of progress notes by staff who received a risk assessment workshop versus staff who attended a workshop not focused on risk assessment.

Method

This study is part of a larger implementation project with the START:AV in a Dutch residential youth care facility. In February 2016, the risk assessment instrument was introduced in a facility that provides medium and high secure care to adolescents. The service, with a capacity of 98 beds, treats boys and girls with severe behavioral and mental health problems to improve their safety (e.g., self-harm, victimization) and/or the safety of others (e.g., physical violence toward others). Youths are admitted under civil law with a child protection order and stay for 262 days on average, ranging from 4 to 717 days (A. Baanders, personal communication, January 31, 2019).

START:AV

The START:AV is an evidence-based risk assessment instrument consisting of 24 items that are rated twice: as strength (protective factor) and as vulnerability (risk factor). Based on the presence of these items and prior history, risk estimates (*low, moderate, high*) are formulated for eight adverse outcomes: violence to others, nonviolent offending, substance abuse, unauthorized absence, suicide, self-harm, victimization, and self-neglect. Furthermore, any imminent and serious risk of harm in relation to violence, suicide, self-harm, and victimization is also assessed and referred to as "THREAT" (see Viljoen et al., 2014/2016). In addition, for strengths and vulnerabilities, evaluators identify between two and six "key strengths" and "critical vulnerabilities" that are considered of particular relevance to the adolescent's risk. After completing the sections with the items and the adverse outcomes, the START:AV concludes with a text-box feature "periods of stability". If applicable, the evaluator describes circumstances in which the adolescent was doing relatively well. Reflecting on periods of stability in the past can help identify strengths in the youth's life. Overall, the START:AV's objective is to guide risk management and intervention planning in order to reduce the occurrence of adverse outcomes.

Implementation

Starting from February 2016, the START:AV was incorporated in the service's treatment cycle: it was integrated in the workflow, treatment plans, and treatment evaluation procedures. Treatment coordinators (i.e., the

"evaluators") were responsible for completing the START:AV rating forms. These professionals, with at least a master's degree in psychology or special needs education, are in charge of the adolescent's treatment process. In consultation with the team on the ward, they decide on the treatment approach in terms of treatment goals and type of therapy, and they evaluate treatment progress.

All treatment coordinators received a two-day START:AV training by the first author and discussed additional practice cases during another two-hour training session, led by the first and third author. Those hired after the initial training received one-on-one training by the first author shortly after starting employment. Within the present setting, the *comprehensive* rating form was implemented: this form includes text boxes in which evaluators are expected to write a rationale for their ratings. For more details on the implementation and the use of the START:AV within this service, we refer to De Beuf et al. (2019).

Development of the START:AV Adherence Rating Scale (STARS)

We developed the STARS after reviewing other fidelity assessment efforts outside the forensic field (McHugo et al., 2007; De Vos et al., 2013) and by following three steps documented in the literature (Mowbray et al., 2003).

Step 1: identifying adherence items and indicators

The primary focus was on risk assessment, therefore, risk management and risk communication features fell outside the scope of the STARS. The risk assessment process includes three phases: 1) gathering information, 2) rating the items/factors, and 3) rating history and future risk of the adverse outcomes. These phases are reflected in the different chapters of the START:AV user guide (Viljoen et al., 2014/2016), as well as in the comprehensive rating form. Each step's key features were translated into adherence items. In total, 11 items were identified (see Table 2): items 1 and 2 assess adherence to information gathering and documentation instructions (e.g., "Information is gathered from multiple sources"); items 3 to 6 assess adherence to instructions for the START:AV item ratings (e.g., "At least two and at most six key strengths are marked"); items 7 to 11 concern adherence to the guidelines for rating adverse outcomes (e.g., "For each adverse outcome rated as moderate or high risk, the risk rating is explained"). This item set was agreed upon by consensus among the authors, all experts in risk assessment research and practice.

Eight STARS items are rated on a 3-point scale as 0 (*insufficient*), 1 (*sufficient*), or 2 (*good*), because Likert-type scales allow detection of variance compared to present/absent ratings (Sutherland et al., 2013). Four STARS items are rated dichotomously as *insufficient* (0) or *good* (2), because they can

only be incomplete or complete (e.g., "T.H.R.E.A.T. tick boxes are completed"). The adherence total score is the sum of the 11 items, ranging from 0 to 22.

The scoring criteria for each response option (0, 1 or 2; see Table 2) are based on instructions in the START:AV user guide. In conjunction with the expert consensus, a research assistant reviewed the STARS items and their response criteria. Moreover, during test rounds, the first author and the research assistant independently rated and discussed multiple START:AVs with the STARS, which resulted in further fine-tuning of the criteria. The STARS was finalized when there were no more ambiguities and discrepancies in the test phase.

Step 2: deciding on how to use the STARS

Our aim was to design a concise and pragmatic adherence measure. Therefore, we constructed a scale that could be rated solely based on a review of the START:AV comprehensive rating forms. Completing the STARS for one form took 4.5 minutes on average (range = 4–5).

Step 3: assessing reliability of the STARS

Prior to applying the STARS, the interrater reliability (IRR) of the scale was assessed by two independent raters. The results reported in this section are based on 86 START:AV forms (28%) evaluated by the first author and the research assistant. Both raters were trained in using the START:AV and had experience completing the comprehensive rating forms and the STARS. Because all cases were coded by the same two raters who represented a larger group of raters of interest, a two-way random-effects model was deemed appropriate for calculating the intraclass correlation coefficient (ICC; Koo & Li, 2016). Furthermore, we were interested in the absolute agreement between the raters. Table 3 reports ICC values for both the "single rater" and the "average" (i.e., mean of multiple raters) measurement type, as well as their 95% confidence intervals. In the present setting, the STARS was designed to be completed by one assessor, therefore, the values of the "single rater" type were of interest. In addition, Koo and Li (2016) argue that it is more appropriate to interpret the confidence intervals rather than the ICC point estimates. Confidence intervals 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. Thus, we interpreted the confidence intervals of the single-rater absolute-agreement ICCs according to Koo and Li's guidelines: ICC < .50 = *poor*; .50–.75 = *moderate*; .75–.90 = *good*; > .90 = *excellent*.

All STARS items demonstrated good to excellent interrater reliability, except for STARS items 3 and 9 that reached moderate to good reliability. Still, STARS 3, which assesses whether all items are completed, had a percentage agreement of 94% between raters, and STARS 9, which assesses whether a rationale is provided for the risk estimates, reached 78%

Table 2. START:AV Adherence Rating Scale.

	Insufficient (0)	Sufficient (1)	Good (2)
Information Gathering			
• Administration Data: All requested information is completed: youth's name, date of birth, gender, assessment date, deadline, status, purpose and evaluator's name.	Information is incomplete or incorrect.	-	All required information is complete.
• Multiple Sources: Information is gathered through multiple sources.	Assessment is completed based on two or less file sources. Or used sources are not identified.	Assessment is completed based on multiple sources and/or collateral interviews. Youth is not interviewed.	Assessment is completed based on multiple sources, collateral interviews and interview with the youth.
START:AV items			
• Item Ratings: All strengths and vulnerabilities are rated (excl. case specific items and the item 'culture')	More than five strengths and/or five vulnerabilities are unrated.	Five strengths and/or five vulnerabilities are unrated.	Less than five strengths and/or five vulnerabilities are unrated.
• Item Rating Rationale: For all strengths and vulnerabilities rated as moderate and high, a written rationale is provided or relevant anchors are marked.	More than 20% of the items with a moderate or high rating have no rationale or marked anchors.	Between 1% and 20% of items with a moderate or high rating have no rationale or marked anchors.	All items with a moderate and high rating have a rationale or marked anchors.
• Key Strengths: At least two and maximum six key strengths are identified.	Less than 2 or more than 6 key strengths are selected.	2 key strengths are selected.	More than 2, but no more than 6 key strengths are selected.
• Critical Vulnerabilities: At least two and maximum six Critical vulnerabilities are identified.	Less than 2 or more than 6 critical vulnerabilities are selected.	2 critical vulnerabilities are selected.	More than 2, but no more than 6 critical vulnerabilities are selected.
START:AV adverse outcomes			
• History Rating: For each adverse outcome, history is rated in a checkbox and explained in a text box. In forms without an 'absent' option, empty checkboxes can signify 'not present' or missing information. This can be inferred from the context. If not, rate '0'.	History is not rated for all adverse outcomes. History is clearly present (e.g., text box), but not rated.	History is rated for each adverse outcome, however not every history rating is explained with details.	For all adverse outcomes, history is rated and details are provided.
• Risk Estimate: For each adverse outcome (incl. case specific if present), a risk estimate is provided.	Not all adverse outcomes have received a risk estimate.	-	For all adverse outcomes, a risk estimate is provided.
• Risk Estimate Rationale: For each moderate and high risk adverse outcome, a rationale for the risk estimate is provided.	A rationale is not provided for *all* moderate and high risk estimates.	A global description is provided for moderate and high risk estimates. (Selecting "type" does not count as a description)	For *each* moderate and high risk estimates, a rationale is provided by linking the items to the adverse outcome.

(*Continued*)

Table 2. (Continued).

	Insufficient (0)	Sufficient (1)	Good (2)
• T.H.R.E.A.T.: The tick boxes are completed.	Not all T.H.R.E.A.T.s are completed (Yes/No). A T.H.R.E.A.T. is rated but not all six conditions are met, or it is unclear.	-	All T.H.R.E.A.T.s are completed. When rated as present, all six conditions are met and illustrated.
• Periods of Stability: The periods of stability text box is completed.	The text box is left empty.	-	Evaluator mentions and explains potential periods of stability. If not present, this is indicated (n/a).

Table 3. Interrater reliability of the START:AV Adherence Rating Scale items and total score.

	Intraclass Correlation Coefficient			
Adherence item	Single Measure	95% CI	Average Measure	95% CI
1. Administration Data	.93	[.89,.95]	.96	[.94,.98]
2. Multiple Sources	.97	[.96, 98]	.99	[.98,.99]
3. Item Ratings	.75	[.64,.83]	.86	[.78,.91]
4. Item Rating Rationale	.84	[.76,.89]	.91	[.86,.94]
5. Key Strengths	.94	[.91,.96]	.97	[.95,.98]
6. Critical Vulnerabilities	.98	[.97,.99]	.99	[.99,.99]
7. History Rating	.89	[.83,.92]	.94	[.91,.96]
8. Risk Estimate	.92	[.88,.95]	.96	[.94,.97]
9. Risk Estimate Rationale	.81	[.72,.87]	.90	[.84,.93]
10. THREAT	.89	[.84,.93]	.94	[.91,.96]
11. Periods of Stability	1.00	[1.00, 1.00]	1.00	[1.00, 1.00]
Adherence Total Score	.96	[.93,.97]	.98	[.97,.99]

CI = confidence interval.

agreement. Further inspection of possible reasons for their lower reliability uncovered multiple cases of negligence on the part of the raters (e.g., miscalculating missing item ratings, overlooking missing risk rationales). After excluding these clear errors from the analysis, ICCs (single rater) improved substantially to 1.00 [1.00, 1.00] for STARS 3 and .93 [.89, .96] for STARS 9. The single-rater ICC for the total score did not change: .96 [.94, .98]. Thus, all STARS items demonstrated good to excellent reliability when carefully assessed.

Procedure and participants

Adherence to the START:AV user guide instructions was assessed for all START:AVs completed from February 2016 to February 2018 as part of the service's treatment process: 306 forms in total. All evaluators (i.e., treatment coordinators) that were employed within the facility at any point over the course of the study were included. Typically, the group of treatment coordinators consisted of 10 professionals, however, due to absences and staff turnover, 17 different evaluators completed the START:AV assessments. Two of the 17 evaluators were trainees and three were temporary replacements. All evaluators were female professionals with on average 5.5 years of service within the facility (range = 0–15). Fifty-six percent had prior experience (i.e., use and/or training) with another structured risk assessment instrument. As mentioned above, all evaluators received training in the START:AV by the first author. After the interrater reliability check, the adherence assessments with the STARS were completed by the first author.

Refresher workshop

In March 2017, thirteen months after the start of the START:AV implementation, the first author organized a refresher workshop for all 10 evaluators

who were employed at the time. Half of the invited evaluators (n = 5) attended the workshop, the other half was absent due to scheduling issues (e.g., maternity leave, sick leave, day off). Findings from the adherence evaluation served as input for the refresher workshop. For example, items that were frequently left empty were discussed, as well as how to rate items and risk estimates, and how to identify key and critical items. In addition, advice was given on how to use the information from the structured risk assessment for intervention planning. The workshop took 1.5 hours and concluded with a review of the instructions and institutional policy related to the START:AV.

Sample size analysis

Although this was a field study in which we relied on the START:AV forms that were available from clinical practice, a sample size analysis was conducted for the second hypothesis. Calculations were based on a statistical power of 80%, with α = .05 and one-tailed tests, assuming equal samples. The expected effect sizes were derived from the studies by Reynolds and Miles (2009; d = 2.08) and McNiel et al. (2008; d = .84). For the pre and post-workshop comparison, we calculated a required sample size of 4 in each group, thus, 8 cases total, whereas for comparison between the refresher and the comparison group, the required sample size was 20 cases in each group, thus, 40 in total. These are the required sample sizes for testing the difference between total scores, however, for the pre and post-workshop comparison, we also consider change in ratings for the separate adherence items. In Reynold and Miles, the smallest effect size found for a subcomponents was d = .54. Using this effect size, the required sample size for assessing change in the items pre versus post-refresher workshop is 46 STARS forms per group.

Analytic strategy

Adherence item and total scores were calculated for 306 forms. Five STARS total scores were identified as outliers, that is, these START:AV forms received a total score below 4.5, indicating an abnormal deviation from the other values in the sample. However, because they reflect relevant variations in adherence, these outliers were retained. As a consequence, central tendency is reported as median rather than mean, the interquartile range (*IQR;* the middle 50% of values) is reported as a measure of spread, and non-parametric tests were used. For the second hypothesis, we assessed whether the refresher workshop increased adherence among the START:AVs of those who attended, and whether the post-refresher total scores of the START:AVs of those who attended (i.e., refresher group) were significantly higher than the START:AVs of those who did not attend the refresher workshop (i.e., comparison group).The Kruskal-Wallis test is a non-parametric alternative to

a two-way analysis of variance for a between-subjects design. However, this test does not show which groups differ significantly and because no post-hoc tests exist for this situation, we conducted Mann-Whitney tests for each pair of groups. The Bonferroni correction was applied to control for multiple testing and a finding was considered statistically significant when $p < .0125$. All analyses were performed using IBM SPSS Statistics 25. For the calculation of the effect sizes for the non-parametric tests, an online calculator was used (Lenhard & Lenhard, 2016).

Ethical considerations

The Ethics Review Committee Psychology and Neuroscience (ERCPN) of Maastricht University approved the research protocol and procedures (ERCPN number 174_06_12_2016). In addition, the director of the youth care facility gave permission to conduct the research within the service. All data were analyzed anonymously and stored according to the university's Data Management Code of Conduct and the institution's data protection guidelines.

Results

Adherence to the START:AV features

The median adherence total score was 15 out of 22 (*IQR* = 5) with total scores ranging from 2 to 22. Table 4 presents the adherence score per STARS item.

Items 3 ("Item Ratings") and 8 ("Risk Estimate") were most often adhered to. More specifically, almost all forms were complete in terms of no more than four strengths and/or four vulnerabilities left blank (STARS 3) and in providing a risk estimate (STARS 8). Furthermore, for most item ratings, a rationale was provided by the evaluator (STARS 4: "Item Rating Rationale"). Still, 10% of the START: AV forms fell above the cutoff for missing item rationales (i.e., more than 20%). This cutoff is based on a benchmark for missing items used in risk assessment research (Sellers et al., 2017). With respect to providing a rationale for the risk estimates, only one in five forms included appropriate arguments for all adverse outcomes rated as moderate or high risk (STARS 9: "Risk Estimate Rationale"). In two-thirds of the assessments, the adolescent was interviewed as a source of information in addition to other sources (e.g., collateral information, files), which was considered good adherence. Thirty-seven forms (12%) relied solely on two or fewer file sources to complete the risk assessment (STARS 2: "Multiple Sources").

Item 11 ("Periods of Stability") was poorly adhered to: 82% did not include periods of stability on the START:AV form. Furthermore, items 1

Table 4. Percentage of START:AV forms in each adherence category and median for the adherence items.

Adherence Item	Insufficient %	Sufficient %	Good %	Median
1. Administration Data	33.3	-	66.7	2
2. Multiple Sources	12.1	21.2	66.7	2
3. Item Ratings	4.3	1.0	94.7	2
4. Item Rating Rationale	10.2	37.5	52.3	2
5. Key Strengths	34.6	21.3	44.1	1
6. Critical Vulnerabilities	29.4	9.8	60.8	2
7. History Rating	28.8	14.7	56.5	2
8. Risk Estimate	8.5	-	91.5	2
9. Risk Estimate Rationale	32.7	45.1	22.2	1
10. THREAT	33.3	-	66.7	2
11. Periods of Stability	82.0	-	18.0	0

Adherence items 1, 8, 10 and 11 have no "sufficient" category.

("Administration Data"), 5 ("Key Strengths"), 6 ("Critical Vulnerabilities"), 7 ("History Rating"), and 10 ("THREAT") were insufficiently adhered to in about one-third of the assessments. Specifically, we found that 33% of the START:AV forms had incomplete or incorrect administrative data (e.g., date, client's name, evaluator's name, purpose, etc.). Almost 30% of the forms did not follow the guideline to rate at least two and at most six critical vulnerabilities; this percentage was 35% for key strengths. Upon further inspection, we found that forms with an "insufficient" adherence rating typically had too few (< 2) key strengths and/or too many (> 6) critical vulnerabilities. With regard to the prior history of adverse outcomes, almost one-third of the forms had missing ratings and almost one in six forms did not provide details about the prior history when rated as present (STARS 7). Lastly, in one-third of the forms, THREAT ratings were missing or coded while not all conditions were met (STARS 10).

Adherence before and after the refresher workshop

The Kruskal-Wallis test demonstrated a statistically significant difference between the four groups in terms of adherence, $H(3) = 12.15$, $p = .007$, $d = .40$. The refresher workshop, consisting of five evaluators, completed 68 forms before and 66 forms after the workshop. The median adherence score for the refresher group was 15 ($IQR = 2.75$) prior to the workshop and 16 ($IQR = 4$) after the workshop. A Mann-Whitney test demonstrated that the adherence in the refresher group increased significantly after the workshop, $U = 1637.5$, $p = .006$, indicating a medium effect ($d = .48$). A post-hoc power analysis found a power of 85%. In addition, we assessed change in adherence for the individual STARS items. Most items showed an increase in adherence, while STARS 2 ("Multiple Sources") and 3 ("Item Ratings") decreased in adherence score, and 11 ("THREAT") did not change. However, change was

statistically significant only for STARS 1: we found increased adherence to completing administration data (U = 1695, p = .002, d = 0.43) after the workshop.

The comparison group, also consisting of five evaluators, completed 45 START:AVs before and 67 after the workshop. Their median adherence was 16 (IQR = 6) prior to the refresher workshop. This was not significantly different from the pre-workshop adherence score of the refresher group (U = 1440, p = .595). In the period after the workshop, the comparison group showed a median adherence score of 15 (IQR = 7) which was not significantly different from their pre-workshop score; U = 1257.5, p = .136. When examining change in adherence for the individual STARS items in the comparison group, we noticed a non-significant increase in adherence for STARS 6 ("Critical Vulnerabilities"), 8 ("Risk Estimate") and 9 (Risk Estimate Rationale), and a status quo for STARS 1 ("Administration Data"). All other items decreased in adherence. This decline was statistically significant for STARS 4: providing a rationale for item ratings (U = 1132, p = .015, d = .43), STARS 10: rating THREAT (U = 1201, p = .029, d = .35), and STARS 11: Periods of Stability (U = 1318, p = .050, d = .21).

The difference in post-refresher adherence total scores between the refresher group (Mdn = 16) and the comparison group (Mdn = 15) was statistically significant: U = 1493, p = .001, with a medium effect size (d = .58). A post-hoc power analysis revealed a statistical power of 95%.

Discussion

Adherence to structured risk assessment practice is defined as the extent to which a risk assessment instrument is completed as recommended by the instructions in the user guide. In general, adherence is important to ensure quality in research and practice. Yet, thus far, there have been limited documented efforts to measure adherence to specific structured risk assessment instruments (Reynolds & Miles, 2009; Sen et al., 2015). Our study addressed this gap and is the first study to assess adherence to a structured risk assessment instrument for youth. In particular, we developed and evaluated an adherence rating scale for the START:AV (i.e., the STARS). Before discussing the findings related to our hypotheses, we reflect on the scale's interrater reliability. We found that all STARS items could be reliably assessed by raters who were familiar with the START:AV. It was also demonstrated that disciplined attention is necessary when completing the STARS; evaluators easily overlooked information relevant to certain adherence items, resulting in inaccurate adherence scores.

Hypothesis 1: adherence varies across START:AV features

We hypothesized that features that required a rationale (i.e., STARS 4, 9 and 11) would be less adhered to than the other features that typically only required checking off a tick box. We found considerable variability in adherence to the individual STARS items. Our expectation was confirmed for STARS 9 ("Risk Estimate Rationale") and STARS 11 ("Periods of Stability"). The feature "periods of stability" demonstrated the lowest adherence rate. This might be due to the layout on the START:AV form: there is no "not applicable" tick box. Thus, an empty tick box could mean that evaluators did not consider the item or that they could not identify periods of stability. Feedback from the evaluators confirmed this: some did not consider the feature because they thought they were finished after completing the adverse outcomes, others indicated that there was often not enough file information to identify periods of stability in the past. Files mainly include accounts of the youth's misconduct and chaotic circumstances. Furthermore, evaluators indicated that the youth's behavior within the secure setting was often more stable than it had been in a long time.

Contrary to what we expected, relatively straightforward features such as "Administration Data", "History Rating", "THREAT", "Key Strengths", and "Critical Vulnerabilities" showed insufficient adherence in almost one-third of the assessed START:AVs. Because the administrative and historical information about an adolescent is readily available in the (electronic) patient file, these omissions seemed to reflect negligence. With respect to the missing THREATs, we learned that some evaluators were not familiar with the conditions for THREAT ratings, while others reasoned that it was not necessary to complete THREATs when the adverse outcome was rated as low risk. In terms of the key and critical items, we found that many evaluators identified too few key strengths and too many critical vulnerabilities according to the recommendations in the user guide. This finding may be explained by the serious problems experienced by the adolescents admitted to this secure setting. In fact, the service is a kind of "last resort" when other, less intensive and invasive interventions have turned out to be ineffective. The low number of strengths and high number of vulnerabilities identified by evaluators could therefore reflect the actual features of the service's caseload.

On a positive note, most START:AV items and all eight adverse outcomes were completed in the majority of START:AV forms. These features are arguably the most essential to the START:AV assessments and they were well adhered to. Furthermore, we observed that in two-thirds of the assessments an interview with the adolescent was used as a source of information. Overall, the adherence to the START:AV instructions was far from perfect, with an average adherence total score of 15/22 or 68%.

In comparison, staff members in Reynolds and Miles' study (2009) scored on average 10.44/12 (87%) for adherence to four broad components of the HCR-20. That is, they rated the presence of historical, clinical and risk management factors, and a risk management plan. The level of detail we required for adequate adherence may have resulted in lower adherence scores, compared to the more general adherence measure in the Reynolds and Miles study.

Hypothesis 2: a refresher workshop improves adherence

We found that evaluators who attended the refresher workshop showed a significant increase in adherence scores, from 15/22 (68%) to 16/22 (73%). These findings are in line with Reynolds and Miles' study (2009) in which the completeness of HCR-20s improved from 9.16/12 (76%) before training to 11.71/12 (98%) after training. In their study, staff had not received any prior training at baseline, which might explain the larger increase in adherence compared to our study in which staff members were trained prior to the refresher workshop.

When we focus on the individual adherence items, only STARS 1 "Administration Data" significantly increased. A longer, more comprehensive workshop might be needed to increase adherence for more features. Furthermore, the working mechanism behind improved adherence remains unclear. We do not know whether the effect was due to repetition, receiving feedback, the opportunity to ask questions, or other factors. Still, not participating in the refresher workshop was followed by a significant decrease in adherence on several features. In sum, the refresher workshop helped increase adherence to the risk assessment guideline, or at least stabilized adherence. Similar to McNiel et al. (2008), we found that START:AVs adhered significantly better to the guidelines when completed by staff members who had attended the refresher workshop compared to staff who had not participated.

Limitations and future directions

First, the items of the STARS were developed based on expert consensus after an in-depth review of the user guide, an approach that has been challenged by Mowbray et al. (2003). They indicate that expert opinions can change over time, that their predictive utility may be low, and that experts have a tendency to perceive the majority of features as important. Furthermore, the adherence items are based on instructions in the user guide, some of which cannot (yet) be considered as evidence-informed rules for risk assessment practice (e.g., preferable number of key and critical items). At present it is unclear from reliability and validity research on risk assessment

instruments which adherence features are essential (Hart & Logan, 2011). Ideally, we would select those features of a structured risk assessment instrument for an adherence measure that have a proven empirical association with the instrument's reliability and validity. However, as implementation science in structured risk assessment research stands today, this is unrealistic.

Second, the sample of evaluators in this field study changed considerably over the course of two years due to staff turnover and pregnancy replacements. This has prevented us from examining the effect of time on adherence. In a large-scale study, McHugo et al. (2007) found that adherence increased during the first 12 months after the start of the implementation, with little further gain over the next 12 months. Their longitudinal study examined adherence to five psychosocial evidence-based programs (e.g., supported employment) in 53 community mental health settings for two years. McHugo et al.'s significant effect of time on adherence was limited to the first year of implementation. In addition, they found that change in adherence depended on the intervention: for one intervention, adherence leveled off at six months, for another at one year, and for other interventions, adherence continued to increase over time. Such variety in adherence trajectories (decrease, increase, or fluctuation) has also been found in longitudinal studies on fidelity of health promotion programs, and might be the result of political, organizational and professional factors (Hoekstra et al., 2017). It would interesting to study these longitudinal effects on adherence to risk assessment instruments. This requires a large and more stable sample of evaluators, as well as a longitudinal design that can account for various potentially moderating variables, such as evaluator characteristics (e.g., demographics, risk assessment experience) and training characteristics (e.g., length, modality).

Third, research on the validity of the STARS was beyond the scope of this study. In future studies, convergent validity can be evaluated by comparing the findings on the STARS (based on file review) with other measures of adherence, for example, based on interviews with START:AV evaluators. In addition, the relationship between STARS scores and psychometric features of the START:AV can be examined. Are forms with a higher adherence more reliable? Do they have greater predictive power? Which adherence items have the strongest impact on the predictive value of the instrument? Does higher adherence to the START:AV instructions lead to more effective risk management strategies? To answer these questions, it is recommended to include a measure of quality of delivery (i.e., evaluator competence) in addition to the adherence measure (Breitenstein et al., 2010). Evaluators may be completing all features of the risk assessment instrument, but in such an inaccurate manner that it results in poor reliability or validity.

Lastly, there is currently no established threshold for adequate adherence in risk assessment practice. Thus, we cannot determine if the adherence levels

we obtained are "good enough". If future research reveals a significant association of adherence to predictive accuracy, this could lead to the establishment of minimum cutoff scores for adherence level. At present, in implementation research, there is no consensus about cutoff scores for fidelity. They are typically chosen based on face validity, for example, McHugo et al. (2007) used 4/5 (80%) or higher as a cutoff for high fidelity. Others have used ranges of total scores, such as the adherence labels created by De Vos et al. (2013): never (0%), seldom (1–33%), sometimes (34–66%), often (67–99%), and always (100%). A cutoff score (or range of scores) would help differentiate the adequate from the inadequate risk assessments.

Implications

This paper introduced a reliable measure to examine adherence to the START:AV user guidelines for assessing adolescents' risk for multiple adverse outcomes. The STARS can be used by practitioners and researchers to inform stakeholders about implementation quality. Users or developers of other risk assessment instruments are invited to apply this approach to their instrument of interest. Within the current setting, the adherence assessments were continued after the study, albeit not for all START:AV forms. Twice a year, the implementation coordinator completes the STARS for four START:AVs per evaluator, and communicates the results to the evaluators and their supervisor. Along with the obtained adherence scores, general advice for improving adherence is provided to all evaluators, and those with the lowest scores receive personalized feedback. Furthermore, the STARS is included in the local START:AV operations manual that contains protocols and policies relevant to the START:AV.

In concordance with other studies, the refresher workshop, although brief, proved advantageous, because evaluators who attended the refresher workshop subsequently showed more complete START:AVs. Refresher workshops might be imperative for sites that fail to reach adequate levels of adherence to risk assessment (McHugo et al., 2007). The content of these refresher workshops can be attuned to the shortcomings identified with an adherence scale. Future research should investigate the working mechanisms behind improving adherence, and the approach that works best (e.g., group sessions vs. personalized feedback).

Conclusion

Although the field of violence risk assessment is highly invested in designing and evaluating instruments, there is much to be gained in terms of assessing and securing adherence, a dimension of implementation fidelity. Moreover, adherence is just one component in the successful implementation of

a structured risk assessment instrument. We have previously addressed other implementation outcomes such as acceptability, adoption, appropriateness, feasibility, and penetration (De Beuf et al., 2019). In closing, we would like to reiterate the importance of applying the existing knowledge base of implementation science to risk assessment research and practice, to further improve the delivery of forensic mental health services.

Acknowledgments

The authors wish to acknowledge M. J. A. Gradussen, research assistant at CONRISQ group, Zetten, for her assistance in the development of the STARS.

Disclosure statement

The O.G. Heldring Institution publishes the Dutch START:AV User Guide and provides START:AV training. All proceeds go to research, without personal revenues for the authors.

ORCID

Tamara L. F. De Beuf http://orcid.org/0000-0001-5273-8523
Vivienne de Vogel http://orcid.org/0000-0001-7671-1675
Corine de Ruiter http://orcid.org/0000-0002-0135-9790

References

American Psychological Association. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

Breitenstein, S. M., Gross, D., Garvey, C., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health*, *33*(2), 164–173. https://doi.org/10.1002/nur.20373

Brown, B., Gude, W. T., Blakeman, T., van der Veer, S. N., Ivers, N., Francis, J. J., Lorencatto, F., Presseau, J., Peek, N., & Daker-White, G. (2019). Clinical performance feedback intervention theory (CP-FIT): A new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implementation Science*, *14*(1), 40–65. https://doi.org/10.1186/s13012-019-0883-5

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1), 40–49. https://doi.org/10.1186/1748-5908-2-40

De Beuf, T. L. F., de Vogel, V., & de Ruiter, C. (2019). Implementing the START:AV in a Dutch residential youth facility: Outcomes of success. *Translational Issues in Psychological Science*, *5*(2), 193–205. https://doi.org/10.1037/tps0000193

De Vos, A. J. B. M., Bakker, T. J., de Vreede, P. L., van Wijngaarden, J. D. H., Steyerberg, E. W., Mackenbach, J. P., & Nieboer, A. P. (2013). The prevention and reactivation care program: Intervention fidelity matters. *BMC Health Services Research*, *13*(1), 29–41. https://doi.org/10.1186/1472-6963-13-29

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20^{V3}: Assessing risk of violence – User guide*. Mental Health, Law, and Policy Institute, Simon Fraser University.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Fallon, L. M., Collier-Meek, M. A., Kurtz, K. D., & DeFouw, E. R. (2018). Emailed implementation support to promote treatment integrity: Comparing the effectiveness and acceptability of prompts and performance feedback. *Journal of School Psychology*, *68*, 113–128. https://doi.org/10.1016/j.jsp.2018.03.001

Gutheil, T. G., & Appelbaum, P. S. (2000). *Clinical handbook of psychiatry and the law* (3rd ed.). Lippincott Williams & Wilkins.

Hart, S. D., Douglas, K. S., & Guy, L. S. (2017). The structured professional judgement approach to violence risk assessment: Origins, nature, and advances. In D. P. Boer, A. R. Beech, T. Ward, L. A. Craig, M. Rettenberger, L. E. Marshall, & W. L. Marshall (Eds.), *The Wiley handbook on the theories, assessment, and treatment of sexual offending* (pp. 643–666). Wiley-Blackwell. https://doi.org/10.1002/9781118574003

Hart, S. D., & Logan, C. (2011). Formulation of violence risk using evidence-based assessments: The structured professional judgment approach. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation* (pp. 83–106). John Wiley & Sons. https://doi.org/10.1002/9781119977018

Hoekstra, F., van Offenbeek, M. A. G., Dekker, R., Hettinga, F. J., Hoekstra, T., van der Woude, L. H. V., & van der Schans, C. P., & ReSpAct group. (2017). Implementation fidelity trajectories of a health promotion program in multidisciplinary settings: Managing tensions in rehabilitation care. *Implementation Science*, *12*(1), 143–158. https://doi.10.1186/s13012-017-0667-8

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

Lenhard, W., & Lenhard, A. (2016). *Calculation of effect sizes*. Psychometrica. https://www.psychometrica.de/effect_size.html

Loy, V., Kwiatt, J., Dodda, A., Martin, E., Dua, A., & Saeian, K. (2016). Performance feedback improves compliance with quality measures. *American Journal of Medical Quality*, *31*(2), 118–124. https://doi.org/10.1177/1062860614556089

McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., Goldman, H. H., Lutz, W. J., & Finnerty, M. T. (2007). Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services*, *58*(10), 1279–1284. https://doi.org/10.1176/appi.ps.58.10.1279

McNiel, D. E., Chamberlain, J. R., Weaver, C. M., Hall, S. E., Fordwood, S. R., & Binder, R. L. (2008). Impact of clinical training on violence risk assessment. *American Journal of Psychiatry*, *165*(2), 199–200. https://doi.org/10.1176/appi.ajp.2007.06081396

McNiel, D. E., Hung, E. K., Cramer, R. J., Hall, S. E., & Binder, R. L. (2011). An approach to evaluating competence in assessing and managing violence risk. *Psychiatric Services*, *62*(1), 90–92. https://doi.org/10.1176/ps.62.1.pss6201_0090

Monahan, J. (1993). Limiting therapist exposure to Tarasoff liability: Guidelines for risk containment. *American Psychologist*, *48*(3), 242–250. https://doi.org/10.1037/0003-066X.48.3.242

Mowbray, C. T., Holter, M., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340. https://doi.org/10.1177/109821400302400303

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & 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*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Reynolds, K., & Miles, H. L. (2009). The effect of training on the quality of HCR-20 violence risk assessments in forensic secure services. *Journal of Forensic Psychiatry & Psychology*, *20*(3), 473–480. https://doi.org/10.1080/14789940802638366

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

Sen, P., Lindsey, S., Chatterjee, N., Rama-Iyer, R., & Picchioni, M. (2015). An audit of the quality of HCR-20 violence risk assessments in a low secure service. *Journal of Psychiatric Intensive Care*, *11*(S1), 1–9. https://doi.org/10.1017/S1742646415000096

Simon, R. I. (2003). Commentary: Think fast, act quickly, and document (maybe). *Journal of American Academy of Psychiatry & the Law*, *31*(1), 65–67. https://www.ncbi.nlm.nih.gov/pubmed/12817845

Sutherland, K. S., McLeod, B. D., Conroy, M. A., & Cox, J. R. (2013). Measuring implementation of evidence-based programs targeting young children at risk for emotional/behavioral disorders: Conceptual issues and recommendations. *Journal of Early Intervention*, *35*(2), 129–149. https://doi.org/10.1177/1053815113515025

Viljoen, J. L., Cochrane, D. M., & Jonnson, M. R. (2018). Do risk assessment tools help manage and reduce risk of reoffending? A systematic review. *Law and Human Behavior*, *42*(3), 181–214. https://doi.org/10.1037%2Flhb0000280

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., Gershenson, B. G., & McCabe, P. (2012). Does risk assessment make a difference? Results of implementing the SAVRY in juvenile probation. *Behavioral Sciences and the Law*, *30*(4), 384–405. https://doi.org/10.1002/bsl.2014

Vincent, G. M., Guy, L. S., Perrault, R. T., & Gershenson, B. (2016). Risk assessment matters, but only when implemented well: A multisite study in juvenile probation. *Law and Human Behavior*, *40*(6), 683–696. https://doi.org/10.1037/lhb0000214

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

Young, D., Moline, K., Farrell, J., & Bierie, D. (2006). Best implementation practices: Disseminating new assessment technologies in a juvenile justice agency. *Crime and Delinquency*, *52*(1), 135–158. https://doi.org/10.1177/0011128705281752