Words: 7,648
Tables: 5

Running Head: PSYCHOMETRIC PROPERTIES OF THE SUICIDE STROOP TASK

Investigating the of the Suicide Stroop Task

Kelly M. Wilson[1]
Alexander J. Millner[2]
Randy P. Auerbach[3]
Catherine R. Glenn[4]
Jaclyn C. Kearns[4]
Olivia J. Kirtley[5]
Sadia Najmi[6]
Rory C. O'Connor[7]
Jeremy G. Stewart[8,9]
Christine B. Cha[1]

[1]Department of Counseling and Clinical Psychology, Teachers College, Columbia University [2]Department of Psychology, Harvard University [3]Department of Psychiatry, Columbia University, College of Physicians and Surgeons [4]Department of Clinical and Social Sciences in Psychology, University of Rochester [5]Center for Contextual Psychiatry, University of Leuven [6]Department of Psychology, San Diego State University Department of Psychiatry [7]Suicidal Behavior Research Laboratory, Institute of Health & Wellbeing, University of Glasgow [8]Center for Depression, Anxiety and Stress Research, McLean Hospital [9]Department of Psychiatry, Harvard Medical School

Correspondence to: Christine B. Cha, Ph.D., Department of Counseling & Clinical Psychology, Teachers College, Columbia University, 525 West 120th Street, Box 102, New York, NY 10027, e-mail: cbc2120@tc.columbia.edu, phone: +1 (212) 678-8212.

Abstract

Behavioral measures are increasingly used to assess suicidal thoughts and behaviors. Some measures, such as the Suicide Stroop task, have yielded mixed findings in the literature. An understudied feature of these behavioral measures has been their psychometric properties, which may affect the probability of detecting significant effects and reproducibility. In the largest investigation of its kind, we tested the internal consistency and concurrent validity of the Suicide Stroop Task, drawing from seven separate studies (*N*=875 participants, 64% female, aged 12 to 81 years). Results indicated that the most common Suicide Stroop scoring approach, interference scores, yielded unacceptably low internal consistency (*r*=-.09-.13) and failed to demonstrate concurrent validity. Internal consistency coefficients for mean reaction times (RT) to each stimulus-type ranged from *r*=.93-.94. All scoring approaches for suicide-related interference demonstrated poor classification accuracy, AUCs=.52-.56 indicating that scores performed near chance in their ability to classify suicide attempters from non-attempters. In the case of mean RTs, we did not find evidence for concurrent validity despite our excellent reliability findings, highlighting that reliability does not guarantee a measure is clinically useful. These results are discussed in the context of the wider implications for testing and reporting psychometric properties of behavioral measures in mental health research.

Investigating the Psychometric Properties of the Suicide Stroop Task

The prediction of suicidal thoughts and behaviors continues to challenge both researchers and clinicians. These life-threatening behaviors, which include suicide ideation, suicide attempts, and suicide deaths, are alarmingly common; 9.3 million adults report suicidal thoughts, 1.3 million adults report suicide attempt, and over 40,000 people die by suicide each year in the U.S. (Centers for Disease Control, 2013; Substance Abuse and Mental Health Services Administration, 2014). The addition of more objective behavioral measures to assess suicide risk may aid in the detection and prediction of suicidal thoughts and behaviors, particularly in clinical settings where patients may conceal suicidal intent or lack insight into their own risk states (Busch, Fawcett, & Jacobs, 2003). Behavioral measures also offer the possibility of capturing relevant cognitive processes that are outside of one's awareness, including transient suicide ideation (Cyders & Coskunpinar, 2011). Despite these strengths, behavioral measures have not been subjected to the same psychometric scrutiny as self-report measures, leaving issues of reliability and validity unaddressed (Green et al., 2016; Sharma, Markon, & Clark, 2013). The current study tackles these issues for the Suicide Stroop task by testing its psychometric properties.

The Suicide Stroop task has been tested as a potential suicide risk assessment tool among adults and adolescents (Cha, Najmi, Park, Finn, & Nock, 2010; Stewart, Glenn, Esposito, Cha, Nock, & Auerbach, 2017; Williams & Broadbent, 1986). It is adapted from the Emotional Stroop Task (EST), which measures the degree to which emotionally-valenced stimuli interfere with the

effortful process of suppressing prepotent responses (Cisler, Bacon, & Williams, 2009). ESTs have been used to capture psychopathology-specific emotional interference among several clinical populations, including those with depression and anxiety-related disorders (e.g., Mitterschiffthaler et al., 2008; McNally, Kaspi, Riemann, & Zeitlin, 1990; Phaf & Kan, 2007). While the specific cognitive processes assessed by ESTs has been a subject of ongoing debate (e.g., Algom et al., 2004, McKenna and Sharma, 2004, Williams et al., 1996), recent evidence suggests that increased response latency on ESTs among patient populations is driven by deficits in the top-down modulation of attention in the context of salient affective stimuli (Feng, Becker, Huang, Wu, Eickhoff, & Chen, 2018; Zimmerman et al., 2017). Within this framework, increased response latencies on the Suicide Stroop task reflect increased salience of suicide-related information, coupled with impairments in regulating attention away from emotional stimuli and toward the external environment (Kaiser et al., 2015). We refer to this increased response latency as *suicide-related attentional bias* to maintain continuity with prior theoretical and empirical literature. Suicide-specific attentional bias is conceptually linked with a cognitive model of suicide (Wenzel & Beck, 2008), wherein individuals with an activated suicide schema have difficulty disengaging from suicide-related thoughts due to an attentional fixation on suicide-relevant information. As such processing biases may accelerate suicidal crises (Wenzel & Beck, 2008), the Suicide Stroop task could be used as a unique source of data that cannot be assessed through subjective ratings, providing an important augmentation to suicide risk assessments and enhance our understanding of cognitive processes associated with the transition from suicidal thoughts to behavior.

Williams and Broadbent (1986) provided initial evidence for suicide-related attentional bias using a manual version of the Suicide Stroop task. They found that psychiatric inpatients

who had recently attempted suicide had greater response latencies for suicide-related words versus neutral words compared to psychiatric non-attempters and healthy controls—a finding replicated by Becker and colleagues (1999). This was eventually adapted to a computerized format and administered to recent suicide attempters in a psychiatric emergency department (Cha, Najmi, Park, Finn, & Nock, 2010). Findings revealed that suicide attempters displayed greater suicide-related attentional biases compared to non-attempter psychiatric controls at baseline and that biased attention predicted suicide attempts six months later. Since then, the computerized Suicide Stroop task has been used in a number of studies (e.g., Chung & Jeglic, 2016; Chung & Jeglic, 2017; Stewart et al., 2017); but to date, findings are mixed. Richard-Devantoy and colleagues' (2016) meta-analysis on computerized Suicide Stroop task studies concluded that there is evidence for suicide-related attentional bias among suicide attempters, but the effect size was small (Hedge's g=.22) and only one study (Cha et al., 2010) included in the meta-analysis found significantly larger suicide-related interference scores among suicide attempters compared to patient controls, suggesting this study alone accounted for significant overall effects.

There are several potential explanations for mixed results across Suicide Stroop task studies. First, the Suicide Stroop task may genuinely be unreliable. Testing its reliability via internal consistency estimates could help determine the degree to which task performance scores are affected by measurement error, which is necessary to accurately interpret inter-individual differences. Since prior studies have reported poor internal consistency for other adaptations of the EST (e.g., emotion faces EST, Brown et al., 2014; substance-related ESTs, Ataya et al., 2012), here we test the internal consistency of the Suicide Stroop task across and within multiple samples.

Second, Suicide Stroop task studies vary substantially in their methodology. As a consequence, no matter how reliable the suicide Stroop is, variable findings across studies are to be expected. Several examples of inconsistent methodological features exist. First, studies have used different scoring approaches for the Suicide Stroop task. ESTs are typically scored based on the difference in reaction time (RT) for emotional words relative to neutral words across trials (i.e., *interference score*; used in Cha et al., 2010). However, other studies used scoring approaches such as extracting RT to the word 'suicide' only (disregarding RT to other words in the 'suicide' category) to calculate an interference score (Chung & Jeglic, 2016, 2017), dividing the interference score by neutral word RT (Stewart et al., 2017), or used mean RTs (Becker et al., 1999). As different scoring approaches may artificially increase the proportion of error variance in reliability calculations due to methodological artifacts (e.g., due to correlation between the two component variables in difference scores), testing reliability across these approaches could explain the emergence of inconsistent findings.

Finally, Suicide Stroop studies have compared distinct clinical groups. Although most studies have compared suicide attempters to psychiatric patients with no suicide attempt history (i.e., regardless of suicide ideation status; Williams & Broadbent, 1986; Becker et al., 1999; Cha et al., 2010; Richard-Devantoy et al., 2016), other studies compared a broad suicidal group (i.e., including history of suicide ideation, plan or attempt) to a non-suicidal group (Chung & Jeglic, 2016; Chung & Jeglic, 2017) or suicide attempters to suicide ideators (Stewart et al., 2017).

In the present study, we pursue two aims pertaining to reliability and validity of the Suicide Stroop task. First, we aim to test internal consistency of the Suicide Stroop task. Second, we aim to test concurrent validity of the Suicide Stroop task. Specifically, we test whether individuals with a history of suicide attempts ("attempters") show suicide-related attentional bias

(above and beyond deficits for other emotion conditions), compared to participants without a history of suicide attempts and, in a second analysis, to those with a history of suicide ideation (i.e., "ideators") and those without a history suicide ideation (i.e., "non-suicidal controls"). Groups are defined by distinct suicide-related behaviors in order to clarify how suicide-related attentional bias may characterize different stages on a pathway from suicidal thoughts to behavior. To further explore the concurrent validity of the Suicide Stroop task, we also estimate the sensitivity and specificity of task scores using receiver operating characteristic (ROC) analysis. Given the aforementioned methodological inconsistencies across studies, we examine reliability and concurrent validity for three different scoring approaches. We also examine these outcomes within different clinical presentations of study samples (psychiatric vs. community-based) and among adults and adolescents.

Method

Sample

The sample was drawn from seven separate studies (n=875) that administered the Suicide Stroop task, the *Self-Injurious Thoughts and Behaviors Interview* (SITBI; Nock, Holmberg, Photos, & Michel, 2007), and the *Beck Scale for Suicidal Ideation* (SSI; Beck, Kovacs, & Weissman, 1979) at the same time point. Study participants included adults and adolescents (M=27.17 years, SD=13.47), of whom 63.9% were Female and 63.0% White, 6.6% Hispanic, 6.3% Asian, 6.2% Black/African-American, and 17.7% Other.[1] They were recruited from surrounding communities (i.e., community sample)[2], and psychiatric emergency department

[1] Demographic survey for Study 6 allowed participants the option of identifying by nationality, which we opted to classify as 'Other' when no other race or ethnicity was reported.
[2] Community-based sample had an average age of M=28.39 years, SD= 14.93, were 61.4% Female and 57.2% White, 6.9% Black/African-American, 8.4% Hispanic, 5.4% Asian, and 21.8% Biracial, Other, or identified by nationality.

(ED) or inpatient settings (i.e., psychiatric sample)[3]. Eligible participants were fluent in English. Exclusion criteria included inability to provide informed consent or assent and presence of gross cognitive impairment. Sample characteristics for each study are summarized in Table 1.

Materials and Procedure

Each study included distinct aims and procedures, with some studies focusing on assessment (Studies 1, 3, 5, 6, and 7) and others on interventions (Studies 2 and 4). Intervention study data were used from the baseline assessment (i.e., pre-intervention). Measures were administered in a several different settings, ranging from a university laboratory (Studies 3-7), to an interview room located within a psychiatric inpatient unit (Study 2), to a psychiatric ED (Study 1). Please see individual studies for information on study-specific procedures [REMOVED IDENTIFYING CITATIONS TO MAINTAIN BLIND].

Suicide Stroop Task. In the Suicide Stroop Task, participants were instructed to name the font color of suicide-related, emotionally-valenced (i.e., positive, negative), and neutral words as quickly and as accurately as they could. Stimuli for the task were presented and reaction times recorded using Empirisoft DirectRT v2004 software (Jarvis, 2004) or SuperLab 4.5.1 (Cedrus Corp, 2011). Instructions were presented on the screen at the beginning of the task. Each trial started with a blank white screen for 4 seconds followed by the presentation of a centered "+" for 1 second, another blank screen for 1 second, and then the word presented in red or blue font. The word remained on the screen until a response was recorded. Participants indicated the color of the word by pressing the red or blue key on the computer keyboard. Participants completed eight practice trials featuring neutral words, followed by 48 critical trials including suicide-related words (e.g., *suicide, dead, funeral*), negatively-valenced words (e.g.,

[3] Psychiatric sample had an average age of *M*= 20.16 years, *SD*= 11.72, were 69.2% Female and 77.7% White, 4.3% Black/African-American, 1.9% Hispanic, 8.5% Asian, and 7.6% Biracial or Other.

alone, rejected, stupid), positively-valenced words (e.g., *happy, success, pleasure*) and neutral words (e.g., *paper, engine, museum*).

Participants' RTs to identify the color of each word were recorded on each trial. Raw data from each study was obtained and analyzed together. Consistent with prior scoring procedures (e.g., Cha et al., 2010; Chung & Jeglic, 2016, 2017; Stewart et al., 2017), only trials with correct responses were included in the analysis, and participants with an error rate greater than 2 standard deviations above the mean error rate for the entire sample were excluded from analysis. Individual trials with RTs ±2 standard deviations from each participant's mean RT were eliminated, and participants for whom the mean RT across all trials was ±2 standard deviations from the mean RT for the sample were also excluded.

Three scoring approaches using these RTs were employed: *Mean RTs*, *Interference Scores*, and *Ratio Scores*. Mean RTs represent the averaged RTs untransformed for each valence type. Averaging each participant's raw RTs yielded Mean RTs for suicide-related words (Mean RT_{Sui}), negative words (Mean RT_{Neg}), positive words (Mean RT_{Pos}), and neutral words (Mean RT_{Neu}). Interference Scores were calculated by subtracting each participant's mean RT for neutral words from their mean RT for suicide-related words ($Interference_{Sui}$), negative words ($Interference_{Neg}$), or positive words ($Interference_{Pos}$). We also computed an Interference Score that consisted of RTs only from the presentation of the word 'suicide' (i.e., not including RTs to the presentation of the words 'dead' and 'funeral' in that category) subtracted from the RT for all neutral words ($Interference_{SuiWord}$), adapted from Chung and Jeglic (2016). Finally, we computed Ratio Scores by further dividing each interference score by the mean RT for neutral words (Stewart et al., 2017). This yielded three Ratio Scores: one for suicide-related words ($Ratio_{Sui}$),

negative words ($Ratio_{Neg}$), and positive words ($Ratio_{Pos}$). See Table 3 for an outline of scoring approaches.

Suicidal Thoughts and Behaviors. The *Self-Injurious Thoughts and Behaviors Interview* (SITBI; Nock et al., 2007) is a structured interview that assesses the presence, frequency, and characteristics of a range of self-injurious thoughts and behaviors. The SITBI has demonstrated strong interrater reliability (average κ =.99, r = 1.0) and test-retest reliability (average κ = .70, intraclass correlation coefficient = .44; Nock et al., 2007). The SITBI was administered to all participants; based on the SITBI data, we coded participants as either lifetime suicide attempters, lifetime suicide ideators (i.e., who have had thoughts of suicide, but never made an attempt), or non-suicidal control participants (i.e., no lifetime history of suicide attempt or suicide ideation). Of note, the two studies reported in [REMOVED IDENTIFYING CITATIONS TO MAINTAIN BLIND] involved the administration of the SITBI through a brief, self-report format.

The *Beck Scale for Suicidal Ideation* (Beck SSI; Beck, Kovacs, & Weissman, 1979) is a 21-item self-report measure that assesses severity of current suicide ideation. The first 19 items are summed to yield a total score ranging from 0-38. The final two items assess the number of previous suicide attempts and desire to die during the most recent suicide attempt and are not included in the total score. Cronbach's alpha for SSI scores was excellent (α=0.95).

Data Analysis

In pursuit of our first aim, we calculated split-half reliability with Spearman-Brown correction. Individual trial RTs for the task were divided into odd-even trials, and two Stroop scores were created and correlated to calculate reliability coefficients for each valence (i.e. suicide, negative, positive and neutral words) and scoring-type (i.e., Mean RTs, Interference,

Ratio). Reliability coefficients were calculated across participants in the full sample (*n*=875), as well as within samples recruited from the community (*n*=340) or psychiatric treatment settings (*n*=535), and across adults (*n*=552) and adolescents (*n*=323).

Aligned with our second aim, we performed mixed design analyses of variance (ANOVAs). We conducted multiple *Group* x *Valence* ANOVAs, but each factor varied depending on the selected experimental groups and scoring approach. For the between-subjects factor, *Group* comparisons included suicide attempters versus non-attempters or suicide attempters vs. suicide ideators vs. non-suicidal controls. For the within-subjects factor, *Valence* had 4 levels in Mean RT analyses (i.e., suicide-related, negative, positive, neutral), and when testing Interference and Ratio Scores there were 3 levels (neutral mean RT was omitted because it was subtracted from the mean RT of the other valences). For these ANOVAs, we estimated a *Group* x *Valence* interaction to determine whether suicide attempters and ideators would show significantly larger RT interference, or ratio scores for suicide-related words relative to the other valences categories (i.e., to test whether participants with suicidal thoughts or behaviors demonstrated attentional bias for suicide-related stimuli above and beyond deficits for other emotion words), compared to participants with no suicidal thoughts or behaviors or ideation only. We conducted post-hoc tests with Holm-Bonferroni corrections (Holm, 1979). Analyses with the different groups and scoring approaches were explored within the full sample, within community-based and psychiatric subsamples, and among adults and adolescent subsamples. To test classification accuracy, Suicide Stroop interference scores were converted to a categorical variable representing positive scores (i.e., longer latencies for suicide-related words relative to neutral words) and negative scores (i.e., longer latencies for neutral relative to suicide-related words). ROC curve analysis was used to estimate the classification accuracy for both the

continuous Suicide Stroop interference scores and the categorical Stroop score variables in differentiating suicide attempters from non-attempters. For a detailed account of these results as well as those based on analyses within subsamples, alternative analytic approaches, and alternative Suicide Stroop scoring approaches, see Supplemental Materials.

Results

Reliability

Across all studies (n=875), Mean RTs for each word valence demonstrated excellent reliability (range=0.93-0.94). For the Interference and Ratio scores, reliability was unacceptably low and near zero (range=-0.09-0.13). This pattern, wherein Mean RT scores demonstrated good reliability and difference score-based (i.e., Interference and Ratio) reliability was poor, remained the same when testing reliability by clinical sample (community vs. psychiatric) and among adults vs. adolescents (Table 4).

Concurrent Validity

To examine concurrent validity, we compared suicide-related attentional bias between suicide attempters and non-attempters, as well as between suicide attempters and suicide ideators and controls.[4] *Group* x *Valence* interactions were not significant when testing Mean RT, Interference, Ratio or scoring for the two-group (attempters vs. non-attempters) or three-group comparisons (attempters vs. ideators vs. controls), ps=0.56-0.88, η_p^2s<0.01. Results were similar within the aforementioned subsamples (community vs. psychiatric; adults vs. adolescents), with suicide attempters failing to demonstrate a significant suicide-related interference effect. For Mean RTs, there was a main effect of *Group* for two-group, $F(1,873)=9.96$, $p<.01$, $\eta_p^2=0.01$, and

[4] ANCOVA analyses were also conducted to account for potential age effects, as RT follows a strong developmental trajectory (Deary & Der, 2005). Controlling for the effects of age had no effects on the statistical significance of tests for any between group comparisons. We also tested for possible gender differences in task performance using ANCOVA and testing ANOVA within male and female participants, which yielded no significant interactions.

three-group comparisons, $F(2,872)=5.72$, $p<.01$, $\eta_p^2=0.01$. Post-hoc Holm-Bonferroni tests across the two separate analyses revealed that suicide attempters showed slower overall RTs compared with non-attempters ($p<0.01$, $d=0.21$), as well as with suicide ideators ($p<0.01$, $d=0.25$), and controls ($p=0.02$, $d=0.15$,); however, ideators and controls did not differ in overall RT ($p=.18$, $d=0.11$). When comparing groups within the psychiatric or adult subsamples, attempters continued to show significantly slower RT compared to non-attempters (ps<0.01, ds ~ .30); however this was not the case within the community or adolescent subsamples (ps=0.46-0.80, ds < .05). For three-group comparisons within Mean RT, Interference and Ratio scoring approaches, there were main effects of *Valence*, Fs=2.86-4.60, ps=<0.01-0.03, η_p^2=0.003-0.01, but post-hoc tests revealed no significant differences across valences via each scoring approach after correcting for multiple comparisons. The exception was for Mean RT among community samples, where participants responded more slowly to neutral words compared to suicide words ($p<.01$, $d = .06$).

We conducted additional analyses that: (1) compared Suicide Stroop scores among participants with active suicide ideation in the past week (those who endorsed item 4 of the SSI, "current ideators") to scores among those with no current or past history of suicide ideation; and (2) tested for the potential association between Suicide Stroop scores and severity of suicide ideation among current ideators (i.e., SSI total score). Results from ANOVAs comparing Suicide Stroop scores between current ideators and non-ideators mirrored results from ANOVAs comparing non-suicidal control participants to suicide attempters and lifetime ideators, in that no score yielded a significant *Group* x *Valence* interaction, and only Mean RTs showed a main effect of *Group*. Correlations between the Beck SSI total scores and Suicide Stroop scores were near zero (rs=-0.09-0.00, ps=0.18-0.98).

Suicide Stroop scores failed to accurately differentiate suicide attempters from non-attempters. Area under the ROC curve (AUC) and 95% confidence intervals for Suicide Stroop scoring approaches[5] were as follows: interference score AUC=0.53 (0.49-0.57), ratio score AUC=0.52 (0.49-0.57), and suicide Mean RT AUC=0.56 (0.52-0.60), indicating that Suicide Stroop task scores performed no better than chance in classifying suicide attempters from non-attempters. Categorical variables derived from suicide-related interference scores showed an estimated sensitivity of 54.8% and a specificity of 48.8%.

Post-Hoc Analyses

We conducted post hoc analyses to address two factors affecting concurrent validity. We first ran analyses to rule out habituation or practice effects, found for other ESTs (e.g., Ashley, Honzel, Larsen, Justus, & Swick, 2013). To address this possibility, we conducted analyses focused on Mean RT for the first presentation of each stimulus only. Results were consistent with Mean RT findings, showing main effects of *Group* and *Valence* with no significant *Group* x *Valence* interaction. Second, we explored the possibility that limitations of interference or difference-based scoring approaches (present in both Interference and Ratio scores), specifically the compounding of the measurement errors of individual scores, would reduce the likelihood of detecting statistically significant group differences in performance (Overall & Woodward, 1975). To address this limitation, we computed standardized residuals by separately regressing RTs for suicide-related, negative, and positive word trials onto RTs for neutral word trials, using the residuals from the predicted value to represent task performance. Tests of scores based on standardized residuals also yielded main effects of *Group* and *Valence* with no significant *Group* x *Valence* interaction.

[5] AUC was also estimated for categorical Suicide Stroop score variables (described under Data Analysis), with similar results to those reported for continuous Suicide Stroop scores.

Discussion

The goal of the current study was to test the reliability and concurrent validity of the Suicide Stroop task. First, we determined that the most common scoring approaches for the Suicide Stroop task, based on the calculation of difference scores, have poor internal consistency and lack concurrent validity. Mean RTs for all stimuli, however, demonstrated good internal consistency. Second, we found that the Suicide Stroop task did not reveal suicide-related attentional biases among suicide attempters or suicide ideators across scoring approaches (even those with excellent reliability), nor were scores able to accurately classify suicidal participants based on ROC curve analysis, indicating a lack of concurrent validity for the task. Finally, results were generally consistent across community vs. psychiatric subsamples, and across adult and adolescent subsamples. These findings demonstrate that commonly used Suicide Stroop scoring approaches have poor reliability and are unable to differentiate participants based on history of suicide attempt or ideation, indicating the Suicide Stroop task may not be useful in predicting suicide... These findings and their larger implication are discussed in further detail.

The present study underscores the need for testing and reporting of psychometric properties of behavioral measures. Reliance on the face validity of tasks aimed at capturing cognitive processes has limited our understanding of whether these tasks are actually measuring the psychological construct(s) of interest and how well they do so. Measures of RT-based performance such as the Suicide Stroop task are sensitive to individual, contextual, and procedural factors which may introduce random measurement error[6] (Ataya et al., 2012; LeBel & Paunonen, 2011). Despite this, reliability estimates for behavioral measures are infrequently

[6] Of note, many RT-based tasks demonstrate excellent reliability and validity (e.g., original Stroop, Erdodi et al., 2018), and the excellent reliability estimates for mean RTs in our study provide circumstantial evidence of overall credible responding among participants.

reported. A recent review reported that fewer than 6-10% of studies report reliability coefficients for behavioral measures (Green et al., 2016). Given that unidentified sources of measurement error in these tasks (and their resultant outcome scores) negatively impact effect size, power of hypothesis tests, and replicability of results across studies, tests of reliability are crucial (Green et al., 2016; LeBel et al., 2013).

Our reliability findings are largely consistent with other studies testing internal consistency of adapted ESTs, showing poor reliability for interference or difference scores and acceptable reliability for Mean RTs (e.g., emotion faces EST, Brown et al., 2014; panic attack EST, Dresler et al., 2012). The use of difference scores remains a topic of debate within the field (e.g., Cronbach, 1958; Overall & Woodward, 1975). Within classical test theory, the reliability of a difference score is often low when component measures are highly correlated and have similar variances, which is typically the case for difference scores calculated from experimental task RTs (May & Hittner, 2003; Hedge, Powell, & Sumner, 2017). Despite these limitations, difference scores have at times been shown to be useful for other measures (e.g., Wechsler Adult Intelligence Scale subtests; Erdodi et al., 2017; Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995). Ultimately, the benefits or potential detriments of using difference scores may depend on the measure, study design, subject matter, and whether corresponding analyses are likely to have appropriate power (Thomas & Zumbo, 2012).

Drawing from the broader EST literature, there are several design features of the Suicide Stroop task that warrant consideration and possible modification. The first design feature pertains to the presentation of stimuli in random order across trials. Studies indicate that the presentation of emotional stimuli prior to that of neutral stimuli disrupts performance on the latter, suggesting the involvement of a ‘slow’ emotional intrusion effect that is sustained across

proximal subsequent trials (McKenna, 1986; McKenna & Sharma 2004; Sharma & McKenna, 2001). The use of a blocked design format may allow for more precise observation and interpretation of the impact of word category on task performance (Ben-Haim et al., 2016). The second design feature pertains to the limited number of words represented within each valence. The Suicide Stroop task presented four words for each valence category (i.e., 16 distinct words total) across 48 trials, so word repetition was high, considering researchers recommend 20-50 words per category (Ben-Haim et al., 2016). Few stimuli that are repeated numerous times invites possible habituation effects, in which emotion-related interference decreases due to faster responding to emotion words and slower responding to neutral words (Ashley, Honzel, Larsen, Justus, & Swick, 2013). Presenting a greater number of stimuli per valence with fewer repetitions may help buffer these effects. Though our analyses targeting first presentation of stimuli did not yield results that differed from analyses of Mean RT, this may also reflect other study design features such as mixed presentation of stimuli. The third design feature pertains to the overall number of trials. Increasing the number of trials administered to each participant may contribute to more adequate reliability estimates, as reliability coefficients are a function of the length of a measure (Hedge, Powell, & Sumner, 2017). Conners Continuous Performance Task II (CPT-II) is an example of a reaction-time-based measure which implements these design features, demonstrates good reliability and construct validity, and is sensitive to a wide range of neuropsychiatric conditions (Raz, Bar-Haim, Sadeh, & Dan, 2014)

There are a number of additional factors that may contribute to the variability in RT-based scores and thus are important to consider in the context of the Suicide Stroop task. First, general RT speed is influenced by age-related differences in cognitive ability (Kiselev, Espy, & Sheffield, 2005; Der & Deary, 2006), with some evidence for gender-related differences as well

(Dane and Erzurumlugoglu, 2003; Der and Deary, 2006). Second, multiple cognitive factors including lexical processing, language ability, English language proficiency[7], as well as individual differences in inattentiveness, motivation and effort may account for some of the variance in task performance (Larsen, Mercer, & Balota, 2006; Robles, López, Salazar, Boone, & Glaser, 2015; Abeare, Messa, Zuccato, Merker, & Erdodi, 2018). As the Suicide Stroop task did not include a measure of pure reaction time, some of the variance in task performance may be related to interindividual differences in lexical processing. In the current study, inclusion and exclusion criteria, as well as data cleaning procedures that eliminate potentially invalid trials and response sets, may mitigate the impact of confounding variables on measures of task performance. However, the lack of objective measures of performance validity in the current study should be considered a limitation, particularly given that the original Stroop task has been shown to be impacted by non-credible responding (Erdodi et al., 2018; Guise, Thompson, Greve, Bianchini, & West, 2014). The present use of data cleaning procedures, which eliminated invalid trials and response sets, may exclude meaningful data and decrease the probability of detecting significant effects and is thus another study limitation.

The present study has a few additional limitations to note. First, we did not systematically sample for specific non-suicidal control groups (e.g., non-suicidal depressed individuals). Thus, we are unable to estimate the degree to which psychiatric symptoms may confound results. Second, we were not able to examine additional suicide attempt characteristics which could be meaningfully related to task performance. Recency of suicide attempt was examined inconsistently across studies and thereby difficult to compare across the overall sample.

[7] As stated in the Method section, participants were fluent in English but studies did not assess whether or not they were native speakers; as such limited. Limited English proficiency has been determined to impact performance on neuropsychological tests (e.g., Erdodi, Jongsma, & Issa, 2017).

Additionally, we used a categorical measure of suicide attempt status and did not capture features relating to severity such as number of suicide attempts and attempt lethality.

While we did not find evidence to support the reliability or validity of frequently used Suicide Stroop task scores, there have been more promising results regarding other behavioral measures of suicide and self-harm risk. Implicit association tests (IAT), for instance, measure the strength of association a person endorses between self and self-harm related constructs (i.e., death/suicide and self-injury). IATs have demonstrated acceptable to excellent internal consistency, strong evidence for concurrent validity, and predictive accuracy (Glenn et al., 2017; Nock et al., 2010; Millner, Coppersmith, Teachman, & Nock, 2018). One potential explanation for the higher reliability found in studies of the IAT is the blocked design of the task, which has been shown to produce much larger effect sizes than mixed task design in ESTs (Phaf & Kan, 2007). The IATs examined by many of these studies had a greater number of trials than the Suicide Stroop task (a minimum of 80-120 compared to 48 based on the standard IAT procedure (Greenwald, Nosek, & Banaji, 2003), which would reduce associated measurement error and increase reliability. These IATs have also featured many more practice trials, which would improve participants' retention of task instructions and accuracy of performance (Schmidt & Bjork, 1992). Future Suicide Stroop research may adopt several of these task design features to determine if they may help improve its reliability and validity, or whether other factors, such as the relevance of the construct to suicidal thoughts and behaviors, explain this difference.

The present study offers important clarifications regarding the interpretation and overall use of the Suicide Stroop task. While the Suicide Stroop task in its current form is not a reliable or valid measure of suicide risk, there remain concrete ways to modify and improve upon its current design. On a broader scale, these findings call attention to the larger issue of testing and

reporting on psychometric properties for behavioral measures. The present findings also offer a critical reminder that reliability does not ensure measure validity or usefulness, thereby underscoring the impact of psychometric properties on our ability to draw clinically significant inferences.

Acknowledgements

XXXX was supported by American Foundation for Suicide Prevention. XXXX was partially supported through funding from the National Institute of Mental Health (NIMH) GRANT # XXXX, the Simches Fund, and the Tommy Fuss Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or NIMH. XXXX was partially supported through funding from the Department of Defense (GRANT #XXXX). Data collection for XXXX was partially supported by the National Institute of Mental Health (GRANT #XXXX).The authors thank XXXX and XXXX for their guidance and helpful feedback throughout the study preparation and interpretation of results. The authors also wish to thank XXXX for her assistance reviewing this manuscript.

References

Abeare, C. A., Messa, I., Zuccato, B. G., Merker, B., & Erdodi, L. (2018). Prevalence of invalid performance on baseline testing for sport-related concussion by age and validity indicator. *JAMA Neurology*, *75*(6), 697-703.

Ashley, V., Honzel, N., Larsen, J., Justus, T., & Swick, D. (2013). Attentional bias for trauma-related words: exaggerated emotional Stroop effect in Afghanistan and Iraq war veterans with PTSD. *BMC Psychiatry*, *13*(1), 86.

Ataya, A.F., Adams, S., Mullings, E., Cooper, R.M., Attwood, A.S., & Munafo, M.R. (2012). Internal reliability of measures of substance-related cognitive bias. *Drug and Alcohol Dependence, 121*(1-2), 148-151.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*(1), 32-55.

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the Scale for Suicide Ideation. *Journal of Consulting and Clinical Psychology*, *47*(2), 343.

Becker, E. S., Strohbach, D., & Rinck, M. (1999). A specific attentional bias in suicide attempters. *Journal of Nervous and Mental Disease*, *187*(12), 730 –735.

Ben-Haim, M. S., Williams, P., Howard, Z., Mama, Y., Eidels, A., Algom, D. The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content. *Journal of Visualized Experiments* (112), e53720.

Brown, H. M., Eley, T. C., Broeren, S., MacLeod, C., Rinck, M. H. J. A., Hadwin, J. A., & Lester, K. J. (2014). Psychometric properties of reaction time based experimental

paradigms measuring anxiety-related information-processing biases in children. *Journal of Anxiety Disorders, 28*(1), 97-107.

Busch, K.A., Fawcett, J., & Jacobs, D.G. (2003). Clinical correlates of inpatient suicide. *Journal of Clinical Psychiatry*, *64*, 14–19.

Centers for Disease Control and Prevention (CDC). Web-based Injury Statistics Query and Reporting System (WISQARS) [Online]. (2013, 2011) National Center for Injury Prevention and Control, CDC (producer).

Cedrus Corp. (2011). *SuperLab* (Version 4.5). [Computer software]. San Pedro, CA: Author.

Cha, C. B., Najmi, S., Park, J. M., Finn, C. T., & Nock, M. K. (2010). Attentional bias toward suicide-related stimuli predicts suicidal behavior. *Journal of Abnormal Psychology, 119*(3), 616–622.

Cha, C. B., Najmi, S., Amir, N., Matthews, J. D., Deming, C. A., Glenn, J. J., Calixte, R. M., Harris, J. A., & Nock, M. K. (2017). Testing the efficacy of attention bias modification for suicidal thoughts: Findings from two experiments. *Archives of Suicide Research, 21*(1), 33-51.

Cha, C. B., O'Connor, R. C., Kirtley, O., Cleare, S., Wetherall, K., Eschle, S., Tezanos, K. M., & Nock, M. K. (in press). Testing mood-dependent cognitive markers for suicidal ideation.

Chung, Y., & Jeglic, E. L. (2016). Use of the modified emotional Stroop task to detect suicidality in college population. *Suicide and Life-Threatening Behavior*, *46*(1), 55-66.

Chung, Y., & Jeglic, E. L. (2017). Detecting suicide risk among college students: a test of the predictive validity of the Modified Emotional Stroop Task. *Suicide and Life-Threatening Behavior*, *47*(4), 398-409.

Cisler, J. M., Bacon, A. K., & Williams, N. L. (2009). Phenomenological characteristics of attentional biases towards threat: A critical review. *Cognitive Therapy and Research, 33*(2), 221-234.

Cyders, M. A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity?. *Clinical Psychology Review, 31*(6), 965-982.

Dane, S., & Erzurumluoglu, A. (2003). Sex and handedness differences in eye-hand visual reaction times in handball players. *International Journal of Neuroscience*, *113*, 923-929.

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. *Psychology and Aging*, *21*(1), 62-73.

Dresler, T., Ehlis, A. C., Attar, C. H., Ernst, L. H., Tupak, S. V., Hahn, T., Warrings, B., Markulin, F., Spitzer, C., Löwe, B., Deckert, J., & Fallgatter, A. J. (2012). Reliability of the emotional Stroop task: an investigation of patients with panic disorder. *Journal of Psychiatric Research*, *46*(9), 1243-1248.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 265–287.

Erdodi, L. A., Abeare, C. A., Lichtenstein, J. D., Tyson, B. T., Kucharski, B., Zuccato, B. G., & Roth, R. M. (2017). Wechsler Adult Intelligence Scale-(WAIS-IV) processing speed scores as measures of noncredible responding: The third generation of embedded performance validity indicators. *Psychological Assessment*, *29*(2), 148-157.

Erdodi, L.A., Jongsma, K.A., & Issa, M. (2017). The 15-item version of the Boston Naming Test as an index of English proficiency. *The Clinical Neuropsychologist, 31*:1, 168-178.

Erdodi, L. A., Sagar, S., Seke, K. Zuccato, B. G., Schwartz, E. S., & Roth, R. M. (2018). The Stroop Test as a measure of performance validity in adults clinically referred for neuropsychological assessment. *Psychological Assessment*, *30*(6), 755-766.

Feng, C., Becker, B., Huang, W., Wu, X., Eickhoff, S. B., & Chen, T. (2018). Neural substrates of the emotion-word and emotional counting Stroop tasks in healthy and clinical populations: A meta-analysis of functional brain imaging studies. *NeuroImage*, *173*, 258-274.

Glenn, C. R., Lanzillo, E. C., Esposito, E., Santee, A. C., Nock, M. K., & Auerbach, R. P. (2017). Examining the course of suicidal and nonsuicidal self-injurious thoughts and behaviors in outpatient and inpatient adolescents. *Journal of Abnormal Child Psychology*, *45*(5), 971-983.

Glenn, J. J., Werntz, A. J., Slama, S. J. K., Steinman, S. A., Teachman, B. A., & Nock, M. K. (2017). Suicide and self-injury-related implicit cognition: A large-scale examination and replication. *Journal of Abnormal Psychology*, *126*, 199–211.

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*(3), 750-763.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216.

Guise, B. J., Thompson, M. D., Greve, K. W., Bianchini, K. J., & West, L. (2014). Assessment of performance validity in the Stroop Color and Word Test in mild traumatic brain injury

patients: A criterion-groups validation design. *Journal of Neuropsychology*, *8*(1), 20-33.

Hedge, C., Powell, G. & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, Advance online publication. doi.org/10.3758/s13428-017-0935-1.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Jarvis, B. (2004). DirectRT (Version 2004) [Computer software]. New York, NY: Empirisoft Corp.

Kaiser, R. H., Andrews-Hanna, J. R., Spielberg, J. M., Warren, S. L., Sutton, B. P., Miller, G. A., . . . Banich, M. T. (2015). Distracted and down: Neural mechanisms of affective interference in subclinical depression. *Social Cognitive and Affective Neuroscience, 10*(5), 654-663.

Kiselev, S., Espy, K. A., & Sheffield, T. (2009). Age-related differences in reaction time task performance in young children. *Journal of Experimental Child Psychology, 102*(2), 150-166.

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion, 6*(1), 62-72.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*(4), 570-583.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8*(4), 424-432.

Malgady, R. G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement, 51*(4), 803-807.

May, K., & Hittner, J. B. (2003). On the relation between power and reliability of difference scores. *Perceptual and Motor Skills*, *97*(3), 905-908.

McKenna, F. P. (1986). Effects of unattended emotional stimuli on color-naming performance. *Current Psychology*, *5*(1), 3-9.

McKenna, F.P., & Sharma, D. (2004). Reversing the emotional stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 382-392.

McNally, R.J., Kaspi, S.P., Riemann, B.C., & Zeitlin, S.B. (1990). Selective processing of threat cutes in posttraumatic stress disorder. *Journal of Abnormal Psychology, 99*(4), 398-402.

Miller, G. M., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48.

Millner, A. J., Coppersmith, D. D. L., Teachman, B. A., & Nock, M. K. (2018). The Brief Death Implicit Association Test: Scoring recommendations, reliability, validity, and comparisons with the Death Implicit Association Test. *Psychological Assessment*, *30*(10), 1356-1366.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale—Revised. *Professional Psychology: Research and Practice*, *26*(5), 491-498.

Mitterschiffthaler, M.T., Williams, S.C.R., Walsh, N.D., Cleare, A.J., Donaldson, C., Scott, J., & Fu, C.H. (2008). Neural basis of the emotional stroop interference effect in major depression. *Psychological Medicine, 38*(2), 247-256.

Nock, M. K., Holmberg, E. B., Photos, V. I., & Michel, B. D. (2007). The self-injurious thoughts and behaviors interview: Development, reliability, and validity in an adolescent sample. *Psychological Assessment, 19*, 309–317.

O'Connor, D. B., Green, J. A., Ferguson, E., O'Carroll. R. E., & O'Connor, R. C. (2017). Cortisol reactivity and suicidal behavior: investigating the role of the hypothalamic-pituitary-adrenal axis responses to stress in suicide attempters and ideators. *Psychoneuroendocrinology*, *75*, 183-191.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*(1), 85-86.

Phaf, R.H & Kan, K.J. (2007). The automaticity of emotional stroop: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 181-199.

Richard-Devantoy, S., Ding, Y., Turecki, G., & Jollant, F. (2016). Attentional bias toward suicide-relevant information in suicide attempters: A cross-sectional study and a meta-analysis. *Journal of Affective Disorders*, *196*, 101-108.

Robles, L., López, E., Salazar, X., Boone, K. B., & Glaser, D. F. (2015). Specificity data for the b Test, dot counting test, Rey-15 item plus recognition, and Rey word recognition test in

monolingual Spanish-speakers. *Journal of Clinical and Experimental Neuropsychology, 37*(6), 614-621.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-218.

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*(2), 374-408.

Sharma, D., & McKenna, F. P. (2001). The role of time pressure on the emotional Stroop task. *British Journal of Psychology, 92*(3), 471-481.

Stewart, J. G., Glenn, C. R., Esposito, E. C., Cha, C. B., Nock, M. K., & Auerbach, R. P. (2017). Cognitive control deficits differentiate adolescent suicide ideators from attempters. *Journal of Clinical Psychiatry, 78*(6), e614–e62.

Substance Abuse and Mental Health Services Administration, *Results from the 2013 National Survey on Drug Use and Health: Mental Health Findings*, NSDUH Series H-49, HHS Publication No. (SMA) 14-4887. Rockville, MD: Substance Abuse and Mental Health Services, 2014.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 23*(1), 9-17.

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37-43.

Wenzel, A., & Beck, A. T. (2008). A cognitive model of suicidal behavior: Theory and treatment. *Applied and Preventive Psychology, 12*(4), 189-201.

Willett, J.B. (1998). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

Williams, J. M. G., & Broadbent, K. (1986). Distraction by emotional stimuli: Use of a Stroop task with suicide attempters. *British Journal of Clinical Psychology, 25*(2), 101–110.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1), 3-24.

Zimmermann, K., Walz, C., Derckx, R. T., Kendrick, K. M., Weber, B., Dore, B., Ochsner, K.N., Becker, B. (2017). Emotion regulation deficits in regular marijuana users. *Human Brain Mapping, 38*(8), 4270-4279.

Table 1. Sample characteristics by study

Study	Sample Size	Age Group	Age (years) *M* (*SD*)	Gender % Female	Race/Ethnicity % White[a]
			Psychiatric Sample		
1	126	Adult	33.75 (11.44)	42.9	81.0
2	37	Adult	41.70 (14.57)	48.7	81.1
3	177	Adolescent	15.60 (1.34)	73.1	77.0
			Community Sample		
4	117	Adult	32.41 (13.60)	51.3	59.0
5	144	Adult	31.74 (13.47)	56.3	72.9
6	128	Adult	33.55 (13.46)	56.3	30.2[b]
7	146	Adolescent	17.37 (1.80)	81.6	63.7

Note. Study 1 = XXXX; Study 2 = XXXX; Study 3 = XXXX; Study 4 = XXXX; Studies 5-6 = XXXX; Study 7 = XXXX. [REMOVED IDENTIFYING CITATIONS TO MAINTAIN BLIND]

[a] For all studies except Study 6, non-White race and ethnicity categories included Black/African American, Hispanic, Asian, and Other or Biracial.

[b] Demographic survey for Study 6 allowed participants the option of identifying by nationality, which we opted to classify as 'Other' when no other race or ethnicity was reported.

Table 2. Between group differences by scoring approach

Score	Suicide Attempters *n*=320 *M* (*SD*)	Suicide Ideators *n*=338 *M* (*SD*)	Non-suicidal controls *n*=217 *M* (*SD*)	Test *F*(2,872)	η_p^2
Mean RT_{Sui}	633.48 (196.33)	586.98 (185.39)	601.46 (175.54)	5.23**	0.01
Mean RT_{Neg}	629.00 (193.76)	580.40 (173.5)	599.6 (166.74)	6.07**	0.01
Mean RT_{Pos}	625.46 (190.93)	577.77 (167.82)	600.26 (171.47)	5.93**	0.01
Mean RT_{Neu}	621.61 (187.64)	579.15 (172.71)	597.51 (171.87)	4.68**	0.01
$Interference_{Sui}$	11.87 (71.68)	7.83 (72.9)	3.96 (58.48)	0.86	0.00
$Interference_{Neg}$	7.39 (71.13)	1.25 (58.82)	2.09 (53.04)	0.89	0.00
$Interference_{Pos}$	3.85 (63.82)	-1.38 (62.79)	2.75 (56.75)	0.64	0.00
$Interference_{SuiWord}$	7.30 (100.05)	11.35 (89.09	9.10 (75.15)	0.17	0.00
$Ratio_{Sui}$	0.02 (0.10)	0.02 (0.10)	0.01 (0.09)	0.96	0.00
$Ratio_{Neg}$	0.02 (0.10)	0.01 (0.09)	0.01 (0.09)	0.88	0.00
$Ratio_{Pos}$	0.01 (0.10)	0.00 (0.10)	0.01 (0.09)	0.46	0.00

***p*<.01

Note. All Suicide Stroop score Means and Standard Deviations reported in milliseconds (ms). Multiple ANOVAs for Mean RT did not produce a significant *Group* x *Valence* interaction, indicating no significant between group differences for RT valence.

Table 3. Suicide Stroop task scoring approaches

Scoring Type	**Calculation Method**	**Name**
1. Mean RT	Mean RT for Suicide, Negative, Positive, and Neutral words	Mean RT_{Sui}, Mean RT_{Neg}, Mean RT_{Pos}, Mean RT_{Neu}
2. Interference	Suicide word RT - Neutral word RT	$Interference_{Sui}$
	Negative word RT - Neutral word RT	$Interference_{Neg}$
	Positive word RT - Neutral word RT	$Interference_{Pos}$
	"Suicide" word only RT-Neutral word RT	$Interference_{SuiWord}$
3. Ratio Score	(Suicide word RT - Neutral word RT)/Neutral RT	$Ratio_{Sui}$
	(Negative word RT - Neutral word RT)/Neutral RT	$Ratio_{Neg}$
	(Positive word RT - Neutral word RT)/Neutral RT	$Ratio_{Pos}$

Note. RT=Reaction Time

Table 4. Split-half reliability for scoring approach by setting and age group

Score	All Subjects	Community	Psychiatric	Adults	Adolescents
	n=875	*n*=340	*n*=535	*n*=552	*n*=323
Mean RT_{Sui}	0.94	0.94	0.94	0.94	0.95
Mean RT_{Neg}	0.93	0.94	0.94	0.93	0.94
Mean RT_{Pos}	0.93	0.94	0.93	0.93	0.94
Mean RT_{Neu}	0.93	0.93	0.93	0.93	0.93
$Interference_{Sui}$	0.13	0.23	-0.02	0.02	0.32
$Interference_{Neg}$	-0.09	-0.24	0.06	-0.05	-0.18
$Interference_{Pos}$	0.02	-0.06	0.11	-0.03	0.13
$Interference_{SuiWord}$	-0.02	0.09	-0.19	0.00	-0.09
$Ratio_{Sui}$	0.04	0.10	-0.05	-0.01	0.11
$Ratio_{Neg}$	-0.05	-0.11	0.01	0.00	-0.15
$Ratio_{Pos}$	-0.04	-0.02	-0.07	-0.07	-0.01

Note. RT= Reaction Time. Community subsample includes participants from studies 4-7; Psychiatric subsample includes participants from studies 1-3; Adult subsample includes participants from studies 1, 2, 4, 5, and 6; Adolescent subsample includes participants from studies 3 & 7

Table 5. ANOVAs comparing Suicide Stroop performance across suicide attempters, suicide ideators, and non-suicidal controls for each scoring approach, by sample-type and age group

Score	Community		Psychiatric		Adults		Adolescents	
	F	η_p^2	*F*	η_p^2	*F*	η_p^2	*F*	η_p^2
Mean RT_{Sui}	0.27	0.00	5.90**	0.03	7.11***	0.03	0.11	0.00
Mean RT_{Neg}	0.71	0.00	5.26**	0.03	7.18***	0.03	0.24	0.00
Mean RT_{Pos}	0.15	0.00	7.41**	0.04	6.05**	0.02	0.48	0.00
Mean RT_{Neu}	0.18	0.00	5.23**	0.03	5.91**	0.02	0.14	0.00
$Interference_{Sui}$	0.28	0.00	1.29	0.01	1.05	0.00	0.03	0.00
$Interference_{Neg}$	1.59	0.01	0.42	0.00	0.44	0.00	0.54	0.00
$Interference_{Pos}$	0.06	0.00	1.14	0.01	0.02	0.01	1.31	0.01
$Interference_{SuiWord}$	0.23	0.00	0.01	0.00	0.38	0.00	0.48	0.00
Ratio $Score_{Sui}$	0.21	0.00	1.31	0.01	0.83	0.00	0.21	0.00
Ratio $Score_{Neg}$	1.49	0.01	0.30	0.00	0.58	0.00	0.33	0.00
Ratio $Score_{Pos}$	0.26	0.00	0.96	0.01	0.03	0.00	0.72	0.00

***p*<.01, ***p<.001

Note. Community subsample includes participants from studies 4-7; Psychiatric subsample includes participants from studies 1-3; Adult subsample includes participants from studies 1, 2, 4, 5, and 6; Adolescent subsample includes participants from studies 3 & 7.