# Toward a More Valid Assessment of Behavioral Aggression: An Open Source Platform and an Empirically Derived Scoring Method for Using the Competitive Reaction Time Task (CRTT)

Lobbestael, J.; Emmerling, F.; Brugman, S.; Broers, N.; Sack, A.T.; Schuhmann, T.; Bonnemayer, C.; Benning, R.; Arntz, A.

**DOI**
10.1177/1073191120959757

**Publication date**
2021

**Document Version**
Final published version

**Published in**
Assessment

*Article*

Assessment
2021, Vol. 28(4) 1065–1079

DOI: 10.1177/1073191120959757
journals.sagepub.com/home/asm
SAGE

# Toward a More Valid Assessment of Behavioral Aggression: An Open Source Platform and an Empirically Derived Scoring Method for Using the Competitive Reaction Time Task (CRTT)

**Jill Lobbestael[1], Franziska Emmerling[2*], Suzanne Brugman[1], Nick Broers[1], Alexander T. Sack[1], Teresa Schuhmann[1], Charlie Bonnemayer[1], Richard Benning[1], and Arnoud Arntz[3]**

**Abstract**

While the Competitive Reaction Time Task (CRTT) is the most used behavioral aggression paradigm, it is characterized by methodological heterogeneity and quantification strategies for its' outcome are unstandardized. Therefore, the standards of measuring aggression should be improved. This article contributes on such an improvement by providing: (a) a freely available CRTT online administration program, and (b) a factor-analytically derived scoring method. Based on a combined sample ($n$ = 423), a two-factor model was fit to the 30-trial CRTT version. The first factor included all trial scores subsequent to the first time the participant received aversive feedback (i.e., provoked factor) and the second factor included all trial scores prior to this first aversive feedback (i.e., unprovoked factor). Construct validity was evidenced based on the factors` differential relationship with self-reported aggression and narcissism. Our factor analytic findings empirically support the superiority of one of the existing CRTT scoring methods, that is, separately averaging all preprovocation versus all postprovocation trials. We discuss practical recommendations for CRTT users and outline future empirical avenues. This article aims at stimulating joint efforts to move toward standardization of CRTT implementation and outcome measure analysis.

**Keywords**

Competitive Reaction Time Task, aggression, scoring algorithm, aggression assessment

Excessive aggression has detrimental effects on an individual (e.g., increased blood pressure; Player et al., 2007), interpersonal (i.e., loss of social relationships; Coie et al., 1991), and societal (i.e., costs for civil trials and incarceration of offenders; Foster & Jones, 2005) level. Consequently, increasing our knowledge about how to predict and decrease aggression is crucial. Self-report measures of aggression are often flawed by socially desirable answer tendencies (Vigil-Colet et al., 2012). The use of behavioral measures of aggression therefore is particularly valuable as behavior is not as easily susceptible to social desirability as is self-report. Developing a controlled but at the same time ecologically valid behavioral aggression task is challenging due to ethical and practical considerations (Ritter & Eslea, 2005). Despite these challenges, a number of creative tasks have been marketed such as the Point Subtraction Aggression Paradigm (Cherek et al., 1997), the Hot Sauce Paradigm (Lieberman et al., 1999), and the Voodoo Doll Task (DeWall et al., 2013). All these paradigms are based on providing participants with the opportunity to negatively affect an alleged opponent, for example, by subtracting points from a competitive score or exposure to aversive food.

A further category of behavioral aggression measures are tasks based on the Taylor Aggression Paradigm (TAP; Taylor, 1967), in which competition between two or more opponents is simulated, during which participants can

[1]Maastricht University, Maastricht, the Netherlands
[2]Technical University Munich, Munich, Germany
[3]University of Amsterdam, Amsterdam, the Netherlands

*Franziska Emmerling previously published as Franziska Dambacher

**Corresponding Author:**
Jill Lobbestael, Department of Clinical Psychological Science, Faculty of Psychology and Neuroscience, Maastricht University, Universiteitssingel 40, 6229 ET Maastricht, Netherlands.
Email: jill.lobbestael@maastrichtuniversity.nl

administer aversive stimuli to their opponents. A widely implemented variant of the TAP is the Competitive Reaction Time Task (CRTT, see, e.g., Bushman & Baumeister, 1998; Warburton & Bushman, 2019). In the CRTT, participants are led to believe that they engage in a competitive task against an opponent. Participants are told that they can administer aversive auditory feedback to each other. In case the participant "loses" a trial, he or she gets administered an aversive white noise through headphones by the opponent. Vice versa, in case the participant "wins," the opponent is administered an aversive tone by the participant. Prior to each trial, the participant can preset both the volume and duration of this aversive tone, which serve as behavioral indications of physical aggression. In reality, there is no actual opponent and the win or loss trials—as well as the volume and duration of the feedback the participant is confronted with—are preprogrammed and standardized for every participant.

Studies largely support the reliability (Cronbach's $\alpha$ values ranging from $\alpha = .85$ to .99, Chester & Lasko, 2019; Ferguson & Rueda, 2009; Schmidt et al., 2015; Thomaes et al., 2008) and validity of the CRTT. Convergent validity is evidenced by the paradigm's positive correlation with other behavioral laboratory aggression measures (Chester & Lasko, 2019), as well as by its responsiveness to factors known to magnify aggressive behavior like provocation (Chester & Lasko, 2019) and alcohol (Giancola & Parrott, 2008). CRTT scores are, furthermore, positively related to self-reported aggression (Giancola & Parrott, 2008; Giancola & Zeichner, 1995), especially state aggression or recent aggressive acts (Chester & Lasko, 2019). A lack of association with traits like competition and experience seeking (Bernstein et al., 1987; Chester & Lasko, 2019; Giancola & Zeichner, 1995) supports the CRTT's discriminant validity. Finally, the CRTT has demonstrated ecological validity because its reactivity after viewing violent scenes (Bushman, 1995) is similar to that of real-life acts of physical aggression.

## Methodological Heterogeneity

The CRTT is predominantly used in social and clinical psychological science. The literature indicates that the CRTT provides scientists with a large degree of methodological flexibility. On the one hand, this flexibility can be considered an advantage as it allows the application of the paradigm in manifold contexts. On the other hand, flexibility seriously threatens the validity of aggression studies; flexibility in study design and data analyses has been shown to (a) increase the likelihood of Type I error rates and inflated effect size estimates (LeBel et al., 2017; Simmons et al., 2011; Wicherts et al., 2016) and (b) dilute the interpretability and applicability of the results (Baker at al., 1992).

The methodological heterogeneity of using the CRTT manifests within three areas. First, many different versions of cover stories are used. Participants physically meet their opponents (e.g., Anderson & Anderson, 2008; Dambacher et al., 2015b; Pickett et al., 2016) or are told that that they are present in an adjacent room (e.g., Wilkowski et al., 2010; Wilkowski et al., 2015). Some are told that a second experimenter guides the opponent (Anderson & Anderson, 2008; DeWall et al., 2007). There is substantial variation in the effort researchers put into convincing their participant of the actual presence of an actively participating opponent. Some experimenters tell the participant that they should start quickly with the experiment because they received a message that the opponent is ready to start the CRTT (Banse et al., 2015; Brugman et al., 2018). Other experimenters present participants with a connection screen before the onset of the CRTT while supposedly waiting for their opponent to start the CRTT (Brugman et al., 2018; Wilkowski et al., 2015) or tell the participants that they are going to check on the progress of the opponent in another room (Brugman et al., 2015). Further safeguard against suspicion of premeditation is sometimes integrated within the CRTT programming, for example, by having participants automatically lose trials in which they respond very slowly or press a wrong key (Imhoff et al. 2013; Sestir & Bartholow, 2010), or by inserting randomized waiting periods after participants' responses during which the participant is told to wait for his or her opponent's selection (Wilkowski et al., 2010). The second area of methodological heterogeneity with respect to CRTT usage concerns trial specifications. Most studies either use a 25- (e.g., Anderson & Anderson, 2008; Elson et al., 2015; Sestir & Bartholow, 2010; Whitaker & Bushman, 2012) or a 30- (e.g., Brugman et al., 2018; Dambacher et al., 2015a, 2015b; Lobbestael et al., 2014; Sherrill et al., 2016) trial version. Most articles do not specify the preset volume, duration, and the win/loss structure per trial (see Banse et al., 2015; Brugman et al., 2015, for exceptions). Many researchers report that trial specifications are assigned in a random yet fixed order (e.g., Hahn-Holbrook et al., 2011; Sherrill et al., 2016) or based on predefined numbers of win versus loose trials (Murphy et al., 1992). Other studies use preset tones that gradually increase in severity and/or duration (e.g., Lansu et al., 2014; Moss & Maner, 2016) or a blocked design starting with less provocative and building up to highly provocative tones (e.g., Anderson & Anderson, 2008; Imhoff et al., 2013; Weisbuch et al., 1999). Elson et al. (2014) empirically addressed the impact of differential trial specification by reanalyzing three CRTT studies and concluded that part of the explained variance was indeed rooted in the differential volume and duration levels. The third area of heterogeneity—that is the focus of the present study—represents the largest source of ambiguity in CRTT results and concerns the vast number of strategies to quantify outcome scores. Elson (2016, see http://www.flexiblemeasures.com/crtt) scrutinized these strategies across 130 publications and

found 157 distinct calculation strategies. Largely, these strategies match four categories, that is, volume-based (e.g., average volume across all trials), duration-based (e.g., duration of first trial only), composite (e.g., standardized and averaged sum of volume and duration across all trials), and other (e.g., number of trials before maximum volume was set for the first time) quantifications (Elson, 2016). While some of these scoring methods are characterized by merely slight differences (e.g., averaging all values of 25 vs. 30 trials), other methods (e.g., only duration after winning trials) are more distinct and likely cause strong impacts on study findings. Most scoring methods are used only once, and many papers report findings from multiple scoring strategies. Over the past 30 years, the number of scoring methods with respect to CRTT outcome measures increased constantly. Justifications for why a certain method was chosen are rarely provided.

## Empirically Derived Scoring Algorithm

The lack of standardization of the CRTT between and within laboratories is extremely problematic because it undermines the paradigm's validity (Chester & Lasko, 2019; Elson et al., 2014). Particularly, the multitude of scoring methods increases the chances of Type I errors (i.e., false positives) because it allows researchers to selectively report results from an (ad hoc) quantification strategy that supports their hypothesis (Chester & Lasko, 2019; Simmons et al., 2011). To overcome these problems, both Elson (2016) and Chester and Lasko (2019) suggested a twofold solution. The first solution includes preregistration of the analyses plan prior to data collection. Second, researchers should specify why a particular quantification method is used. While both approaches undoubtedly avoid ad hoc cherry-picking of findings that are either significant or in line with ones' hypothesis (also referred to as *p*-Hacking, see Warburton & Bushman, 2019), they do not contribute to the development of a standardized scoring method. Researchers, thus, remain in the dark as to what quantification method is favorable. Currently, empirical evidence on which of the many existing CRTT scoring options is optimal, is lacking (McCarthy & Elson, 2018). Therefore, the main aim of this study is the development of an empirically derived scoring algorithm for the CRTT. To our knowledge, this is the first data-driven solution toward such a standardized CRTT scoring method. For this purpose, we collapsed several data sets into one large sample including both patients and nonpatients and used a factor-analytic method to delineate the structure underlying the CRTT outcome measures. External validity was assessed by comparing the factor-based scoring method with self-reported aggression and the personality trait of narcissism. In the following section, we describe the specifics of the study leading to the construction of this scoring algorithm.

## Method

### Materials: CRTT

Participants were told that they were competing against an opponent and that they had to try to outperform this opponent by mouse-clicking a rectangle when it turned from yellow to red. The amount of time it took before the rectangle changed from yellow to red was randomized, ranging from 1,000 milliseconds to 2,000 milliseconds. The amount of win or lose trials as well as volume and duration levels of the tones administered by the opponent, were preprogrammed in the same order for every participant (see Tables 1 and 2). A valid reaction time range was set between 0 and 2,000 milliseconds; when participants reacted slower, they automatically lost the trial to ensure credibility. Participants were told that the winner of a trial could administer a loud noise to his or her opponent and that this noise could influence the performance of the opponent on the next trial. Before each trial, participants were asked to choose the duration and volume of this noise blast. Sliders with a range between 0 and 10 were used to set both, duration and volume. Regarding volume, 0 represented no noise at all and 10 was equal to 100 dB, which exceeds the United States Safety and Health Standards recommendations for sustained 8-hour exposure of 90 decibels for full-time workers, however, well within the pain threshold of 125 decibels. Regarding duration, 0 represented 0 seconds and 10 represented 5 seconds. To provide participants with the opportunity to refrain from aggressive responding, we explicitly instructed that they could also choose to refrain from giving their opponent an aversive tone by moving both sliders but setting them back to the zero level. The CRTT was programmed in an online custom-made software tool, using HTML5, JavaScript, and CSS. This tool can be hosted on any web server and is compatible with any web browser. The program exists of three parts: (a) the settings.xml file where researchers can specify volume, duration, and win/loss characteristics of each trial, (b) the actual CRTT task for the participant to complete, and (c) the final page, showing a systematic overview of the outcomes. We opted for a user-friendly version based on the TAP (Taylor, 1967) and the Competitive Reaction Task Reward and Punish (version 3.2.5, Bushman & Baumeister, 1998; Warburton & Bushman, 2019). Our online version is free to use and can be accessed from https://soto.maastrichtuniversity.nl/users/crtt/crtt-en.html. A zip file containing the source code can be downloaded from https://soto.maastrichtuniversity.nl/users/crtt/crtt.zip.

When using this version, please use the following reference: Lobbestael et al., 2020.

### Materials: RPQ

Self-reported reactive and proactive aggression were measured with the Reactive Proactive Questionnaire (RPQ;

**Table 1.** Preprogramed Trials of the CRTT (25-Trial Version).

| Trial number | Volume | Duration | Win/lose |
|---|---|---|---|
| 1 | 0 | 0 | Win |
| 2 | 6 | 7 | Lose |
| 3 | 1 | 1 | Win |
| 4 | 6 | 5 | Lose |
| 5 | 3 | 7 | Lose |
| 6 | 5 | 2 | Lose |
| 7 | 5 | 9 | Win |
| 8 | 2 | 6 | Lose |
| 9 | 1 | 3 | Win |
| 10 | 3 | 3 | Win |
| 11 | 6 | 5 | Lose |
| 12 | 10 | 2 | Win |
| 13 | 4 | 6 | Win |
| 14 | 7 | 9 | Lose |
| 15 | 3 | 10 | Lose |
| 16 | 6 | 5 | Win |
| 17 | 1 | 10 | Lose |
| 18 | 10 | 6 | Lose |
| 19 | 4 | 10 | Win |
| 20 | 9 | 10 | Lose |
| 21 | 6 | 4 | Win |
| 22 | 2 | 3 | Lose |
| 23 | 9 | 7 | Lose |
| 24 | 10 | 3 | Win |
| 25 | 2 | 6 | Lose |

*Note*. CRTT = Competitive Reaction Time Task.

Raine et al., 2006). The RPQ consists of 23 items, 11 assessing reactive aggression and 12 assessing proactive aggression. Items have to be rated on frequency (0 = never, 1 = sometimes, 2 = often). The RPQ subscales and total scores showed good internal reliability (Cronbach's $\alpha > .75$) and factor analyses demonstrated that the two-factor solution outperformed a one-factor solution (Cima & Raine, 2009; Raine et al., 2006). The RPQ showed adequate convergent and criterion validity as well as temporal stability (Cima et al., 2013).

## Materials: NPI

Narcissism was measured with the Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979). The NPI consists of 37 items that are rated on a 7-point Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The NPI has a four-factor loading solution comprising of: Leadership/Authority, Self-Absorption/Self-Administration, Superiority/Arrogance, and Exploitiveness/Entitlement (Emmons, 1987). The NPI approaches narcissism dimensionally and focusses primarily on the grandiose form as conceptualized in the *Diagnostic and Statistical Manual of Mental Disorders–Fifth edition* (American Psychiatric Association, 2013). Good construct validity and internal consistencies (between $\alpha = .83$ and $\alpha = .86$) have been reported (e.g., Raskin & Terry, 1988).

**Table 2.** Preprogramed Trials of the CRTT (30-Trial Version).

| Trial number | Volume | Duration | Win/lose |
|---|---|---|---|
| 1 | 0 | 0 | Win |
| 2 | 0 | 0 | Win |
| 3 | 0 | 0 | Win |
| 4 | 0 | 0 | Lose |
| 5 | 0 | 0 | Lose |
| 6 | 0 | 0 | Win |
| 7 | 6 | 7 | Lose |
| 8 | 1 | 1 | Win |
| 9 | 6 | 5 | Lose |
| 10 | 3 | 7 | Lose |
| 11 | 5 | 2 | Lose |
| 12 | 5 | 9 | Win |
| 13 | 2 | 6 | Lose |
| 14 | 1 | 3 | Win |
| 15 | 3 | 3 | Win |
| 16 | 6 | 5 | Lose |
| 17 | 10 | 2 | Win |
| 18 | 4 | 6 | Win |
| 19 | 7 | 9 | Lose |
| 20 | 3 | 10 | Lose |
| 21 | 6 | 5 | Win |
| 22 | 2 | 10 | Lose |
| 23 | 10 | 6 | Lose |
| 24 | 4 | 10 | Win |
| 25 | 9 | 10 | Lose |
| 26 | 6 | 4 | Win |
| 27 | 2 | 3 | Lose |
| 28 | 9 | 7 | Lose |
| 29 | 10 | 3 | Win |
| 30 | 2 | 6 | Lose |

*Note*. CRTT = Competitive Reaction Time Task.

## Procedure

Healthy volunteers were recruited through advertisements on posters, flyers, and social media. They participated in exchange for study credits or €7.50 per hour in vouchers. Clinicians were asked for permission to contact patients meeting the inclusion criteria. Next, selected participants were provided with information about the study. If the participant agreed to participate, meetings were planned to conduct the experiment. All relevant studies were approved by the local ethical committee of Maastricht University (ERCPN codes ECP-138, 17_03_2014; ECP-119 06_09_2012; ECP-107 05_10_2011; ECP 145 07_10_2014; ECP-129 05_06_2013; ECP-107- 05_10_2011) and by the Florida State University`s Human Subject Committee (code 2010.5111).

### Data Sources

CRTT data were collected within eight different studies conducted in our lab and that of Florida State University over the past 6 years. Seven studies included nonpatients (Brugman et al., 2015; Dambacher et al., 2015a; Dambacher et al., 2015b; Lobbestael et al., 2014; Lobbestael & Brugman, in preparation; Lobbestael & Vancleef, in preparation; Van Teffelen et al., in press), and one study included patients from a forensic psychiatric clinic (Brugman et al., 2018). Together, $N = 620$ participants were included in these studies. Per study, the participant numbers ranged between $n = 26$ and $n = 100$. Seventeen (2.7%) participants were excluded because of extreme trial values (i.e., scores of 0 or of 10 on all trials), resulting in a total sample of $N = 603$. Five studies used the 30-trial version of the CRTT (Brugman et al., 2015; Brugman et al., 2018; Dambacher et al., 2015a; Dambacher et al., 2015b; Lobbestael et al., 2014), whereas two studies (Lobbestael & Vancleef, in preparation; Van Teffelen et al., in press) used the 25-trial version, and one study (Lobbestael & Brugman, in preparation) used both versions.

### Data Analyses

Data were checked for normality. Since participants had to determine both volume and duration of each trial, the number of trials per participants was either 50 (in case of the 25-trial version) or 60 (in case of the 30-trial version). Because the 30-trial CRTT version incorporates multiple (seven) unprovoked trials and the 25-trial version only two, the data from the 30-trial version are more reliable and thus could not be clustered directly with that of the 25-trial version. The following analyses were therefore performed on the 30-trial subsample only ($n = 423$). The 60 variables were entered into an exploratory factor analysis (EFA). All EFA analyses were conducted with M*plus* version 7.2 (Muthén & Muthén, 1998-2020). Apart from enabling the researcher to apply conventional criteria like checking eigenvalues, screeplot and interpretability, M*plus* also provides results for a model Chi Square goodness-of-fit test and output on other fit criteria like the root mean squared error of approximation (RMSEA), the Tucker–Lewis index (TLI), and the standardized root mean square residual (SRMR). Although such fit measures are most powerfully used in a confirmatory rather than an exploratory context, they can be of assistance in gauging how well the original correlation matrix can be described with a given factor solution. The RMSEA is a measure of "badness of fit," with lower values indicating better fit. Traditionally, RMSEA values smaller than .08 are taken to be indicative of a fair fitting model, and values smaller than .05 reflect a close fit. TLI is an incremental fit index, showing how the current model has succeeded in improving the reproduction of the correlation matrix, relative to the performance of the independence (i.e., worst possible) model. TLI values above .90 are traditionally considered as good fitting models. For SRMR, a measure that roughly corresponds to the average standardized residual covariance, values below .08 are considered as reflecting a good fit. In addition, M*plus* enables the use of the robust maximum likelihood estimator MLR to correct for violations of the normality assumption. This estimator calculates robust standard errors for all parameter estimates and provides a scaled correction of the model chi-square test and related goodness-of-fit statistics like RMSEA and SRMR. As a further addition, M*plus* provides robust tests for the statistical comparison of different models (i.e., factor solutions).

Since provoked and unprovoked aggression have been shown to consistently correlate highly across methodologies (Polman et al., 2007), we rotated the solution using the default M*plus* Geomin method, an oblique rotation method that allows factors to be correlated.

As a follow up, the factor analysis was replicated for the 25-trial version only, as well as for other subsets (i.e., excluding patients, suspicious participants and participants first having underwent an experimental manipulation).

To assess construct validity of the CRTT scores, Pearson correlations were calculated with the RPQ and the NPI in a subsample of the 30-trial CRTT version of respectively $n = 337$ and $n = 153$.

## Results

### Demographic Information

In the 30-trial version, $n = 423$, the majority (84.2%) was male, with a mean age of 26.86 years ($SD = 10.89$, range from 18 to 68 years). The sample was predominantly Caucasian (71.2%), followed by Hispanic (2.6%), Asian (1.7%), African American (1.4%), Arabian (0.7%), and others (0.5%), 22% unknown. Furthermore, 23.9% of participants completed a bachelor master education, while 70.4% completed high school, 5.4% elementary school, and 0.2% no education. In the subsample of forensic psychiatric patients ($n = 84$), the majority suffered from antisocial (44%), borderline (19%), or narcissistic (10.7%) personality disorders. The most prevalent clinical disorders were substance abuse (42.9%), sexual disorders (25%), attention deficit disorders (13.8%), mood disorders (11.9%), impulse control disorders (11.9%), anxiety disorders (8.3%), or developmental disorders (6%). Other disorders were diagnosed in less than 5% of the forensic subsample.

### Factor Structure

Inspection of the descriptive statistics ($n = 423$) showed that the kurtosis values of the later trials gradually became

**Table 3.** List of all Eigenvalues >1 (Based on the 30-Trial CRTT Version, $n = 423$).

| Factor | Eigenvalue | % Explained variance | % Cumulative explained variance |
|---|---|---|---|
| 1 | 20.341 | 33.902 | 33.902 |
| 2 | 4.553 | 7.588 | 41.490 |
| 3 | 2.038 | 3.397 | 44.887 |
| 4 | 1.852 | 3.087 | 47.975 |
| 5 | 1.735 | 2.892 | 50.866 |
| 6 | 1.496 | 2.493 | 53.360 |
| 7 | 1.296 | 2.161 | 55.520 |
| 8 | 1.265 | 2.108 | 57.628 |
| 9 | 1.173 | 1.954 | 59.583 |
| 10 | 1.096 | 1.872 | 61.410 |
| 11 | 1.001 | 1.668 | 63.077 |

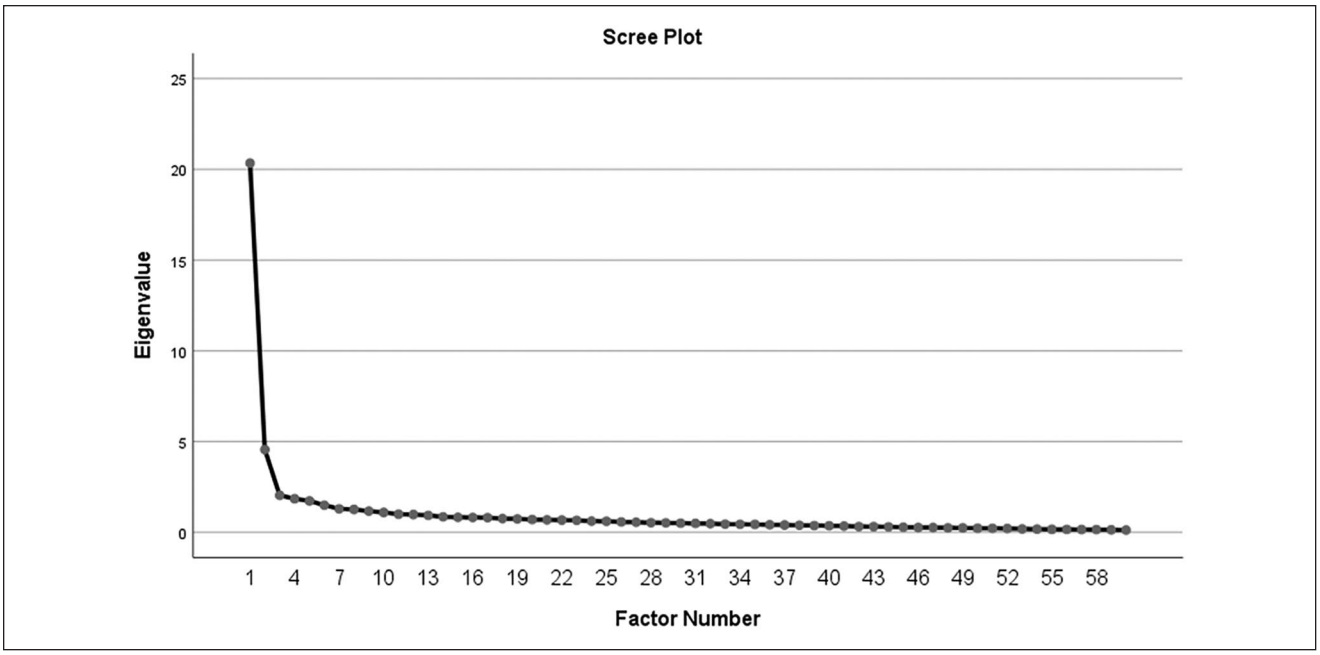


**Figure 1.** Scree plot of EFA of the 30-trial CRTT version with 60 trials ($n = 423$).
*Note.* EFA = exploratory factor analysis; CRTT = Competitive Reaction Time Task.

more negative (i.e., not within ±1 *SE* range; Howell, 2007). Inspection of the histograms revealed that participants administer trials of maximum duration and volume toward the end of the CRTT task (i.e., starting around trial 20), pointing to a ceiling effect.

The Kaiser–Meyer–Olkin measure of sampling adequacy was *0.94*, thus exceeding the recommended value of *0.60*, and Bartlett's test of Sphericity was significant, $\chi^2(1770) = 15501.54, p \leq .001$. Eigendecomposition of the original correlation matrix showed there to be 11 eigenvalues larger than 1 (see Table 3). Parallel analyses using O`Conner`s (2000) SPSS syntax indicated the eigenvalues of the true data (i.e., .498) became lower than those simulated in the random models (i.e., .517) from factor 13 on. This implies that 12 factors can be considered the upper limit of potential number of existing factors in the file. The screeplot (see Figure 1) however clearly reflected that after extraction of the first two principal components, the remaining components added only marginally to the 41.49% of the variance that was explained by the first two components (see Table 3). This suggested that although more of the variance could be explained by extracting more principal components, the objective of data reduction for practical purposes would not be met by simply adding all components with eigenvalues larger than 1.

To gain a richer view of the potential of more complex factor structures, we compared all factor solutions that were suggested by the parallel analysis (12) on a number of

**Table 4.** Goodness-of-Fit Measures of All 12-Factor Solutions.

| #Factors | Model chi square (*df*) | RMSEA (*p* close fit) | TLI | SRMR |
|---|---|---|---|---|
| 1 | 5478.217 (1710), $p < .001$ | .072 ($p < .001$) | .611 | .083 |
| 2 | 4085.187 (1651), $p < .001$ | .059 ($p < .001$) | .739 | .053 |
| 3 | 3902.608 (1593), $p < .001$ | .059 ($p < .001$) | .744 | .048 |
| 4 | 3667.054 (1536), $p < .001$ | .057 ($p < .001$) | .755 | .043 |
| 5 | 3405.356 (1480), $p < .001$ | .055 ($p < .001$) | .770 | .039 |
| 6 | 3169.272 (1425), $p < .001$ | .054 ($p < .01$) | .784 | .036 |
| 7 | 2887.980 (1371), $p < .001$ | .051 ($p = .232$) | .804 | .034 |
| 8 | 2763.813 (1318), $p < .001$ | .051 ($p = .281$) | .806 | .032 |
| 9 | 2741.643 (1266), $p < .001$ | .052 ($p = .063$) | .794 | .029 |
| 10 | 2535.895 (1215), $p < .001$ | .051 ($p = .336$) | .808 | .028 |
| 11 | 2322.189 (1165), $p < .001$ | .048 ($p = .809$) | .824 | .026 |
| 12 | 2465.526 (1116), $p < .001$ | .053 ($p = .023$) | .786 | .024 |
| | | Chi square model comparison tests (corrected for normality violation): | | |
| 1 Factor vs. 2 factors | | 1206.202 (59), $p < .001$ | | |
| 2 Factors vs. 3 factors | | 160.468 (58), $p < .001$ | | |
| 3 Factors vs. 4 factors | | 187.669 (57), $p < .001$ | | |
| 4 Factors vs. 5 factors | | 218.333 (56), $p < .001$ | | |
| 5 Factors vs. 6 factors | | 203.881 (55), $p < .001$ | | |
| 6 Factors vs. 7 factors | | 370.922 (54), $p < .001$ | | |
| 7 Factors vs. 8 factors | | 118.371 (53), $p < .001$ | | |
| 8 Factors vs. 9 factors | | 76.899 (52), $p = .014$ | | |
| 9 Factors vs. 10 factors | | 212.254 (51), $p < .001$ | | |
| 10 Factors vs. 11 factors | | 264.794 (50), $p < .001$ | | |
| 11 Factors vs. 12 factors | | 37.296 (49), $p = .889$ | | |

*Note.* RMSEA = root mean squared error of approximation; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual; *df* = degrees of freedom.

commonly used goodness-of-fit measures, summarized in Table 4. The model chi-square goodness-of-fit test, which tests the null hypothesis that the model is correct for describing the population correlation matrix, is known to be extremely sensitive to small discrepancies between predicted and observed correlations, in case the sample size is large. For this reason, the consistent rejection of good model fit should not come as a surprise. It is however noteworthy that after the extraction of 11 factors no further improvement in the fit can be detected. Not only does the Chi Square test for a comparison between models return a nonsignificant result for the difference in fit between the 11- and the 12-factor solutions, but it is also conspicuous that the 12-factor solution actually becomes worse than the 11-factor solution and even higher solutions do not converge. However, a different story is told both by the RMSEA values, which take into account both goodness-of-fit and parsimony, and the SRMR values. For the two-factor solution, these two indices already have values that could be considered as indicative of reasonable to good fit, although the accompanying test of close fit only becomes nonsignificant after extraction of seven factors. Considering the 90% confidence interval for the population RMSEA value, it runs from .057 to .061 for the two-factor model and from .049 to .054 for the seven-factor solution. The more complex model can therefore be considered to be objectively better on a statistical basis, but the gain in terms of SRMR, the standardized measure of the average residual covariance, is not very large. This can be more clearly understood by examining the number of residual correlations larger than .10. For the two-factor solution, we found 107 out of a total of 1,770 residual correlations to have an absolute residual correlation larger than 0.10. Of these 107 large residuals, 14 even had an absolute value larger than 0.20. On closer inspection, 57 of the large residual correlations could be explained by the fact that they concern consecutive trials (e.g., the residual correlation between volume of the first and that of the second trial) or because they concerned volume and duration trials of the same trial number (e.g., volume of Trial 1 and duration of Trial 1). These 57 trials therefore can be considered methodological artefacts, leaving only 50/1,770 (4.3%) of all residual correlations as being too high without a clear explanation. Compared with the other models, the one-factor solution shows suboptimal

fit values, with RMSEA only being in the acceptable range, the SMRM above the .08 threshold, and a lower TLI.

The last goodness-of-fit measure reported in Table 4, the TLI, consistently returns unsatisfactory low values for all factor models considered. These poor results are inconsistent with the values produced by the RMSEA and SRMR indices, which all reflect good fitting models. It is important to notice that TLI is an incremental measure of goodness-of-fit, which quantifies the performance of a model relative to the independence model, a model that posits no relationships between the variables at population level and can therefore be considered as the worst possible baseline model. In situations where the independence model is not too extravagant for describing the population correlation matrix, the TLI loses its informative value. In our case, we can see that the TLI does not favor any of the models considered, although necessarily more complex models get higher TLI values.

Although the chi-square tests for model comparison identify an 11-factor model as the solution after which further adding of factors no longer give a statistically noticeable improvement of the goodness-of-fit, and the test for close fit suggests the 7-factor model as the solution that has obtained statistical evidence for a close fitting model, RMSEA and SRMR values are already acceptable for the 2-factor solution and change only marginally across the various factor solutions. This means that a premium must be placed on the interpretability of the various factors that can be extracted.

The pattern of loadings for the two-factor solution (see Table 5) approximated a simple structure. The factor loadings of the items on the two-factor solution showed that the volume and duration of Trials 1 to 7 load >.40 on a first, unprovoked factor, whereas the volume and duration of later trials (Trials 8 to 30) load >.40 on a second, provoked factor. Conspicuously, however, the first two provoked trials show approximately equally sized cross loadings on both factors, suggestive of a lingering effect of the previous trials. Evidence for the existence of such lingering effects was also produced by the fact that many residual correlations between consecutive trials were relatively large. Inspection of the factor loading of more complex factor solutions showed that adding factors did not lead to grouping items in a theoretically meaningful way. Specifically, several of the items do not load >.40 on one of the proposed factors (i.e., for the three-factor model, volume of Trials 9, 22, 23 and duration of Trials 9, 15, 22, 25; for the four-factor model, volume of Trials 9, 11, 22, 23, 24 and duration of Trials 8, 9, 10, 15, 22-25; increasing even more in the five- to seven-factor models). Likewise, in the + four factor models, some factors were formed by one item only (i.e., in the four-factor model), while others seemed random, and cannot be theoretically explained at all.

Taken together, we believe that there is some evidence for the statistical superiority of +two-factor solutions; but that at the content level, the only factor solution making sense is that of provoked versus unprovoked aggression. The other characteristics (volume/duration and win/lose) did not have any discernible influence on the pattern of the correlations and did not show up in any of the more complex factor solutions. The two factors are further referred to as "provoked aggression" (scores subsequent to first provocation) and "unprovoked aggression" (scores prior to the first provocation).

Due to the heterogeneous nature of our different data sets (i.e., 30-trial versus 25-trial version; nonpatients vs. forensic patients; participants not suspicious of premeditation/suspicious of premeditation; preceded vs. not preceded by experimental manipulations), additional analyses were conducted to assess the stability of the two-factor structure. Specifically, six new EFA analyses were run (see Table 6). The five EFAs of the subgroup analyses (i.e., in- and excluding patients, participants who first underwent experimental manipulations and participants suspicious of premeditation) revealed two-factor fits identical to that based on the complete data set. When comparing the factorial loading patterns of the five EFA analyses with those of the total EFA, it was revealed that the factor loading of the subanalyses deviated on average .02 from the first-factor loading of the total EFA, and .02 from the second-factor loading of the total EFA. This provided evidence that the two-factor solution is invariant irrespective of the nature of the included participants (patients/nonpatients, suspicious/not), and the precedence by experimental manipulation. In contrast to the previous analyses, the EFA with the independent sample of only the 25-trial CRTT version ($n$ = 180), revealed a one-factor fit. This implies that the 25-trial version cannot be characterized by an unprovoked versus provoked factor, but instead, one composite score averaging all trials is preferable.

## *Derivation of the Two CRTT Subscores*

Based on the EFA analyses described above, we now provide a scoring method for the 30-trial CRTT version that researchers can use to calculate the average scores across all trials with respect to volume (TV) and duration (TD) assessed with the CRTT in their samples. Averaging of the raw scores is justified because (a) the *SD*'s of the variables are comparable, (b) the beta coefficients of their contribution to a factor score are comparable, and (c) the correlations between the average scores and the factor scores are near perfect, Pearson $r$ = .998 for unprovoked, and $r$ = .999 for provoked aggression.

## *Scoring Method 30-Trial Version*

Provoked aggression = (TV8 + TV9 + TV10 + TV11 + TV12 + TV13 + TV14 + TV15 + TV16 + TV17 + TV18 + TV19 + TV20 + TV21 + TV22 + TV23 + TV24 +

**Table 5.** Factor Loadings Based on EFA With Geomin Rotation.

| | Component provoked aggression | Component unprovoked aggression |
|---|---|---|
| Volume Trial 1 | 0.626 | 0.089 |
| Volume Trial 2 | 0.792 | −0.037 |
| Volume Trial 3 | 0.776 | −0.013 |
| Volume Trial 4 | 0.719 | 0.041 |
| Volume Trial 5 | 0.806 | −0.013 |
| Volume Trial 6 | 0.622 | 0.065 |
| Volume Trial 7 | 0.640 | 0.186 |
| Volume Trial 8 | 0.425 | 0.432 |
| Volume Trial 9 | 0.397 | 0.315 |
| Volume Trial 10 | 0.160 | 0.514 |
| Volume Trial 11 | 0.17 | 0.531 |
| Volume Trial 12 | 0.105 | 0.532 |
| Volume Trial 13 | 0.112 | 0.557 |
| Volume Trial 14 | 0.191 | 0.471 |
| Volume Trial 15 | 0.197 | 0.501 |
| Volume Trial 16 | 0.128 | 0.55 |
| Volume Trial 17 | −0.008 | 0.605 |
| Volume Trial 18 | −0.007 | 0.678 |
| Volume Trial 19 | 0.016 | 0.566 |
| Volume Trial 20 | −0.11 | 0.72 |
| Volume Trial 21 | 0.073 | 0.562 |
| Volume Trial 22 | 0.117 | 0.512 |
| Volume Trial 23 | 0.158 | 0.473 |
| Volume Trial 24 | −0.104 | 0.684 |
| Volume Trial 25 | 0.003 | 0.608 |
| Volume Trial 26 | −0.213 | 0.75 |
| Volume Trial 27 | −0.01 | 0.593 |
| Volume Trial 28 | 0.016 | 0.567 |
| Volume Trial 29 | −0.154 | 0.71 |
| Volume Trial 30 | 0.01 | 0.602 |
| Duration trial 1 | 0.643 | 0.007 |
| Duration Trial 2 | 0.669 | −0.022 |
| Duration Trial 3 | 0.727 | −0.001 |
| Duration Trial 4 | 0.684 | 0.052 |
| Duration Trial 5 | 0.764 | 0.026 |
| Duration Trial 6 | 0.684 | −0.003 |
| Duration Trial 7 | 0.595 | 0.16 |
| Duration Trial 8 | 0.325 | 0.458 |
| Duration Trial 9 | 0.415 | 0.301 |
| Duration Trial 10 | 0.13 | 0.474 |
| Duration Trial 11 | 0.169 | 0.477 |
| Duration Trial 12 | 0.12 | 0.437 |
| Duration Trial 13 | 0.1 | 0.471 |
| Duration Trial 14 | 0.089 | 0.518 |
| Duration Trial 15 | 0.166 | 0.431 |
| Duration Trial 16 | 0.039 | 0.529 |
| Duration Trial 17 | 0.006 | 0.534 |
| Duration Trial 18 | 0.093 | 0.555 |
| Duration Trial 19 | 0.055 | 0.52 |
| Duration Trial 20 | −0.099 | 0.643 |
| Duration Trial 21 | −0.065 | 0.565 |
| Duration Trial 22 | 0.057 | 0.447 |
| Duration Trial 23 | 0.031 | 0.459 |
| Duration Trial 24 | −0.09 | 0.654 |
| Duration Trial 25 | 0.013 | 0.545 |
| Duration Trial 26 | −0.283 | 0.734 |
| Duration Trial 27 | −0.048 | 0.614 |
| Duration Trial 28 | −0.019 | 0.576 |
| Duration Trial 29 | −0.16 | 0.721 |
| Duration Trial 30 | 0.019 | 0.546 |

*Note*. EFA = exploratory factor analysis.

TV25 + TV26 + TV27 + TV28 + TV29 + TV30 + TD8 + TD9 + TD10 + TD11 + TD12 + TD13 + TD14 + TD15 + TD16 + TD17 + TD18 + TD19 + TD20 + TD21 + TD22 + TD23 + TD24 + TD25 + TD26 + TD27 + TD28 + TD29 + TD30) / 46.

Unprovoked aggression = (TV1 + TV2 + TV3 + TV4 + TV5 + TV6 + TV7 + TD1 + TD2 + TD3 + TD4 + TD5 + TD6 + TD7) / 14.

### *Reliability of and Intercorrelation Between the Two CRTT Subscales*

Cronbach`s α calculations (30-trial version, $n = 422$) show that all trials prior to the first provocation are reliably measuring the same construct (i.e., unprovoked aggression), $\alpha = .938$, while the same applies to the trials subsequent to the first provocation (i.e., provoked aggression), $\alpha = .962$. The intercorrelation between the provoked and unprovoked scores was Pearson $r = .61^{**}$, $p < .001$.

### *Construct Validity*

The CRTT averaged provoked score correlated significantly with both RPQ reactive and proactive aggression (see Table 7, 30-trial version, $n = 423$) with the first correlation significantly higher than the latter ($Z = 2.01^{*}$, $p = .02$). Provoked CRTT scores also correlated positively with the Exploitiveness/Entitlement scale of the NPI. The CRTT averaged unprovoked score correlated significantly with both RPQ reactive and proactive aggression with no significant differences between both correlations ($Z = .59$, $p = .28$). Unprovoked CRTT scores also correlated positively with both the Superiority/Arrogance and Exploitiveness/Entitlement subscales of the NPI.

## Discussion

The present study addressed the validity of behavioral aggression paradigms. It specifically focused on the CRTT, in which participants are provided with the opportunity to noise-blast an alleged opponent as a form of physical aggression. The field of aggression research needs empirically derived recommendations that can support more standardized use of the CRTT.

### *CRTT Online Administration Tool and Recommendations*

The first aim of the current article was to provide a freely available online CRTT administration tool (Lobbestael et al., 2020). It is available in English, German, and Dutch, and can be administered via online PC access. Data are automatically stored after closing the task. The default version of the administration tool uses the 30-trial CRTT version, which

**Table 6.** Summary of Six Additional EFA Analyses in Subpopulations.

| | | | Initial eigenvalues | | Cumulative percentage of variance | |
|---|---|---|---|---|---|---|
| | *n* | Sample description | Provoked | Unprovoked | Provoked | Unprovoked |
| EFA 1 | 340 | Excluding patients | 32.198 | 7.274 | 32.198 | 39.473 |
| EFA 2 | 366 | Excluding manipulation conditions, i.e., TMS brain stimulation or social exclusion | 33.847 | 7.823 | 33.847 | 41.670 |
| EFA 3 | 283 | Excluding patients and the manipulation conditions, e.g., TMS brain stimulation or social exclusion | 31.790 | 7.598 | 31.790 | 39.388 |
| EFA 4 | 83 | Patient sample only | 40.054 | 8.954 | 40.054 | 49.008 |
| EFA 5 | 405 | Excluding suspicious participants | 32.612 | 7.749 | 32.612 | 40.361 |
| EFA 6 | 180 | 25-Trial only | 37.124 | 6.853 | 37.124 | 43.977 |

*Note*. EFA = exploratory factor analysis; TMS = transcranial magnetic stimulation.

**Table 7.** Pearson Correlations Between CRTT Scores and the RPQ and NPI to Assess Construct Validity, 30-Trial Version.

| | CRTT provoked | CRTT unprovoked |
|---|---|---|
| RPQ, $n = 337$ | | |
| Reactive aggression | .26** | .14* |
| Proactive aggression | .16* | .11* |
| NPI, $n = 153$ | | |
| Total score | .13 | .16 |
| Leadership/authority | −.01 | −.02 |
| Self-absorption/self-administration | .13 | .08 |
| Superiority/arrogance | .08 | .19* |
| Exploitiveness/entitlement | .22* | .25* |

*Note*. CRTT = Competitive Reaction Time Task; RPQ = Reactive Proactive Questionnaire; NPI = Narcissistic Personality Inventory.
*$p < .05$. **$p < .001$.

(compared with the 25-trial version) has the advantage of including multiple assessments of unprovoked aggression (i.e., duration and volume of six trials compared with only one). There are two further advantages of this online CRTT version. First, safeguard elements are integrated to maximize the credibility of the presence of an actual opponent. This is done by (a) having participants lose trials in which their reaction time exceeds 2,000 milliseconds or they press a wrong key, and (b) randomizing the time it takes before the rectangle changes from yellow to red (i.e., where the participant has to wait until the opponent has "set" the duration and volume) in order to induce the impression that an opponent naturally varies in his or her time needed to set the volume and duration sliders. Second, participants are provided with the opportunity for a nonaggressive response (i.e., to refrain from giving an aversive white noise to their opponent; see early criticism on this by Tedeschi and Quigley [1996]). This is done by explicitly informing the participant about this option and demonstrating that they can do this by moving both the volume and durations sliders but then put them back to zero. This avoids demand characteristics (i.e., the participant feeling obliged by the instructions to provide the opponent with some form of aversive feedback).

Aside from this, we formulate two other recommendations for using the CRTT that will foster adequate ethical use of the CRTT and increase replicability of research findings. First, each personal computer and the volume settings should be pretested with a decibel meter to ensure that the maximum sound (i.e., Level 10) indeed does not exert 100 dB(A). 100 dB(A) exceeds the United States Safety and Health Standards recommendations for sustained 8-hour exposure of 90 dB(A) for full-time workers, however is well within the pain threshold of 125 dB(A). In Europe, the sustained 8-hour expose is set at 80 dB(A), and the pain threshold at 120 dB(A). Participants with hearing problems should be excluded from participation. Second, researchers should publish the precise trial specifications implemented in their CRTT. This practice is necessary to stimulate methodological transparency and facilitates replication studies.

## *CRTT Scoring Algorithm*

Literature evidences high levels of methodological heterogeneity in CRTT studies due to large variations in cover stories, trial specifications, and outcome scores. While such heterogeneity is not unique to the CRTT or to aggression

research (e.g., see Scarpina & Tagini, 2017; Nyongesa et al., 2019, for similar issues with the Stroop test and executive functioning tasks), such flexibility in data collection, analyses, and reporting has been shown to increase false-positive findings (Simmons et al., 2011). To overcome this problem of abundant CRTT scoring methods, the second main goal of this study was to present an empirically derived CRTT scoring algorithm. EFA in a large sample evidenced the presence of two distinct CRTT factors in the 30-trial version, that is, provoked aggression (volume and duration scores subsequent to first provocation) and unprovoked aggression (volume and duration scores prior to first provocation).

The provoked CRTT factor showed to be particularly correlated to its' self-reported reactive counterpart, as well as to the Exploitiveness/Entitlement subcomponent of grandiose narcissism, thereby providing strong initial support for its construct validity. Moderate construct validity was shown for the unprovoked CRTT score in that it correlated comparably to both self-reported reactive and proactive aggression, as well as to the Exploitiveness/Entitlement and Superiority/Arrogance subscales of grandiose narcissism. The lack of a unique relationship between CRTT unprovoked aggression and self-reported proactive aggression can likely be ascribed to the fact that the latter construct incorporates a reward-driven focus that is less pronounced in the CRTT. It would be interesting for future studies to further address the external validity of the two CRTT scores by assessing their correlations with other real-world behavioral aggression outcomes (e.g., violent crimes, antisocial behavior). To contribute toward making aggression assessment more valid and standardized, we advise researchers using the 30-trial CRTT to apply our evidence-based scoring method (i.e., averaging all preprovocation vs. all postprovocation trials). We also advice researchers to preregister this intent. In contrast to the 30-trial version, the 25-trial CRTT version showed to be underlined by one single factor, clustering the duration and volume of both provoked and unprovoked trials, irrespective of whether these were preceded by win or lose experiences. This implies that unprovoked aggression can only be assessed in a valid way as a separate construct with the 30-trial CRTT version where multiple (in this case, seven) unprovoked trials are presented to participants, before switching to provoked aggression.

All EFA findings indicate that both duration and volume load on the same factor. This is in line with the only other sufficiently powered principal component analysis report which also delineated duration and volume as indicators of a single component (Chester & Lasko, 2019) and with reports of high intercorrelations between duration and volume levels (e.g., Hahn-Holbrook et al., 2011). However, preliminary findings from a previous study puzzle in suggesting that only noise duration is related to increased planning abilities and verbal intelligence (Ferguson et al., 2008). Clearly, more studies are needed to determine which correlates are shared by CRTT volume and duration scores. The factors derived from Chester and Lasko (2019)'s data deviates from our findings in that their data did not support differentiating unprovoked from provoked trials (i.e., only one factor was found). This outcome is probably due to the fact that the first trial in Chester and Lasko (2019)'s CRTT version was based on the opponent setting maximum volume and duration right away, which renders the assessment of unprovoked aggression impossible and contrasts with our 30-trial CRTT version with started with several unprovoked (i.e., duration and volume 0) trials.

## *Suggestions for Future Studies*

While it is generally assumed to be crucial for the validity of the CRTT that participants believe its cover story, systematic research on how to (a) best effectuate this and (b) assess the impact of believe in the cover story on research findings is currently lacking. The most effective way to maximize the likelihood that participants believe in the actual presence of an opponent is to introduce participants to real-life confederates before initiating the CRTT. This implies extensive training of the confederates to ensure that they act equally toward each participant. Using this method of social deception in our lab enabled us to entirely avoid dropout due to disbelief in the cover story (Dambacher et al., 2014), compared with the usual 3% to 10%. In case resources do not allow for the physical presence of confederates (i.e., restricted budget or availability of confederates), the likelihood that participants believe in the actual presence of an opponent can be maximized by (a) urging the participant to start the CRTT because the opponent is waiting; (b) telling the participant that the opponent is late and, therefore, cannot be introduced to the participant; (c) having the participant "overhear" the experimenter giving instructions to the other opponent in an adjunct room by talking loud or using an audiotape; (d) leaving some belongings (jacket/bag) of the "opponent" in the hallway; and/or (e) introducing the participant to the opponent via a webcam or a supposedly previously taken picture. An important avenue for future research is to empirically assess the impact of participants actually meeting an opponent, or of other actions to increase the believability of actual presence of an opponent on deception belief. Such studies could randomly assign participants to a condition where they physically meet their opponent versus where efforts are maximized to convince participants of the presence of an opponent without a physical encounter, and compare these conditions with a condition in which no credibility measures are implemented. Furthermore, at the end of any CRTT study, participants should be carefully probed for suspicion about

the cover story. In our experience, it is important to use open and vague probing questions (e.g., "What was your impression about your opponent?" "Are there any other things that you noticed about the opponent or the experiment?") to avoid participants becoming suspicious by the questions themselves. We always have two raters categorize the participants as either nonsuspicious (i.e., participants do not report any doubt about the presence of an actual opponent), possibly suspicious (i.e., participants express doubts about whether an opponent was actually there; participants mention they might have been playing against a computer), or suspicious (i.e., participants report to be sure that there is no real other opponent). As the impact of excluding suspicious participants from data analyses has not been systematically addressed, the best way for now seems to report data both in- and excluding (possibly) suspicious participants in order to be transparent about respective effects on the findings. Studies complying with this suggestion (e.g., Gitter et al., 2013; Lobbestael et al., 2014) found that excluding suspicious participants did not or only marginally impacted the findings. Future studies are needed that empirically address how much knowledge on the side of the participants is acceptable. A final general recommendation for CRTT studies is to include a sufficient number of participants. A recent review on the CRTT-personality relationship (Hyatt et al., 2019) pointed out that many CRTT studies are underpowered to assess bivariate relationships, let alone interaction effects (see also Hyatt et al., 2019).

### *Strengths and Limitations*

To our knowledge, no previous study has resorted to the factor-analytic approach of analyzing CRTT data using sufficiently large samples, which we employed in this article. We, nevertheless, want to proactively acknowledge the limitations of this project. A first limitation concerns the fact that the discovered factor structure is based on our eight studies all using the same volume, duration, and win/loss CRTT settings. Recently, Elson et al. (2014) assessed whether CRTT outcomes could be explained by the opponent's volume, duration and win/lose settings in the previous trials and concluded that around one third of the variance of participant's responses was explained by such consecution effects. Likewise, Anderson et al. (2004) showed that ambiguous noise pattern led to significantly different CRTT responses compared with increasing noise patterns. Whether the same factor structure underlies CRTT versions with different specifications is an open empirical question. The second limitation of the current study is the use of a predominantly male Caucasian sample. Replications in more gender- and culture-balanced samples are, therefore, needed.

### *Implications*

The use of valid aggression assessment has crucial implications for the clinical, forensic, and medical field. Ben-Porath and Taylor (2002), for example, randomly assigned participants to either a diazepam administration or placebo condition and had them engage in the CRTT and concluded that diazepam-intoxication caused increased aggression. Likewise, a substantial part of studies on the impact of violent media on aggressive acts is based on the CRTT (Anderson et al., 2010), so the quality of expert witness' defenses inherently depends on the reliability and validity of behavioral aggression assessment. Until we improve our standards of aggression assessment—echoing Elson et al. (2014) and Ferguson and Rueda (2009)—we warn of drawing premature implications based on CRTT studies.

## Conclusion

In conclusion, the current study addressed and pinpointed several prominent threats to the validity of behavioral aggression assessment using the CRTT. These include large variabilities in cover stories, trial specifications, and outcome quantification strategies. Admittedly, it is precisely this versatility that makes the CRTT an appealing research tool for many distinct subtopics within the field of aggression research. Nonetheless, it is our task as researchers to retain a critical attitude toward its validity. Our factor analytic findings support the superiority of one of the existing CRTT scoring methods for the 30-trial versions, that is, averaging all preprovocation versus all postprovocation trials. By providing researchers with an empirically derived scoring method, an online administration tool, concrete recommendations for using the CRTT, and avenues for future research, we want to contribute to further stimulate and ameliorate accurate and standardized behavioral aggression assessment.

### Acknowledgments

We are grateful for the feedback of the anonymous reviewers on this article, which substantially improved the structure of this article and the statistical rigor of the analyses.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO; HCMI 10-19).

### ORCID iD


Jill Lobbestael https://orcid.org/0000-0001-9205-3115

## References


American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Author

Anderson, C. A., & Anderson, K. B. (2008). Men who target women: Specificity of target, generality of aggressive behavior. *Aggressive Behavior*, *34*(6), 605-622. https://doi.org/10.1002/ab.20274

Anderson, C. A., Carnagey, N. L., Flanagan, M., Benjamin, A. J., Eubanks, J., & Valentine, J. C. (2004). Violent video games: Specific effects of violent content on aggressive thoughts and behavior. In M. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 36, pp. 199-249). Elsevier.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R., & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151-173. https://doi.org/10.1037/a0018251

Baker, C., Wuest, J., & Noerager Stern, P. (1992). Method slurring: The grounded theory/phenomenology example. *Journal of Advanced Nursing*, *17*(11), 1355-1360. https://doi.org/10.1111/j.1365-2648.1992.tb01859.x

Banse, R., Messer, M., & Fischer, I. (2015). Predicting aggressive behavior with the aggressiveness-IAT. *Aggressive Behavior*, *41*(1), 65-83. https://doi.org/10.1002/ab.21574

Ben-Porath, D. D., & Taylor, S. P. (2002). The effects of diazepam (valium) and aggressive disposition on human aggression: an experimental investigation. *Addictive Behaviors*, *27*(2), 167-177. https://doi.org/10.1016/S0306-4603(00)00175-1

Bernstein, S., Richardson, D., & Hammock, G. (1987). Convergent and discriminant validity of the Taylor and Buss measures of physical aggression. *Aggressive Behavior*, *13*(1), 15-24. https://doi.org/10.1002/1098-2337(1987)13:1<15::AID-AB2480130104>3.0.CO;2-K

Brugman, S., Lobbestael, J., Arntz, A., Cima, M., Schuhmann, T., Dambacher, F., & Sack, A. T. (2015). Identifying cognitive predictors of reactive and proactive aggression. *Aggressive Behavior*, *41*(1), 51-64. https://doi.org/10.1002/ab.21573

Brugman, S., Lobbestael, J., Cima, M., Schuhmann, T., Dambacher, F., Sack, A. T., & Arntz, A. (2018). Cognitive predictors of reactive and proactive aggression in a forensic sample: A comparison with a non-clinical sample. *Psychiatry Research*, *269*(November), 610-620. https://doi.org/10.1016/j.psychres.2018.08.095

Bushman, B. J. (1995). Moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of Personality and Social Psychology*, *69*(5), 950-960. https://doi.org/10.1037/0022-3514.69.5.950

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology*, *75*(1), 219-229. https://doi.org/10.1037/0022-3514.75.1.219

Cherek, D. R., Moeller, F. G., Schnapp, W., & Dougherty, D. M. (1997). Studies of violent and nonviolent male parolees: I. Laboratory and psychometric measurements of aggression. *Biological Psychiatry*, *41*(5), 514-522. https://doi.org/10.1016/S0006-3223(96)00059-5

Chester, D. S., & Lasko, E. N. (2019). Validating a standardized approach to the Taylor Aggression Paradigm. *Social Psychological and Personality Science*, *10*(5), 620-631. https://doi.org/10.1177/1948550618775408

Cima, M., & Raine, A. (2009). Do distinct characteristics of psychopathy relate to different subtypes of aggression? *Personality and Individual Differences*, *47*(8), 835-840. https://doi.org/10.1016/j.paid.2009.06.031

Cima, M., Raine, A., Meesters, C., & Popma, A. (2013). Validation of the Dutch Reactive Proactive Questionnaire (RPQ): Differential correlates of reactive and proactive aggression from childhood to adulthood. *Aggressive Behavior*, *39*(2), 99-113. https://doi.org/10.1002/ab.21458

Coie, J. D., Dodge, K. A., Terry, R., & Wright, V. (1991). The role of aggression in peer relations: An analysis of aggression episodes in boys' play groups. *Child Development*, *62*(4), 812-826. https://doi.org/10.2307/1131179

Dambacher, F., Sack, A. T., Lobbestael, J., Arntz, A., Brugman, S., & Schuhmann, T. (2014). Out of control: Evidence for anterior insula involvement in motor impulsivity and reactive aggression. *Social Cognitive and Affective Neuroscience*, *10*(4), 508-516. https://doi.org/10.1093/scan/nsu077

Dambacher, F., Schuhmann, T., Lobbestael, J., Arnzt, A., Brugman, S., & Sack, A. T. (2015a). No effects of bilateral tDCS over inferior frontal gyrus on response inhibition and aggression. *PLOS ONE*, *10*(7), Article e0132170. https://doi.org/10.1371/journal.pone.0132170

Dambacher, F., Schuhmann, T., Lobbestael, J., Arntz, A., Brugman, S., & Sack, A. T. (2015b). Reducing proactive aggression through non-invasive brain stimulation. *Social Cognitive and Affective Neuroscience*, *10*(10), 1303-1309. https://doi.org/10.1093/scan/nsv018

DeWall, C. N., Baumeister, R. F., Stillman, T. F., & Gailliot, M. T. (2007). Violence restrained: Effects of self-regulation and its depletion on aggression. *Journal of Experimental Social Psychology*, *43*(1), 62-76.

DeWall, C. N., Finkel, E. J., Lambert, N. M., Slotter, E. B., Bodenhausen, G. V., Pond, R. S., Renzetti, C. M., & Fincham, F. D. (2013). The voodoo doll task: Introducing and validating a novel method for studying aggressive inclinations. *Aggressive Behavior*, *39*(6), 419-439. https://doi.org/10.1002/ab.21496

Elson, M. (2016). *Competitive reaction time task*. http://crtt.flexiblemeasures.com/

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2015). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, *4*(2), 112-125. https://doi.org/10.1037/ppm0000010

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, *26*(2), 419-432. https://doi.org/10.1037/a0035569

Emmons, R. A. (1987). Narcissism: Theory and measurement. *Journal of Personality and Social Psychology*, *52*(1), 11-17. https://doi.org/10.1037/0022-3514.52.1.11

Ferguson, C. J., & Rueda, S. M. (2009). Examining the validity of the modified Taylor competitive reaction time test

of aggression. *Journal of Experimental Criminology*, *5*(2), 121-137. https://doi.org/10.1007/s11292-009-9069-5

Ferguson, C. J., Smith, S., Miller-Stratton, H., Fritz, S., & Heinrich, E. (2008). Aggression in the laboratory: Problems with the validity of the modified Taylor Competitive Reaction Time Test as a measure of aggression in media violence studies. *Journal of Aggression, Maltreatment & Trauma*, *17*(1), 118-132. https://doi.org/10.1080/10926770802250678

Foster, E. M., & Jones, D. E. (2005). The high costs of aggression: Public expenditures resulting from conduct disorder. *American Journal of Public Health*, *95*(10), 1767-1772. https://doi.org/10.2105/AJPH.2004.061424

Giancola, P. R., & Parrott, D. J. (2008). Further evidence for the validity of the Taylor aggression paradigm. *Aggressive Behavior*, *34*(2), 214-229. https://doi.org/10.1002/ab.20235

Giancola, P. R., & Zeichner, A. (1995). Construct validity of a competitive reaction-time aggression paradigm. *Aggressive Behavior*, *21*(3), 199-204. https://doi.org/10.1002/1098-2337(1995)21:3<199::AID-AB2480210303>3.0.CO;2-Q

Gitter, S. A., Ewell, P. J., Guadagno, R. E., Stillman, T. F., & Baumeister, R. F. (2013). Virtually justifiable homicide: The effects of prosocial contexts on the link between violent video games, aggression, and prosocial and hostile cognition. *Aggressive Behavior*, *39*(5), 346-354. https://doi.org/10.1002/ab.21487

Hahn-Holbrook, J., Holt-Lunstad, J., Holbrook, C., Coyne, S. M., & Lawson, E. T. (2011). Maternal defense: Breast feeding increases aggression by reducing stress. *Psychological Science*, *22*(10), 1288-1295. https://doi.org/10.1177/0956797611420729

Howell, D. C. (2007). *Statistical methods for psychologists* (6th ed.). Duxbury.

Hyatt, C. S., Chester, D. S., Zeichner, A., & Miller, J. D. (2019). Analytic flexibility in laboratory aggression paradigms: Relations with personality traits vary (slightly) by operationalization of aggression. *Aggressive Behavior*, *45*(4), 377-388. https://doi.org/10.1002/ab.21830

Hyatt, C. S., Zeichner, A., & Miller, J. D. (2019). Laboratory aggression and personality traits: A meta-analytic review. *Psychology of Violence*, *9*(6), 675-689. https://doi.org/10.1037/vio0000236

Imhoff, R., Bergmann, X., Banse, R., & Schmidt, A. F. (2013). Exploring the automatic undercurrents of sexual narcissism: Individual differences in the sex-aggression link. *Archives of Sexual Behavior*, *42*(6), 1033-1041. https://doi.org/10.1007/s10508-012-0065-x

Lansu, T. A., Cillessen, A. H., & Sandstrom, M. J. (2014). From classroom to dyad: Actor and partner effects of aggression and victim reputation. *Social Development*, *23*(4), 651-665. https://doi.org/10.1111/sode.12055

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*(2), 230-243. https://doi.org/10.1037/pspi0000049

Lieberman, J. D., Solomon, S., Greenberg, J., & McGregor, H. A. (1999). A hot new way to measure aggression: Hot sauce allocation. *Aggressive Behavior*, *25*(5), 331-348. https://doi.org/10.1002/(SICI)1098-2337(1999)25:5<331::AID-AB2>3.0.CO;2-1

Lobbestael, J., Baumeister, R. F., Fiebig, T., & Eckel, L. A. (2014). The role of grandiose and vulnerable narcissism in self-reported and laboratory aggression and hormonal reactivity. *Personality and Individual Differences*, *69*(October), 22-27. https://doi.org/10.1016/j.paid.2014.05.007

Lobbestael, J., & Brugman, S. (in preparation). *Comparison of four behavioral aggression paradigms* [Manuscript in preparation].

Lobbestael, J., Emmerling, F., Brugman, S., Sack, A. T., Schuhmann, T., Arntz, A., Benning, R., & Bonnemayer, C. (2020). *Competitive Reaction Time Task (CRTT), online administration program*. Maastricht University.

Lobbestael, J., & Vancleef, L. (in preparation). *The effect of psychopathic traits on pain and aggression* [Manuscript in preparation].

McCarthy, R. J., & Elson, M. (2018). A conceptual review of lab-based aggression paradigms. *Collabra: Psychology*, *4*(1). Article 4. https://doi.org/10.1525/collabra.104

Moss, J. H., & Maner, J. K. (2016). Biased sex ratios influence fundamental aspects of human mating. *Personality and Social Psychology Bulletin*, *42*(1), 72-80. https://doi.org/10.1177/0146167215612744

Murphy, D. A., Pelham, W. E., & Lang, A. R. (1992). Aggression in boys with attention deficit-hyperactivity disorder: Methylphenidate effects on naturalistically observed aggression, response to provocation, and social information processing. *Journal of Abnormal Child Psychology*, *20*(5), 451-466. https://doi.org/10.1007/BF00916809

Muthén, L. K., & Muthén, B. O. (1998-2020). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Nyongesa, M. K., Ssewanyana, D., Mutindi, A., Chongwo, E., Scerif, G., Newton, C., & Abubakar, A. (2019). Assessing executive function in adolescence: A scoping review of existing measures and their psychometric robustness. *Frontiers in Psychology*, *10*, Article 311. https://doi.org/10.3389/fpsyg.2019.00311

O'Connor's, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, & Computers*, *32*(3), 396-402. https://doi.org/10.3758/BF03200807

Pickett, S. M., Parkhill, M. R., & Kirwan, M. (2016). The influence of sexual aggression perpetration history and emotion regulation on men's aggressive responding following social stress. *Psychology of Men & Masculinity*, *17*(4), 363-372. https://doi.org/10.1037/men0000032

Player, M. S., King, D. E., Mainous, A. G., & Geesey, M. E. (2007). Psychosocial factors and progression from prehypertension to hypertension or coronary heart disease. *Annals of Family Medicine*, *5*(5), 403-411. https://doi.org/10.1370/afm.738

Polman, H., de Castro, B. O., Koops, W., van Boxtel, H. W., & Merk, W. W. (2007). A meta-analysis of the distinction between reactive and proactive aggression in children and adolescents. *Journal of Abnormal Child Psychology*, *35*(4), 522-535. https://doi.org/10.1007/s10802-007-9109-4

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., Stouthamer-Loeber, M., & Liu, J. (2006). The reactive–proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys.

*Aggressive Behavior*, *32*(2), 159-171. https://doi.org/10.1002/ab.20115

Raskin, R. N., & Hall, C. S. (1979). A Narcissistic Personality Inventory. *Psychological Reports*, *45*(2), 590. https://doi.org/10.2466/pr0.1979.45.2.590

Raskin, R. N., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*(5), 890-902. https://doi.org/10.1037/0022-3514.54.5.890

Ritter, D., & Eslea, M. (2005). Hot sauce, toy guns, and graffiti: A critical account of current laboratory aggression paradigms. *Aggressive Behavior*, *31*(5), 407-419. https://doi.org/10.1002/ab.20066

Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in psychology*, *8*, Article 557. https://doi.org/10.3389/fpsyg.2017.00557

Schmidt, A. F., Zimmermann, P. S., Banse, R., & Imhoff, R. (2015). Ego depletion moderates the influence of automatic and controlled precursors of reactive aggression. *Social Psychology*, *46*, 132-141. https://doi.org/10.1027/1864-9335/a000233

Sestir, M. A., & Bartholow, B. D. (2010). Violent and nonviolent video games produce opposing effects on aggressive and prosocial outcomes. *Journal of Experimental Social Psychology*, *46*(6), 934-942. https://doi.org/10.1016/j.jesp.2010.06.005

Sherrill, A. M., Magliano, J. P., Rosenbaum, A., Bell, K. M., & Wallace, P. S. (2016). Trait aggressiveness and aggressive behavior in the context of provocation and inhibition. *Journal of Aggression, Maltreatment & Trauma*, *25*(5), 487-502. https://doi.org/10.1080/10926771.2015.1121192

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, *35*(2), 297-310. https://doi.org/10.1111/j.1467-6494.1967.tb01430.x

Tedeschi, J. T., & Quigley, B. M. (1996). Limitations of laboratory paradigms for studying aggression. *Aggression and Violent Behavior*, *1*(2), 163-177. https://doi.org/10.1016/1359-1789(95)00014-3

Thomaes, S., Bushman, B. J., Stegge, H., & Olthof, T. (2008). Trumping shame by blasts of noise: Narcissism, self-esteem, shame, and aggression in young adolescents. *Child Development*, *79*(6), 1792-1801. https://doi.org/10.1111/j.1467-8624.2008.01226.x

Van Teffelen, M. W., Vancleef, L., & Lobbestael, J. (in press). Ego-threatening provocations in individuals with narcissistic and psychopathic traits: A comparison of social exclusion and insult. *Psychology of Violence*.

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema*, *24*(2), 310-315. http://www.psicothema.com/pdf/4016.pdf

Warburton, W. A., & Bushman, B. J. (2019). The competitive reaction time task: The development and scientific utility of a flexible laboratory aggression paradigm. *Aggressive Behavior*, *45*(4), 389-396. https://doi.org/10.1002/ab.21829

Weisbuch, M., Beal, D., & O'neal, E. C. (1999). How masculine ought I be? Men's masculinity and aggression. *Sex Roles*, *40*(7-8), 583-592. https://doi.org/10.1023/A:1018840130646

Whitaker, J. L., & Bushman, B. J. (2012). "Remain calm. Be kind." Effects of relaxing video games on aggressive and prosocial behavior. *Social Psychological and Personality Science*, *3*(1), 88-92. https://doi.org/10.1177/1948550611409760

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A .L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-Hacking. *Frontiers in Psychology*, *7*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkowski, B. M., Crowe, S. E., & Ferguson, E. L. (2015). Learning to keep your cool: Reducing aggression through the experimental modification of cognitive control. *Cognition and Emotion*, *29*(2), 251-265. https://doi.org/10.1080/02699931.2014.911146

Wilkowski, B. M., Robinson, M. D., & Troop-Gordon, W. (2010). How does cognitive control reduce anger and aggression? The role of conflict monitoring and forgiveness processes. *Journal of Personality and Social Psychology*, *98*(5), 830-840. https://doi.org/10.1037/a0018962