Psychological Assessment
2014, Vol. 26, No. 3, 990–1002

1040-3590/14/$12.00 http://dx.doi.org/10.1037/a0036794

Short-Term Assessment of Risk and Treatability (START): Systematic Review and Meta-Analysis

Laura E. O'Shea
St Andrew's Healthcare, Northampton, United Kingdom

Geoffrey L. Dickens
St Andrew's Healthcare, Northampton, United Kingdom, and University of Northampton

This article describes a systematic review of the psychometric properties of the Short-Term Assessment of Risk and Treatability (START) and a meta-analysis to assess its predictive efficacy for the 7 risk domains identified in the manual (violence to others, self-harm, suicide, substance abuse, victimization, unauthorized leave, and self-neglect) among institutionalized patients with mental disorder and/or personality disorder. Comprehensive terms were used to search 5 electronic databases up to January 2013. Additional articles were located by examining references lists and hand-searching. Twenty-three papers were selected to include in the narrative review of START's properties, whereas 9 studies involving 543 participants were included in the meta-analysis. Studies about the feasibility and utility of the tool had positive results but lacked comparators. START ratings demonstrated high internal consistency, interrater reliability, and convergent validity with other risk measures. There was a lack of information about the variability of START ratings over time. Its use in an intervention to reduce violence in forensic psychiatric outpatients was not better than standard care. START risk estimates demonstrated strong predictive validity for various aggressive outcomes and good predictive validity for self-harm. Predictive validity for self-neglect and victimization was no better than chance, whereas evidence for the remaining outcomes is derived from a single, small study. Only 3 of the studies included in the meta-analysis were rated to be at a low risk of bias. Future research should aim to investigate the predictive validity of the START for the full range of adverse outcomes, using well-designed methodologies, and validated outcome tools.

Keywords: Short-Term Assessment of Risk and Treatability (START), risk assessment, psychometric properties, predictive validity

Supplemental materials: http://dx.doi.org/10.1037/a0036794.supp

Historically, the assessment of risk in institutional psychiatric and correctional populations relied solely on either unstructured clinical opinion or on actuarial risk instruments that were mainly based on static historical information (Webster, Martin, Brink, Nicholls, & Desmarais, 2009). More recently, structured professional judgment schemes, such as the Historical, Clinical and Risk-Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997; Webster & Eaves, 1995), have gained traction among practitioners. Like actuarial tools, these schemes contain empirically derived risk factors together with extensive guidelines to facilitate accurate interpretation and scoring; however, they also retain the flexibility to allow raters to consider case-specific factors and promote professional discretion (Doyle & Dolan, 2002). This approach moves beyond risk prediction by encouraging an active process of risk management and prevention (Doyle & Dolan, 2002). It also promotes consideration of the dynamic changeable and/or fluctuating factors that contribute to risk in addition to more static historical factors (Webster et al., 2009).

Despite their popularity, structured judgment schemes have been criticized for their exclusive focus on the presence of factors that are associated with increased risk while ignoring the presence of protective factors that may diminish risk (Hart, 2001). There is, however, a lack of consensus in the literature about what constitutes a protective factor (Braithwaite, Charette, Crocker, & Reyes, 2010). For example, protective factors have been described variously as the absence of a risk factor (Costa, Jessor, & Turbin, 1999), as a separate factor with no corresponding risk factor (Farrington & Loeber, 2000), and as a factor lying at the opposing end of a continuum to a risk factor (Hawkins, Catalano, & Miller, 1992). Further, there is no empirical support for any particular theoretical model of how protective factors might operate to ameliorate risk (de Vogel et al., 2012). Nevertheless, the consideration of protective factors may have clinical advantages, including promotion of therapeutic relationships and identification of areas for personal growth (de Ruiter & Nicholls, 2011), and may also represent potential areas of intervention for inclusion in risk management and treatment plans (Nonstad et

This article was published Online First May 5, 2014.

Laura E. O'Shea, St Andrew's Academic Centre, King's College London Institute of Psychiatry, St Andrew's Healthcare, Northampton, United Kingdom; Geoffrey L. Dickens, St Andrew's Academic Centre, King's College London Institute of Psychiatry, St Andrew's Healthcare, Northampton, United Kingdom, and School of Health, University of Northampton.

Correspondence concerning this article should be addressed to Geoffrey L. Dickens, St Andrew's Academic Centre, Priory Cottage, Billing Road, Northampton NN1 5DG, United Kingdom. E-mail: gdickens@standrew.co.uk

al., 2010). Concurrently, consideration of protective factors may reduce the likelihood of a negative bias that can result in an overestimation of risk that contributes to unnecessary restriction and detention (de Ruiter & Nicholls, 2011). To date, a small number of structured clinical judgment tools have been developed that explicitly support the consideration of protective factors in risk assessment and that therefore may facilitate further examination of their role. A further criticism of many structured judgment schemes is their exclusive focus on risk outcomes related to aggression and violence. Adequate care and management of patients in secure settings requires the consideration of a broad range of clinical issues, including risk to self and risk from others (Webster et al., 2009).

The START

The Short-Term Assessment of Risk and Treatability (START; Webster et al., 2009) is one risk assessment tool that has attempted to combat the above criticisms and "address the needs of mentally and personality disordered clients in a more complete fashion than has been attempted in other structured professional guidelines" (Webster et al., 2009, p. 3). Raters are required to consider 20 dynamic items in terms of risk (termed *vulnerabilities*) and protective factors (termed *strengths*). The START authors define protective factors as "assets at the disposal of the individual (e.g., a supportive family), which become protective factors when the client makes use of them to reduce the risk" (Webster, Nicholls, Martin, Desmarais, & Brink, 2006, p. 756). The START's authors suggest that strengths and vulnerabilities can coexist in relation to each item (Webster et al., 2009). Reflecting this theoretical standpoint, each START item is scored separately in relation to both strengths and vulnerabilities on two unipolar 3-point scales, where 0 indicates no/minimal vulnerability or strength evident, 1 indicates moderate vulnerability/strength, and 2 indicates high vulnerability/strength. For example, a patient who abuses substances, but is seeking treatment and recognizes the consequences of addiction would warrant rating on both the strength and the vulnerability scale for the "substance abuse" item (Webster et al., 2006). The START also allows clinicians to identify any additional case-specific factors, critical vulnerabilities, key strengths, signature risk signs, and medical conditions an individual may hold. Finally, raters are required to make specific risk estimates (low, medium, or high) about the likelihood of each one of seven identified risk outcomes occurring: violence to others, self-harm, suicide, substance abuse, victimization, self-neglect, and unauthorized absence. There are few guidelines about how these estimates should be made, only that "reliance is placed not on the summed START vulnerability or strength scores but on the overall impression after all factors have been considered and taken into account" (Webster et al., 2009, p. 32). A rating of low risk indicates no or minimal risk, moderate indicates greater than average risk, and high indicates a relatively imminent and serious threat. When an urgent decision is needed and there is insufficient time for a thorough review of the evidence, clinicians are advised to make a dichotomous decision about whether there are Threats of Harm that are Real, Enactable, Acute, and Targeted (T.H.R.E.A.T). These risk estimates should be used to predict the likelihood of each outcome occurring over a maximum of 3 months. The START should then be repeated as it is intended as a measure of dynamic risk to predict short-term behaviors and document change over time (Webster et al., 2009).

Contribution of the Current Study

Research suggests that the START can be scored reliably (Desmarais, Nicholls, Wilson, & Brink, 2012; Nicholls, Brink, Desmarais, Webster, & Martin, 2006), and it has received high utility ratings among mental health professionals working in medium secure mental health units in England (Khiroya, Weaver, & Maden, 2009). Individual studies present evidence of its predictive ability for some outcomes (e.g., Braithwaite et al., 2010; Chu, Thomas, Ogloff, & Daffern, 2011; Desmarais, Nicholls, et al., 2012; Gray et al., 2011). However, to date there have been no systematic reviews of its psychometric properties, nor any systematic review and meta-analysis of its predictive validity. We, therefore, have conducted a systematic review to identify relevant studies; a narrative review of the psychometric properties, predictive validity, and user evaluations of the feasibility and utility of the START; and a meta-analysis to investigate the predictive efficacy of the seven specific risk estimates and the total Strengths and Vulnerabilities scores for each of the seven outcomes.

Method

Review Protocol

The current review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-analyses statement (Moher, Liberati, Tetzlaff, Altman, & The Prisma Group, 2009). This is a 27-item checklist to facilitate transparent reporting of results.

Search Strategy

The aim of our literature search was to identify all empirical studies of the START, including user evaluations and feasibility studies; investigations of its psychometric properties; and investigations of its predictive efficacy for any of the seven identified risk outcomes. All studies, including gray literature such as unpublished manuscripts and conference presentations, were eligible for inclusion. Multiple computerized databases (PsycINFO, Scopus, Web of Knowledge, Cochrane Library, and CINAHL) were searched for articles published prior to January 10, 2013. Search terms related to the START were paired with terms relating to psychometric properties and the seven risk outcomes identified by the START (see example in the Appendix). We used wild card terms (ending with an *) to return all permutations of each search term. Additional studies were located through personal correspondence and by hand searching the references lists of studies identified through the electronic search strategy. We subsequently included one study identified by hand searching that was published after January 10, 2013 but accepted for publication before that date.

Study Selection

The title and abstract of all articles returned by the search strategy were reviewed by the first author and the full text versions

of any that described an empirical investigation of the START were obtained. The first author of all unpublished studies was contacted on up to two occasions to obtain the full text. Eligibility of full-text articles was then independently assessed by both authors, resulting in a Cohen's kappa of .87, with any discrepancy (n = 4) resolved by discussion.

Inclusion and Exclusion Criteria

Narrative review of psychometric properties. In order to be included in the narrative review, the article must have contained an account of an original empirical study using the START. Studies were excluded from the narrative analysis if they did not contain original empirical work (i.e., reviews) and, due to difficulty in obtaining reliable translations, if they were written in languages other than English. Studies were also excluded if they used the adolescent version of the START (START: AV; Nicholls, Viljoen, Cruise, Desmarais, & Webster, 2010) or the previous version of the START (START [Version 1]; Webster, Martin, Brink, Nicholls, & Middleton, 2004), which did not comprise separate Strength and Vulnerability scales.

Meta-analysis of predictive efficacy. In addition to those criteria identified above, for a study to be included in the meta-analysis, it must have examined the predictive efficacy of the START Strength scale, Vulnerability scale, or specific risk estimates for one or more of the seven risk outcomes. Further, the area under the receiver operating characteristic curve (AUC) must have been presented, or the article must have contained sufficient information to allow calculation. Where studies contained overlapping samples, the study with the smallest sample size was excluded to avoid including the same participants twice.

Data Extraction

Narrative review. For each included study, we extracted results from investigations of (a) user evaluations of feasibility and utility of the START, (b) change over time, (c) internal consistency, (d) convergent validity, (e) correlations between subscales and risk estimates, (f) interrater reliability, (g) predictive validity, (h) incremental validity, and (i) effectiveness as an intervention.

Meta-analysis. For each study included in the meta-analysis, we extracted information regarding the number of participants, country of data collection, the START components that were used as predictors, type of risk outcomes measured, study setting and design, length of follow-up, the demographic and clinical characteristics of the sample, and the AUC value for each risk outcome examined.

Risk of Bias

Both the quality of individual studies and the methods used for study selection have the potential to introduce bias to the meta-analysis. The quality of primary studies included in the meta-analysis was assessed independently by both authors using the procedure adopted by Haney et al. (2012). This examined potential sources of bias at each stage of the study, including (a) clear definition and valid/reliable measurement of each outcome predicted; (b) clear definition of the study population and appropriate sampling; (c) validity and reliability of the START assessment including, independence from outcomes assessment; and (d) other confounders including a robust study design that establishes that risk assessment precedes outcomes measurement. For each study, each domain was rated as "yes," "unclear/unsure," or "no," and the overall risk of bias was coded as "low," "unclear," or "high" based on the authors' judgment of the likelihood that identified biases have lowered the confidence that can be placed on the results (see Table S1 in the online supplemental materials). This produced a weighted kappa of .72, and discrepancies (n = 6) were resolved through discussion. All studies were included in the meta-analysis irrespective of bias.

Meta-Analysis

All but one of the included studies (Gray et al., 2011) inverted scores for the Strength scale, such that higher scores represented less strength. We therefore inverted AUC values in the remaining study in order to pool the results. We classified the magnitude of AUC values in accordance with the boundaries suggested by Dolan and Doyle (2000), such that >.75 represents a large effect size; .50 represents chance prediction.

Effect sizes were extracted for each predictor and risk outcome combination reported by the individual studies. However, these were independent, as only one AUC value was included per sample in any given analysis. The predictor domains were categorized as Vulnerability total score, Strength total score, or one of the seven specific risk estimates. Each independent study contributed an effect size to the category of "any aggression." Four studies (Abidin et al., 2013; Gray et al., 2011; Nonstad et al., 2010; C. M. Wilson, Desmarais, Nicholls, Hart, & Brink, 2013) did not report a general criterion measure of aggression, but did report more specific categories of aggression. In these cases, the mean of all nonoverlapping effect sizes related to aggression was coded into an "any aggression" category. The more specific categories of aggression were also pooled for analysis such that the final outcome categories were "physical aggression against others," "physical aggression against objects," "verbal aggression," "any aggression," "self-harm," "suicidality," "unauthorized leave," "substance use," "self-neglect," and "victimization." Outcomes were defined and classified by the individual study authors; therefore, there may be some variation in the operational definitions used across studies. We subsequently excluded suicidality, unauthorized leave, and substance use from the analysis, as only one study had examined these outcomes. Additionally, investigation of the predictive efficacy of the violence risk estimate for physical aggression against objects was not possible, as only one study had investigated this combination.

The analysis was conducted following the procedure outlined by Guy, Douglas, and Hendry (2010). AUC values were weighted by their sample size and aggregated on the basis of a random-effects model using the MeanES SPSS macro (Lipsey & Wilson, 2001; D. B. Wilson, 2012). This macro is intended for use with any effect size and calculates the mean weighted effect size according to the procedures described by Hedges and Olkin (1985). The macro also computes 95% confidence intervals and conducts a homogeneity test using the Q statistic, which is distributed as a chi-square with

$k-1$ degrees of freedom (where k equals the number of independent effect sizes).

Results

Study Characteristics

In total, our search identified 160 records, of which 130 remained for review after the removal of duplicates (see Figure 1). Application of inclusion and exclusion criteria at the abstract level resulted in the exclusion of 93 studies, and we were unable to obtain three conference presentations despite repeated attempts at communication with the first author. Therefore, the full texts of 34 records were reviewed. Eleven were excluded from the narrative review; four did not include accounts of original empirical research, and seven reported research using the adolescent version or a previous version of the START.

Nine studies met criteria for inclusion in the meta-analysis (see Figure 1 for exclusion reasons). These comprised records representing seven nonoverlapping data sets and one pair of studies with overlapping samples, both of which were included, as they examined the predictive validity of different aspects of the START assessment. The total sample size was 543 (mean $N = 60$). Eight studies were journal articles published between 2010 and 2013, and one was an unpublished master's degree dissertation. Most studies ($n = 7$) were conducted in secure psychiatric settings, one in a civil psychiatric hospital, and one in both civil and secure settings. Studies were conducted in Canada ($n = 3$), Australia ($n = 2$), United Kingdom ($n = 2$), Ireland ($n = 1$), and Norway ($n = 1$). In two studies, risk assessments and outcome data were completed as part of routine clinical practice. One study used risk assessments that were completed by the clinical team, but trained researchers coded the outcome data. In the remaining six studies, risk assessments and outcome data were both completed by research teams. The length of follow-up period varied between 30 days and 1 year. Despite the START authors intending that the risk estimates should be used to predict outcomes for a maximum of 3 months, the mean length of follow-up period was 4.55 months (see Table S2 in the online supplemental materials for study characteristics).

Narrative Review

Feasibility/utility. Seven studies examined user evaluations of feasibility and utility. Overall, users were positive about the START. Most (81.6%) users agreed they had sufficient time to score the measure, which, on average, took between 25 and 40 min (Desmarais, Collins, Nicholls, & Brink, 2011). Quinn, Miles, and Kinane (2013) reported that the mean time to complete the START decreased on second and subsequent assessments, and, with monthly completion, mean time to complete the START after 18 months was 5 min. Between 88% and 100% of staff thought that the information required to score the START items was readily

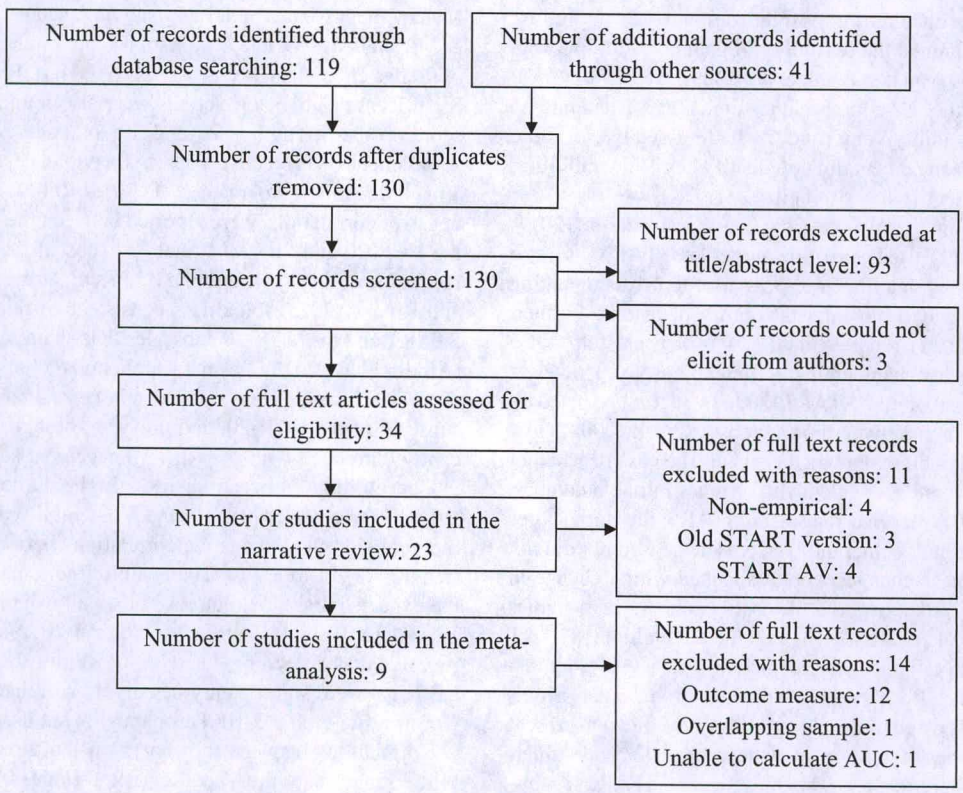

Figure 1. Flow diagram of literature search: Modified from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement flow diagram (Moher et al., 2009). START = Short-Term Assessment of Risk and Treatability; START: AV = START, adolescent version; AUC = area under the receiver operating characteristic curve.

available, and 70%–75% reported consulting more than one source of information. Between 74% and 86% of staff expressed no difficulties in completing the START Strength and Vulnerability items, signature risk signs, specific risk estimates, and risk formulations. However, Desmarais et al. (2011) reported that only 56% felt that it was easy to make the finer distinctions between high, moderate, or no/minimal evidence for each item (i.e., 0, 1, or 2).

Between 62% and 92.5% of users endorsed statements regarding the START's clinical utility. Highest endorsement was found for "key strength items," "focusing on strength and vulnerability items," and the "specific risk estimates" (Desmarais et al., 2011). The perceived utility of "signature risk signs" differed by profession; 81.2% of nurses reported they were useful compared with 37.5% of psychiatrists (Desmarais et al., 2011). Crocker et al. (2008) reported that the "T.H.R.E.A.T." and "Health Concerns/ Medical Tests" sections were not considered useful.

Confidence in ratings of START items and signature risk signs was generally high; 65%–95% of users were moderately to very confident. However, respondents reported that it was difficult to rate new patients (Crocker et al., 2011). Confidence was lowest (12.5%) for the completion of the specific risk estimates, with users reporting the classifications of low, medium, and high were too vague (Crocker et al., 2008; Doyle, Lewis, & Brisbane, 2008).

Users recommended that more emphasis was needed on the intervention section of the START, that night staff should be involved in the assessment process to provide a complete picture of patients, and that more, and iterated, training was necessary (Crocker et al., 2011). The recommended frequency of use varied from weekly to just once during hospitalization.

A final study examined the feasibility of patients' self-appraisal of the START items (van den Brink et al., 2009). Nearly all (98%) patients completed the self-appraisal questionnaire; 2% had missed two or more Vulnerability items, and 6% had missed two or more Strength items. Fifteen percent did not identify key Vulnerabilities, and 14% did not identify key Strengths.

Change over time. Two studies (Desmarais et al., 2011; Kroppan et al., 2011) used verbatim quotes of START users to demonstrate that they felt the tool was useful in documenting change over time, and could inform security decisions. Further, Desmarais et al. (2011) reported that 81.1% of users felt that a focus on dynamic, changeable items is clinically useful. However, despite the intention of the START authors that it be used to document change in risk over time (Webster et al., 2009), few studies have directly addressed whether this is the case or whether changes in rated risk are associated with changes in risk outcomes. Nonstad et al. (2010) reported that mean START Strength scores increased over time and Vulnerability scores decreased but did not examine whether these changes were associated with a change in risk outcomes, or predictive efficacy. In contrast, two studies (C. M. Wilson et al., 2010, 2013) conducted multiple START assessments and investigated the predictive efficacy of such assessments over time. Both studies examined group mean scores over time, which showed little change, but they did not present data about the number of participants whose strengths and vulnerabilities changed significantly. The most recent START assessment was not the most accurate at predicting violence in the immediate follow-up period for one third of comparisons, suggesting that if the scores have changed over time, then this is not always reflected by a change in risk outcomes.

Internal consistency. Seven studies included at least one of three different measures of internal consistency: Cronbach's alpha, mean interitem correlations (MICs), and mean-corrected item-total correlations (CITCs). Mean scores for each of the three measures indicated high internal consistency. Although high internal consistency does not necessarily indicate unidimensionality, a fact that is frequently confused in the literature (Gardner, 1995), it has been suggested that Cronbach's alpha values greater than .80, MIC values between .15 and .50, and CITCs greater than .30 reflect unidimensional scales (Clark & Watson, 1995; Nunnally & Bernstein, 1994). Cronbach's alpha for the Strength and Vulnerability scales ranged from .8 to .95 and from .76 to .95, respectively. MICs for the Strength scale were between .27 and .47; for the Vulnerability scale, they were between .26 and .40. Finally, CITCs for the Strength and Vulnerability scale ranged from .36 to .67 and from .31 to .63, respectively. Viljoen, Launeanu, Hendry, Nicholls, and Brink (2011) found that ratings made by mental health professionals were higher on all measures of internal consistency than those made by subject matter experts (those with training in forensic structured professional judgment and research experience with the START; S. Viljoen, personal communication, March 13, 2014).

Convergent validity. Three studies (Abidin et al., 2013; Desmarais, Nicholls, et al., 2012; Quinn et al., 2013) examined convergent validity between the START and other established risk assessment tools. Significant positive correlations were found between the Vulnerability scale total score and (a) the HCR-20 total and all subscale scores, (b) the Suicide Risk Assessment and Management Manual total and subscale scores (SRAMM; Bouch & Marshall, 2003), and (c) the total score on the screening version of the Psychopathy Checklist (PCL:SV; Hart, Hare, & Cox, 1995). Significant positive correlations were also found between the total score on the Strength scale and the total score on the Structured Assessment of Protective Factors for violence risk (SAPROF; de Vogel, de Ruiter, Bouman, & de Vries Robbé, 2012). Significant negative correlations were reported between the Strength scale and the HCR-20 and SRAMM total and subscale scores, and between the Vulnerability scale and the SAPROF. However, Quinn, Miles, and Kinane (2013) found less consistent correlations between the START and the HCR-20 subscales than the remaining two studies (Abidin et al., 2013; Desmarais, Nicholls, et al., 2012); neither of the START scales were significantly correlated with the historical subscale of the HCR-20, and only the Strength scale was significantly correlated with the Risk-Management subscale.

Correlations between scales. Eight studies examined correlations between the Strength and Vulnerability scales. All but one reported significant negative correlations between the scale totals (range = −.51 to −.95). The result of the remaining study was in the expected direction, but not statistically significant (Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012). Two studies also revealed that the Strength and Vulnerability scores for all individual items were significantly negatively correlated with each other (Braithwaite et al., 2010; Desmarais, Nicholls, et al., 2012).

Correlations between the scores on the Strength and Vulnerability scales and the specific risk estimates varied across the included studies. Significant associations were in the expected directions, such that Strength total scores were negatively correlated with risk estimates and Vulnerability scores positively correlated with risk estimates, with one exception: Gray et al. (2011)

reported a significant positive correlation between the total score on the Strength scale and the specific risk estimate for suicide. The most frequently reported significant associations were between the Strength and Vulnerability total scores, respectively, and the specific risk estimates for violence toward others, victimization, and unauthorized leave. No study reported a significant association between the Strength and Vulnerability total scores and the specific risk estimate for self-harm.

One study investigated correlations between each of the specific risk estimates; all were significantly correlated with one another apart from victimization with suicide and unauthorized leave with self-harm, suicide, and victimization (Nicholls, Petersen, Brink, & Webster, 2011).

Interrater reliability. Seven studies examined interrater reliability. Mean intraclass correlation coefficients (ICCs) were all in the good-excellent range. The highest agreement was observed for the Vulnerability scale, mean ICC of .86; the mean ICCs for the Strength scale and the specific risk estimates were .78 and .82, respectively. Viljoen, Launeanu et al. (2011) found that although the ICCs for Vulnerability ratings made by subject matter experts and mental health professionals were equivalent (.834 and .839), subject matter experts had substantially better agreement on Strength ratings than mental health professionals (.925 vs. .513). Abidin et al. (2013) reported Spearman's *r*, rather than an ICC; agreement for both the Strength (.69) and Vulnerability (.83) scales was in the moderate to strong range. Finally, van den Brink et al. (2009) examined the association between the Strength and Vulnerability items that were selected as most important by patients and case managers (key items). At a group level, the most and least frequently chosen items were similar. Four of the most frequently chosen key risks were the same for patients and case managers: "emotional state," "impulse control," "relationships," and "social support." For key strengths, only two of the most frequently chosen items were the same for patients and case managers: "occupational" and "insight." On an individual basis, however, mean kappas for Strength and Vulnerability items were .07 and .13, respectively, suggesting that patients and case managers disagreed about what the most important items were.

Predictive validity. Twelve studies examined the predictive validity of the START. Ten studies revealed that the Vulnerability scale significantly predicted various categories of inpatient aggression, including any aggression, physical aggression against others, verbal aggression, and physical aggression toward objects. However, Morris (2013) reported that the Vulnerability scale predicted verbal aggression and physical aggression toward objects, but not any aggression or physical aggression against others. Eight studies reported that the Strength scale significantly predicted the categories of aggressive behavior identified above. In contrast, Chu et al. (2011) found that the Strength scale significantly predicted any inpatient aggression and physical aggression toward others, but not verbal threat, whereas Morris (2013) reported that it did not predict any of the aggressive outcomes.

In terms of predicting the other risk outcomes identified in the START manual (Webster et al., 2009), Braithwaite et al. (2010) found that both the Strength and Vulnerability scales significantly predicted unauthorized leave and substance abuse. Gray et al. (2011) found that both scales significantly predicted self-neglect, but no study reported that either of the scales predicted self-harm, victimization, or suicidality. However, Abidin et al. (2013) reported that Strength and Vulnerability scores for the item "mental state" were significant predictors of self-harm.

Most (n = 8) of the included studies reported higher AUC values for the Vulnerability scale than the Strength scale for the majority of the examined outcomes, suggesting that the former is the stronger predictor. However, three studies reported higher AUC values for the prediction of physical aggression toward others for the Strength scale than the Vulnerability scale. Further, two studies revealed that the Strength scale remained a significant predictor over longer time periods (Crocker et al., 2008; C. M. Wilson, Desmarais, Nicholls, & Brink, 2010).

Evidence about the predictive efficacy of the specific risk estimates is mixed. Three studies revealed that the violence risk estimate significantly predicted any aggression or physical aggression toward others (Desmarais, Nicholls, et al., 2012; C. M. Wilson et al., 2010; C. M. Wilson et al., 2013), and three revealed that it did not (Chu et al., 2011; Gray et al., 2011; Morris, 2013). Two studies examined the predictive ability of the violence risk estimate for verbal aggression; both revealed significant results (Desmarais, Nicholls, et al., 2012; Gray et al., 2011).

Two out of three studies (Gray et al., 2011; Morris, 2013) reported that the specific risk estimate for self-harm was a statistically significant predictor. Gray et al. (2011) reported that specific risk estimates for self-neglect and victimization predicted their intended outcomes; however, Braithwaite et al. (2010) reported that they were not significant. Braithwaite et al. (2010) also examined the predictive efficacy of the remaining specific risk estimates: substance abuse, suicidality, and unauthorized leave. The substance use risk estimate was significant, but the others did not predict their intended outcomes.

Braithwaite et al. (2010) developed "optimized" Strength and Vulnerability scales using only items that were significantly associated with the occurrence of each risk outcome. The optimized Vulnerability scale significantly predicted six of the seven risk outcomes, the exception being self-harm. Similarly, the optimized Strength scale significantly predicted all risk outcomes except victimization.

Finally, two studies investigated the ability of the START to predict outcomes other than those intended by the START authors (Quinn et al., 2013; Viljoen, Nicholls, Greaves, de Ruiter, & Brink, 2011). Viljoen et al. (2011) found that Strength and Vulnerability scores significantly predicted nonreadmission following conditional discharge from a forensic psychiatric hospital and the receipt of an absolute discharge during the 3-year follow-up period. Quinn et al. (2013) examined the ability of the START to predict any aversive incident, as recorded on START report sheets and NHS Incident Reporting Information System forms. Strength and Vulnerability scores were both significantly predictive of this outcome over a period of 1 month, but neither were over a 3- or 6-month period.

Incremental validity. C. M. Wilson et al. (2013) found that the Vulnerability scale had incremental validity over the Historical scale of the HCR-20 but that the Strength scale did not. Desmarais, Nicholls et al. (2012) reported that the Strength scale had incremental validity over the Historical scale of the HCR-20 and the PCL:SV for the prediction of physical aggression against others, whereas the Vulnerability scale had incremental validity over the same measures for the prediction of any aggression and verbal aggression. Neither scale had incremental validity for the predic-

tion of aggression toward objects. The specific risk estimate for violence had incremental validity over the Strength and Vulnerability scores combined with the Historical scale of the HCR-20 or the PCL:SV for all four aggressive outcomes. Neither the Strength nor Vulnerability scale was found to have incremental validity over the other (C. M. Wilson et al., 2010).

START as an intervention to reduce risk behavior. Troquete et al. (2013) aimed to ascertain whether an intervention containing a structured risk assessment with the Dutch version of the START (t'Lam, Lancel, & Hildebrand, 2009) led to a reduction in violent or criminal incidents in the follow-up period compared with standard treatment. In the trial, forensic psychiatric outpatients (n = 632) were assigned to case managers (n = 58) who had been randomized to provide the intervention or treatment as usual. The study intervention arm involved risk assessment with the START at baseline by the case manager; patients also rated themselves on a client version of the START. Results were discussed by the case manager and client, with particular emphasis on identified key strengths and vulnerabilities. Subsequently, a treatment plan was devised to satisfy both parties. Outcome information about violent and criminal behaviors at baseline and follow-up for all patients was recorded routinely by case managers in each patient's case file; further information for a subset of patients in each study arm was collected at baseline and follow-up by interviewers blinded to intervention arm. Decisions about whether incidents constituted violent or criminal behavior was conducted by a panel of experts who were blind to the randomization. Mean follow-up period was 16.2 months (SD = 5.3, range = 6–38), and on this measure, there was no significant difference between control and intervention patients. Of the 310 patients in the intervention group, 201 (64.8%) received the intervention at least once (range = 1–6), including 72 (23.2%) who received it two or more times. Analysis revealed that incidents of violence or criminal behavior among patients in both groups reduced significantly between baseline and follow-up. Intention-to-treat analysis, including all participants in the intervention group irrespective of whether the intervention had been received, revealed no significant effect for patients in the intervention group to be more or less likely to have been involved in an incident at follow-up than control group patients. Further analysis of intervention group patients who had received any intervention, or those who had received multiple (2+) interventions as per the study protocol, did not alter the result.

Meta-Analysis

Results of individual studies. In total, 76 AUC values from the nine studies were included in the analysis; the smallest number of effect sizes contributed by a single study was three and the largest was 18. The magnitude of the AUC values ranged from .39 to .89 (see Table 1).

Mean effect sizes. The results of the meta-analysis are consistent with those of the narrative review. All weighted mean effect sizes (AUC_w) were significantly greater than chance (based on inspection of 95% confidence intervals), with the exception of (a) self-neglect as predicted by the Strength scale, Vulnerability scale, and the specific risk estimate for self-neglect; (b) victimization as predicted by the Strength scale, Vulnerability scale, and the risk estimate for victimization; and (c) self-harm as predicted by the Strength scale and the Vulnerability scale (see Table 2). Only the mean weighted effect sizes for the prediction of physical aggression against objects and verbal aggression by the Vulnerability scale and the risk estimate for violence reached the threshold for a large effect size (Dolan & Doyle, 2000). With the exception of physical aggression toward others, the Vulnerability scale produced larger mean weighted effect sizes than the Strength scale. In all cases, the specific risk estimates produced larger mean weighted effect sizes than the Strength and Vulnerability scores.

The largest mean weighted effect size obtained using the Strength scale was for physical aggression against others (.749), and the smallest was for victimization (.530). For the Vulnerability scale, the prediction of verbal aggression produced the largest mean weighted effect size (.777), and the smallest mean weighted effect size (.572) was obtained for self-harm. The violence risk estimate produced larger mean weighted effect size for physical aggression against others (.760) than for any aggression (.736).

Further analyses using the meta-analytic analog to the analysis of variance (MetaF; Lipsey & Wilson, 2001; D. B. Wilson, 2012) indicated that mean effect sizes did not significantly differ as a result of which predictor scale was used (Q_b [5] = 3.36, p = .644). However, they did differ significantly on the basis of the outcome being predicted (Q_b [6] = 16.27, p = .012); the largest mean effect size was obtained for the prediction of verbal aggression (.748), and the smallest was for the prediction of victimization (.570).

Risk of Bias

The quality assessment of the primary studies included in the meta-analysis indicated that only three had a low risk of bias, four had an unclear risk of bias, and one had a high risk of bias. The most common cause of potential bias was either having the same person collect the outcome data as completed the risk assessment or providing no description of assessor independence or blinding (see Table S1 in the online supplemental materials).

Discussion

The results of both the narrative review and meta-analysis provided mixed news about the psychometric properties of the START. User confidence ratings and evaluations about the utility of the START were largely positive, although users found it difficult to make fine distinctions between scores on items and specific risk estimates and felt they could benefit from further training. Interrater reliability was acceptable, and the measures used suggested the tool is internally consistent. Additionally, the START had convergent validity with other established assessments of risk and protective factors. These results suggest that the START is accepted by mental health professionals and can be scored reliably. Given that the START is intended to predict the risk of seven different outcomes, the finding of a high degree of internal consistency may be unexpected. However, the issue of internal consistency may well be irrelevant for risk assessment instruments that, like the START, are expressly designed to predict outcome rather than to measure some underlying construct that is related to risk (Rosenfeld & Penrod, 2011).

Most studies revealed significant negative correlations between the Strength and Vulnerability total scores, which has resulted in debate about the utility and necessity of rating items in terms of both strengths and vulnerabilities. However, examination of the

Table 1
Individual Study AUC Values

Study	*N*	Outcome	S	V	Violence RE	Self-harm RE	Victimization RE	Self-neglect RE
Abidin et al. (2013)	98	Any aggression	.71	.74	—	—	—	—
		Self-harm	.64	.65	—	—	—	—
Braithwaite et al. (2010)	34	Any aggression	.65	.66	.52	—	—	—
		Self-harm	.57	.58	—	.54	—	—
		Self-neglect	.53	.52	—	—	—	.55
		Victimization	.53	.55	—	—	.51	—
Chu et al. (2011)	50	Any aggression	.71	—	—	—	—	—
		Physical–others	.75	—	.69	—	—	—
		Verbal aggression	.64	—	—	—	—	—
Chu et al. (2013)	66	Any aggression	—	.74	—	—	—	—
		Physical–others	—	.75	—	—	—	—
		Verbal aggression	—	.79	—	—	—	—
Desmarais, Nicholls, et al. (2012)	120	Any aggression	.76	.79	.80	—	—	—
		Physical–others	.80	.77	.85	—	—	—
		Physical–objects	.77	.80	—	—	—	—
		Verbal aggression	.75	.79	.78	—	—	—
Gray et al. (2011)	44	Any aggression	.63	.63	.74	—	—	—
		Physical–others	.79	.68	.65	—	—	—
		Verbal aggression	.72	.74	.70	—	—	—
		Self-harm	.39	.48	—	.86	—	—
		Self-neglect	.66	.67	—	—	—	.80
		Victimization	.53	.60	—	—	.67	—
Morris (2013)	54	Any aggression	.61	.68	.64	.65	—	—
		Physical–others	.53	.55	.64	—	—	—
		Physical–objects	.57	.70	—	—	—	—
		Verbal aggression	.68	.76	—	—	—	—
		Self-harm	.45	.50	—	.77	—	—
Nonstad et al. (2010)	47	Any aggression	.77	.77	—	—	—	—
		Physica–others	.77	.77	—	—	—	—
C. M. Wilson et al. (2013)	30	Any aggression	.84	.82	.89	—	—	—
		Physical–others	.84	.82	.89	—	—	—

Note. AUC = area under the receiver operating curve; *N* = number of participants; S = Strength scale; V = Vulnerability scale; Violence RE = specific risk estimate for violence; Self-harm RE = specific risk estimate for self-harm; Victimization RE = specific risk estimate for victimization; Self-neglect RE = specific risk estimate for self-neglect. Dashes indicate that data were not reported in the original study.

correlations between individual items reveals considerable variation, suggesting that some items are more readily conceptualized on both dimensions than others. For example, Braithwaite et al. (2010) found that the correlation between the Strength and Vulnerability score for "relationships" was −.36, suggesting that an individual can have both supportive and damaging relationships, compared with −.87 for "substance use," which suggests individuals are unlikely to be rated as simultaneously holding both strengths and vulnerabilities in this domain. However, even if one concludes that the two scales are duplicating each other statistically, the consideration of strengths is considered clinically advantageous for their promotion of therapeutic relationships and facilitation of the development of risk management and treatment plans (de Ruiter & Nicholls, 2011; Nonstad et al., 2010).

There were also a large number of significant correlations between the various specific risk estimates, and between the risk estimates and Strength and Vulnerability total scores. It may be expected that people who are deemed at risk of exhibiting aggression and other challenging behaviors will be at risk in multiple domains (e.g., Hillbrand, 2001; Kooyman, Dean, Harvey, & Walsh, 2007; Nicholls et al., 2006). However, not all of the correlations were significant, and those that were significant were not always strong. This suggests that raters do not merely assign equal risk to all outcomes. Combined with evidence that the specific risk estimates have incremental validity over the Strength and Vulnerability scores, this suggests that the START is correctly being used as intended and providing a framework to guide decisions based on the overall clinical impression, as per guidance in the START manual (Webster et al., 2009).

Only one study has attempted to ascertain whether the START, as part of a targeted intervention, can reduce risk behavior to a greater extent than standard care with no structured risk assessment (Troquete et al., 2013). Structured judgment schemes like the START are not intended solely to predict the likelihood of risk, but also aim to facilitate treatment planning and management. In the case of the START, this includes the identification of risk and protective factors in order that they can be targeted with interventions that are aimed at their amelioration or amplification. Such targeted intervention should, if effective, result in reduced risk behavior. However, Troquete et al. (2013) found no evidence that an intervention involving the START was effective. Further, the study findings suggested that although there was poor fidelity to the study protocol, failure to implement treatment as planned could not solely account for its lack of effectiveness compared with standard care. This is an important finding and one that provides little positive evidence for the use of the START as an intervention

Table 2
Weighted Mean Effect Sizes of the START for the Prediction of Risk Outcomes

Outcome	*k*	*n*	AUC_w	95% CI_w	*Q*
Strength Scale					
Any aggression	8	477	.714	[.624, .803]	1.91
Physical–others	6	345	.749	[.643, .854]	3.25
Physical–objects	2	174	.696	[.506, .885]	1.49
Verbal	4	268	.710	[.591, .830]	0.49
Self-harm	4	230	.537	[.408, .667]	2.44
Self-neglect	2	78	.603	[.381, .825]	0.32
Victimization	2	78	.530	[.308, .751]	0.00
Vulnerability Scale					
Any aggression	8	493	.738	[.650, .826]	1.48
Physical–others	6	361	.727	[.624, .830]	2.39
Physical–objects	2	174	.769	[.620, .918]	0.37
Verbal	4	284	.777	[.660, .893]	0.11
Self-harm	4	230	.572	[.443, .701]	1.25
Self-neglect	2	78	.605	[.383, .827]	0.43
Victimization	2	78	.578	[.356, .800]	0.05
Violence risk estimate					
Any aggression	5	282	.736	[.619, .853]	3.29
Physical–others	5	298	.760	[.646, .873]	3.03
Verbal	2	164	.759	[.606, .912]	0.21
Self-harm risk estimate					
Self-harm	3	132	.741	[.568, .913]	2.04
Victimization risk estimate					
Victimization	2	78	.600	[.378, .822]	0.49
Self-neglect risk estimate					
Self-neglect	2	78	.688	[.445, .932]	1.19

Note. START = Short-Term Assessment of Risk and Treatability; *k* = number of effect sizes; *n* = number of participants, AUC_w = mean weighted area under the receiver operating curve; 95% CI_w = 95% confidence interval of mean weighted AUC; *Q* = homogeneity test.

for the reduction of violence and criminal behavior among forensic psychiatric outpatients. The START, however, has been largely developed and validated in institutional settings. Further, the trial only examined one outcome domain that the START aims to address (i.e., violence) and a range of more general criminal behaviors for which the authors of the START make no claims regarding the utility of their tool. Further research is required to test the START as an intervention for the full range of outcomes.

Both the narrative review and meta-analyses suggested stronger predictive ability of the Strength scores, Vulnerability scores, and specific risk estimates for various indices of aggression than for the other risk outcomes. Mean weighted effect sizes for aggressive outcomes for both scales were large, whereas for self-neglect and victimization, they were not significantly different from chance. Self-harm was significantly predicted by its specific risk estimate, but not by the Strength or Vulnerability score. There were insufficient studies to calculate mean weighted effect sizes for the prediction of suicide, unauthorized leave, and substance abuse, as only one study had investigated these outcomes. It is disappointing that the current evidence from meta-analysis suggests that five of the specific risk estimates either do not predict their intended outcomes or there is insufficient evidence in the current literature to establish that they do. Further, the only other available information regarding the predictive efficacy of the START for suicidality, unauthorized leave, and substance abuse suggests that the START may predict the latter two outcomes, but there is no evidence that it predicts suicidality (Braithwaite et al., 2010). Interestingly, although this study was small, and conducted among civil psychiatric inpatients, it was one of only three studies to be rated low in terms of potential bias.

Our finding that, for all the violence and self-harm outcomes, the specific risk estimates produced larger effect sizes than total scores for both scales is important because it is the approach advocated for use in clinical practice by the authors of the START (Webster et al., 2009) and proponents of the structured professional judgment approach on the whole (e.g., Douglas, Webster, Hart, Eaves, & Ogloff, 2001; Doyle & Dolan, 2002). Collapsing across predictors, the largest effect size was obtained for the prediction of verbal aggression, followed closely by physical aggression against others. This is promising, as physical aggression is generally regarded as having the most severe consequences to the victim (Nijman et al., 1999; Yudofsky, Silver, Jackson, Endicott, & Williams, 1986).

There are a number of reasons why the START may be better at predicting aggressive over other outcomes. First, there are a lack of established standardized measures for outcomes such as self-neglect and victimization (Nonstad et al., 2010) that may result in less reliable and consistent outcome data. The START Outcomes Scale (SOS; Nicholls et al., 2007), based on a modified version of the Overt Aggression Scale (OAS; Yudofsky et al., 1986), was designed to measure all the risk outcomes identified by the START, plus sexual aggression and stalking. There is evidence that the SOS can be reliably scored (ICC .70) (Nicholls et al., 2006); however, it has not been widely used (only used in four of the studies included in the current review). Second, it may be a result of using a measure designed to predict multiple outcomes; not all of the items will be of equal importance in predicting each of the risk outcomes. This idea is supported by the logistic regression analysis of Braithwaite et al. (2010) that led to the development of the optimized START scales. This may also help account for why, in most cases, superior predictive accuracy was obtained for the specific risk estimates, compared with the Strength and Vulnerability scores used actuarially, as clinicians can consider the items most pertinent to the outcome in question. Further, accurate prediction of the suicide outcome may be affected by its very low incidence (13.7 per 10,000 admissions; Powell, Geddes, Hawton, Deeks, & Goldacre, 2000) such that it is unlikely that there would be a single case in any of the studies included in the current analysis.

Limitations

One of the limitations of the meta-analysis is that it was not possible to fully investigate some of the unique factors of the START, such as the prediction of outcomes other than aggression and the ability to document change in risk over time due to the small number of studies examining these outcomes. Also, although the predictive validity of the Strength scale was investigated, only one study (C. M. Wilson et al., 2010) examined whether the consideration of Strengths was incremental in the prediction of risk outcomes. Although user ratings of the START were high, none of the included studies investigated how users rated the START in comparison with other risk assessment tools; therefore, it is not possible to conclude whether users think the START is an improvement over previous measures. Khiroya et al. (2009) did find

that the START had the highest utility rating out of 19 risk assessments, along with the Sexual Violence Risk-20 (Boer, Hart, Kropp, & Webster, 1997) and Stable 2000 (Hanson & Harris, 2001); however, this finding was not based on a direct comparison between tools, and only one unit reported using the START.

We could not obtain three unpublished records despite repeated attempted communication with authors; however, we did obtain eight unpublished documents that were considered for inclusion. A further source of potential bias comes from the exclusion of non-English language studies (Song et al., 2010). However, a recent study revealed no evidence of systematic bias from the use of restricted languages in meta-analyses (Morrison et al., 2012), suggesting any effect of excluding non-English language studies in the current analyses is minimal. The current study also revealed sources of potential bias in the literature, mainly due to nonindependent collection of outcome data from risk assessment data. It was not possible to investigate whether risk of bias moderated the mean weighted effect size as the homogeneity test revealed no significant heterogeneity in the AUC values contributed by individual studies. However, Cochran's Q has limited power to detect heterogeneity when there are a small number of studies (Heudo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006); therefore, results should be interpreted cautiously. Investigations have revealed that clinical, demographic, and methodological factors affect the performance of similar structured professional judgment schemes, such as the HCR-20 (e.g., O'Shea, Mitchell, Picchioni, & Dickens, 2013). It is reasonable to assume that the START may be similarly affected by these factors, and they should be investigated in future research. This is important as it will inform us about the extent to which we can trust its performance for an individual who is substantially different to the validation population. Further, an allegiance effect has been identified for other risk assessment tools, such that effect sizes are significantly larger in studies conducted by authors of the tools than in those conducted by independent researchers (Blair, Marcus, & Boccaccini, 2008).

Implications and Future Considerations

Results of the current analysis suggest that the START can be scored reliably and is a strong predictor of various indices of aggression and of self-harm but is not accurate in predicting self-neglect or victimization. However, these results are based on a small number of studies, and the predictive efficacy of the START for suicide, unauthorized leave, and substance abuse is relatively untested. Future research should determine whether differences in predictive accuracy based on outcome are a result of the predictive capabilities of the START itself, differences in the effectiveness of management of these outcomes, or a lack of standardized outcome measures. Other directions for research include investigations of whether the SOS can be scored reliably and is a valid measure of the intended risk outcomes; the development and validation of additional standardized outcome measures; establishing whether START risk factors vary over time, and whether any changes are associated with a change in risk outcomes; investigating the utility and feasibility of the START in comparison with other risk assessment tools; investigating how the predictive efficacy of the START may be moderated by methodological, demographic, and clinical characteristics; and further investigations of the predictive efficacy of the optimized scales identified by Braithwaite et al. (2010). Results from the one study to explore use of the START as an intervention are disappointing but do not preclude further research into its effectiveness in inpatient settings, with more rigorous protocol fidelity, and incorporating outcomes beyond criminality and violence. We therefore encourage researchers and clinicians to conduct further research into the predictive validity of the START and into its effectiveness as an intervention for the full range of adverse outcomes, using well-designed methodologies and validated outcome tools. These studies should be done in as wide a range of settings and with as diverse samples as possible.

References

References marked with one asterisk indicate studies included in the narrative review; references marked with two asterisks indicate studies included in both the narrative review and meta-analysis.

**Abidin, Z., Davoren, M., Naughton, L., Gibbons, O., Nulty, A., & Kennedy, H. G. (2013). Susceptibility (risk and protective) factors for in-patient violence and self-harm: Prospective study of structured professional judgement instruments START and SAPROF, DUNDRUM-3 and DUNDRUM-4 in forensic mental health services. *BMC Psychiatry, 13,* 197–214. doi:10.1186/1471-244X-13-197

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15,* 346–360. doi:10.1111/j.1468-2850.2008.00147.x

Boer, D., Hart, S., Kropp, P., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20.* Burnaby, British Columbia, Canada: Simon Fraser University, Mental Health, Law, and Policy Institute.

Bouch, J., & Marshall, J. J. (2003). *S-RAMM: Suicide Risk Assessment and Management Manual* (Research ed.). Glamorgan, Wales: The Cognitive Centre Foundation 2003.

**Braithwaite, E., Charette, Y., Crocker, A. G., & Reyes, A. (2010). The predictive validity of clinical ratings of the Short-Term Assessment of Risk and Treatability (START). *The International Journal of Forensic Mental Health, 9,* 271–281. doi:10.1080/14999013.2010.534378

**Chu, C. M., Thomas, S. D. M., Ogloff, J. R. P., & Daffern, M. (2011). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) in a secure forensic hospital: Risk factors and strengths. *The International Journal of Forensic Mental Health, 10,* 337–345. doi:10.1080/14999013.2011.629715

**Chu, C. M., Thomas, S. D. M., Ogloff, J. R. P., & Daffern, M. (2013). The short- to medium-term predictive accuracy of static and dynamic risk assessment measures in a secure forensic hospital. *Assessment, 20,* 230–241. doi:10.1177/1073191111418298

Clark, L., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. doi:10.1037/1040-3590.7.3.309

Costa, F. M., Jessor, R., & Turbin, M. S. (1999). Transition into adolescent problem drinking: The role of psychosocial risk and protective factors. *Journal of Studies on Alcohol, 60,* 480–490.

*Crocker, A. G., Braithwaite, E., Laferriere, D., Gagnon, D., Venegas, C., & Jenkins, T. (2011). START changing practice: Implementing a risk assessment and management tool in a civil psychiatric setting. *The International Journal of Forensic Mental Health, 10,* 13–28. doi:10.1080/14999013.2011.553146

*Crocker, A. G., Garcia, A., Israel, M., Hindle, Y., Gagnon, D., & Venegas, C. (2008). *Implementing and using a systematic risk assessment scheme to increase patient safety on a risk management unit for individuals with severe mental illness: A demonstration project.* Edmonton, Canada: Canadian Patient Safety Institute.

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *The International Journal of Forensic Mental Health, 10,* 160–170. doi:10.1080/14999013.2011.600602

*Desmarais, S. L., Collins, M., Nicholls, T., & Brink, J. (2011). *Perceptions of the Short-Term Assessment of Risk and Treatability (START) as implemented in forensic psychiatric practice*. Unpublished manuscript.

**Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24,* 685–700. doi:10.1037/a0026668

*Desmarais, S. L., Van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences and the Law, 30,* 448–469. doi:10.1002/bsl.2022

de Vogel, V., de Ruiter, C., Bouman, Y., & de Vries Robbé, M. (2012). *SAPROF. Guidelines for the assessment of protective factors for violence risk* (English version, 2nd ed.). Utrecht, the Netherlands: Forum Educatief.

Dolan, M., & Doyle, M. (2000). Violence risk prediction: Clinical and actuarial measures and the role of the psychopathy checklist. *The British Journal of Psychiatry, 177,* 303–311. doi:10.1192/bjp.177.4.303

Douglas, K. S., Webster, C. D., Hart, S. D., Eaves, D., & Ogloff, J. R. P. (Eds.). (2001). *HCR-20: Violence risk management companion guide*. Burnaby, British Columbia, Canada: Simon Fraser University, Mental Health, Law, and Policy Institute, University of South Florida, Department of Mental Health Law & Policy.

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing, 9,* 649–657. doi:10.1046/j.1365-2850.2002.00535.x

*Doyle, M., Lewis, G., & Brisbane, M. (2008). Implementing the Short-Term Assessment of Risk and Treatability (START) in a forensic mental health service. *Psychiatric Bulletin, 32,* 406–408. doi:10.1192/pb.bp.108.019794

Farrington, D. P., & Loeber, R. (2000). Epidemiology of juvenile violence. *Child and Adolescent Psychiatric Clinics of North America, 9,* 733–748.

Gardner, P. (1995). Measuring attitudes to science: Unidimensionality and internal consistency revisited. *Research in Science Education, 25,* 283–289. doi:10.1007/BF02357402

**Gray, N. S., Benson, R., Craig, R., Davies, H., Fitzgerald, S., Huckle, P., . . . Snowden, R. J. (2011). The Short-Term Assessment of Risk and Treatability (START): A prospective study of inpatient behavior. *The International Journal of Forensic Mental Health, 10,* 305–313. doi:10.1080/14999013.2011.631692

Guy, L. S., Douglas, K. S., & Hendry, M. C. (2010). The role of psychopathic personality disorder in violence risk assessments using the HCR-20. *Journal of Personality Disorders, 24,* 551–580. doi:10.1521/pedi.2010.24.5.551

Haney, E. M., O'Neil, M. E., Carson, S., Low, A., Peterson, K., Denneson, L. M., . . . Kansagara, D. (2012). *Suicide risk factors and risk assessment tools: A systematic review* (VA-ESP Project 05–225). Washington, DC: Department of Veterans Affairs.

Hanson, R. K., & Harris, A. J. (2001). A structured approach to evaluating change among sexual offenders. *Sex Abuse, 13,* 105–122.

Hart, S. D. (2001). Assessing and managing violence risk. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves & J. R. P. Ogloff (Eds.), *HCR-20 violence risk management companion guide* (pp. 13–26). Burnaby, British Columbia, Canada: Simon Fraser University, Mental Health, Law, and Policy Institute, and University of South Florida, Department of Mental Health Law & Policy.

Hart, S. D., Hare, R. D., & Cox, D. (1995). *The Hare Psychopathy Checklist: Screening version (PCL:SV)*. Toronto, Ontario, Canada: Multi-Health Systems.

Hawkins, J. D., Catalano, R. F., & Miller, J. Y. (1992). Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychological Bulletin, 112,* 64–105. doi:10.1037/0033-2909.112.1.64

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Heudo-Medina, T., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11,* 193–206.

Hillbrand, M. (2001). Homicide–suicide and other forms of co-occurring aggression against self and against others. *Professional Psychology: Research and Practice, 32,* 626–635. doi:10.1037/0735-7028.32.6.626

Khiroya, R., Weaver, T., & Maden, T. (2009). Use and perceived utility of structured violence risk assessments in English medium secure forensic units. *Psychiatric Bulletin, 33,* 129–132. doi:10.1192/pb.bp.108.019810

Kooyman, I., Dean, K., Harvey, S., & Walsh, E. (2007). Outcomes of public concern in schizophrenia. *The British Journal of Psychiatry, 191,* s29–s36. doi:10.1192/bjp.191.50.s29

*Kroppan, E., Nesset, M. B., Nonstad, K., Pedersen, T. W., Almvik, R., & Palmstierna, T. (2011). Implementation of the Short Term Assessment of Risk and Treatability (START) in a forensic high secure unit. *International Journal of Forensic Mental Health, 10,* 7–12. doi:10.1080/14999013.2011.552368

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis: Applied social research methods* (Vol. 49). Thousand Oaks, CA: Sage.

**Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLOS Medicine, 6,* 1–6. doi:10.1371/journal.pmed.1000097

**Morris, D. (2013). *The predictive validity of the Short Term Assessment of Risk and Treatability in an inpatient female forensic population* (Unpublished master's thesis). School of Psychology, University of Birmingham, Birmingham, United Kingdom.

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., . . . Rabb, D. (2012). The effect of English-language restriction on systematic review-based meta-analyses: A systematic review of empirical studies. *International Journal of Technology Assessment in Health Care, 28,* 138–144. doi:10.1017/S0266462312000086

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,* 313–327. doi:10.1177/1073191106290559

*Nicholls, T. L., Desmarais, S. L., & Brink, J. (2009, June). Implementation in a forensic psychiatric service. In T. Nicholls (Chair), *Implementation and evaluation of START in civil and forensic psychiatric services*. Symposium conducted at the 9th Annual Conference of the International Association of Forensic Mental Health Services, Edinburgh, Scotland.

Nicholls, T., Gagnon, N., Crocker, A., Brink, J., Desmarais, S., & Webster, C. (2007). *START Outcomes Scale (SOS)*. Vancouver, Canada: BC Mental Health & Addiction Services.

*Nicholls, T., Petersen, K. L., Brink, J., & Webster, C. (2011). A clinical and risk profile of forensic psychiatric patients: Treatment team STARTs in a Canadian service. *The International Journal of Forensic Mental Health, 10,* 187–199. doi:10.1080/14999013.2011.600234

Nicholls, T. L., Viljoen, J. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (2010). *Short-Term Assessment of Risk and Treatability: Adolescent Version (START: AV)* (Abbreviated manual). Coquitlam, British Columbia, Canada: Mental Health and Addiction Services.

Nijman, H. L. I., Muris, P., Merckelbach, H. L. G. J., Palmstierna, T., Wistedt, B., Vos, A., . . . Allertz, W. (1999). The Staff Observation Aggression Scale - Revised (SOAS-R). *Aggressive Behavior, 25,* 197–209. doi:10.1002/(SICI)1098-2337(1999)25:3<197::AID-AB4>3.0.CO;2-C

**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. *The International Journal of Forensic Mental Health, 9,* 294–299. doi:10.1080/14999013.2010.534958

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

O'Shea, L. E., Mitchell, A. E., Picchioni, M. M., & Dickens, G. L. (2013). Moderators of the predictive efficacy of the historical, clinical and risk management-20 for aggression in psychiatric facilities: Systematic review and meta-analysis. *Aggression and Violent Behavior, 18,* 255–270. doi:10.1016/j.avb.2012.11.016

Powell, J., Geddes, J., Hawton, K., Deeks, J., & Goldacre, M. (2000). Suicide in psychiatric hospital in-patients: Risk factors and their predictive power. *The British Journal of Psychiatry, 176,* 266–272. doi:10.1192/bjp.176.3.266

*Quinn, R., Miles, H., & Kinane, C. (2013). The Validity of the Short-Term Assessment of Risk and Treatability (START) in a UK medium secure forensic mental health service. *The International Journal of Forensic Mental Health, 12,* 215–224. doi:10.1080/14999013.2013.832714

Rosenfeld, B., & Penrod, S. D. (2011). *Research methods in forensic psychology*. Hoboken, NJ: John Wiley & Sons.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., & Sutton, A. J. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8). doi:10.3310/hta14080

t'Lam, K., Lancel, M., & Hildebrand, M. (2009). Manual for the Short Term Assessment of Risk and Treatability (START): Guidelines for assessment of short-term risks and treatment opportunities (Dutch translation). Drenthe, the Netherlands: GGZ.

*Troquete, N. A. C., van den Brink, R. H. S., Beintema, T., Mulder, T. W. D. P., van Os, R. A., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient foresnic psychiatry: Cluster randomised controlled trial. *British Journal of Psychiatry, 202,* 365–371. doi:10.1192/bjp.bp.112.113043

*van den Brink, R., Troquete, N., van Os, T., Schaafsma, G., Schram, A., & Wiersma, D. (2009). *Patient self-appraisal of the risk and protective factors of the START*. Paper presented at the 9th Annual Conference of the International Association of Forensic Mental Health Services, Edinburgh, United Kingdom.

*Viljoen, S., Launeanu, M., Hendry, M., Nicholls, T. L., & Brink, J. (2011). *Comparing ratings of subject matter experts and clinicians on the START in a forensic clinic*. Poster presented at the American Psychology-Law Society conference, Miami, FL.

*Viljoen, S., Nicholls, T., Greaves, C., De Ruiter, C., & Brink, J. (2011). Resilience and successful community reintegration among female forensic psychiatric patients: A preliminary investigation. *Behavioral Sciences and the Law, 29,* 752–770. doi:10.1002/bsl.1001

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

Webster, C. D., & Eaves, D. (1995). *The HCR-20 scheme: The assessment of dangerousness and risk*. Burnaby, British Columbia, Canada: Simon Fraser University, Mental Health, Law, and Policy Institute, and Forensic Psychiatric Services Commission of British Columbia.

Webster, C. D., Martin, M., Brink, J., Nicholls, T. L., & Desmarais, S. L. (2009). *Manual for the Short Term Assessment of Risk and Treatability (START)* (Version 1.1). Coquitlam, Canada: British Columbia Mental Health & Addiction Services.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Middleton, C. (2004). *Short-Term Assessment of Risk and Treatability (START)*. Port Coquitlam, British Columbia, Canada: St. Joseph's Healthcare Hamilton, Ontario, and Forensic Psychiatric Services Commission.

Webster, C. D., Nicholls, T. L., Martin, M. L., Desmarais, S. L., & Brink, J. (2006). Short-Term Assessment of Risk and Treatability (START): The case for a new structured professional judgment scheme. *Behavioral Sciences and the Law, 24,* 747–766. doi:10.1002/bsl.737

*Wilson, C. M., Desmarais, S. L., Nicholls, T. L., & Brink, J. (2010). The role of client strengths in assessments of violence risk using the Short-Term Assessment of Risk and Treatability (START). *The International Journal of Forensic Mental Health, 9,* 282–293. doi:10.1080/14999013.2010.534694

**Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior, 37,* 377–388. doi:10.1037/lhb0000025

Wilson, D. B. (2012). *Meta-analysis macros for SAS, SPSS, and STATA*. Retrieved from http://mason.gmu.due/~dwilsonb/ma.html

Yudofsky, S. C., Silver, J. M., Jackson, W., Endicott, J., & Williams, D. (1986). The Overt Aggression Scale for the objective rating of verbal and physical aggression. *The American Journal of Psychiatry, 143,* 35–39.

(*Appendix follows*)

Appendix

Example of Electronic Search Strategy: PsycINFO 10/01/13

Search term	Results
1) START AND WEBSTER	29
2) "short term assessment of risk and treatability"	19
3) "START assessment tool"	0
4) 1, 2 OR 3	42
5) Violen*	59,072
6) Agg*	80,862
7) Nonviolen*	2,313
8) Infract*	677
9) Physical*	187,028
10) Verbal*	99,097
11) Misconduct	1,568
12) Misbehav*	1,533
13) Assault	9,981
14) "Self-harm"	2,712
15) Suicide	35,495
16) Abscond*	150
17) Escap*	12,673
18) "Unauthorised leave"	1
19) Victim*	42,480
20) "Self-neglect"	262
21) "Substance abuse"	25,231
22) Reliab*	104,072
23) Valid*	164,476
24) Responsiv*	29,358
25) Psychometric	30,649
26) 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 OR 25	698,308
27) 4 AND 26	34

Received August 9, 2013
Revision received March 18, 2014
Accepted March 28, 2014 ■