Press CRTT to measure aggressive behavior: the unstandardized use of the competitive reaction time task in aggression research

Elson, Malte; Mohseni, M. Rohangis; Breuer, Johannes; Scharkow, Michael; Quandt, Thorsten

Postprint / Postprint
Zeitschriftenartikel / journal article

This is an accepted manuscript of an article published by American Psychological Association in
Psychological Assessment, Vol. 26, 2014, No. 2, pp. 419-432
available online at: https://doi.org/10.1037/a0035569

Press CRTT to Measure Aggressive Behavior: The Unstandardized Use of the Competitive Reaction Time Task in Aggression Research

Malte Elson, University of Münster

Johannes Breuer, University of Münster

M. Rohangis Mohseni, University of Osnabrück

Michael Scharkow, University of Hohenheim

Thorsten Quandt, University of Münster

The Competitive Reaction Time Task (CRTT) is the measure of aggressive behavior most commonly used in laboratory research. However, the test has been criticized for issues in standardization because there are many different test procedures and at least 13 variants to calculate a score for aggressive behavior. We compared the different published analyses of the CRTT using data from 3 different studies to scrutinize whether it would yield the same results. The comparisons revealed large differences in significance levels and effect sizes between analysis procedures, suggesting that the unstandardized use and analysis of the CRTT have substantial impacts on the results obtained, as well as their interpretations. Based on the outcome of our comparisons, we provide suggestions on how to address some of the issues associated with the CRTT, as well as a guideline for researchers studying aggressive behavior in the laboratory.

Keywords: aggression, standardization, measurement, methodology, Competitive Reaction Time Task
Supplemental materials

Supplemental materials: http://dx.doi.org/10.1037/a0035569.supp

Researchers have been looking into predictors and antecedents of aggression for a long time, but the field is still struggling to find valid methods to measure aggressive behaviors in the laboratory. The problem of measurement already begins with finding an appropriate definition of aggression. Baron and Richardson (1994) defined aggression as any behavior that is intended to cause harm to another person who intends to avoid this harm. To distinguish between aggression and violence, Ferguson and Rueda (2009) pointed out that "violent behaviors … are typically restricted to acts which are intended to cause serious physical harm" (p. 121), while "aggression as a class of behavior is much broader than violent behavior and can include numerous acts... which are neither physically injurious nor illegal" (pp. 121-122). Similar to the difficulty of defining aggression, finding a valid, reliable, and ethically acceptable measure of aggressive behavior for laboratory research is an intricate undertaking.

Despite such difficulties, numerous studies have been conducted and published in which laboratory measurements are employed to provide evidence that a large variety of stimuli increase aggressiveness. These stimuli include displayed violence in digital games (Anderson & Dill, 2000) and gory passages from the Bible (Bushman, Ridge, Das, Busath, & Key, 2007) but also alcohol (Bond & Lader, 1986) and other psychoactive drugs (Bond & Lader, 1988). The test most commonly used in laboratory research on aggressive behavior in adults, especially in studies on media violence, is a modification of Taylor's (1967) Competitive Reaction Time Task (CRTT). In this article, we briefly describe this test and how it has been used and modified for aggression research. Afterward, by applying the different analyses to data from three independent studies that employed the CRTT as a measure of aggressive behavior, we show that the many analysis procedures that have been used for the CRTT can lead to contradictory results even within the same data set.

Malte Elson, Department of Communication, University of Münster, Münster, Germany; M. Rohangis Mohseni, Center for Information Management and Virtual Teaching, University of Osnabrück, Osnabrück, Germany; Johannes Breuer, Department of Communication, University of Münster, Münster, Germany; Michael Scharkow, Department of Communication Studies, University of Hohenheim, Stuttgart, Germany; Thorsten Quandt, Department of Communication, University of Münster, Münster, Germany.

The research leading to the results of Studies 2 and 3 has received funding from the European Union's Seventh Framework Programme (FP7/ 2007-2013) under Grant 240864 (SOFOGA).

Correspondence concerning this article should be addressed to Malte Elson, Department of Communication, University of Münster, Bispinghof 9-14, 48143 Münster, Germany. E-mail: malte.elson@uni-muenster.de

The Competitive Reaction Time Task

In the original version of the CRTT by Taylor (1967), the Taylor Aggression Paradigm, participants were led to believe that they would be playing a reaction time game against another participant in which the winner of a round would punish the loser with an electric shock. Participants who lost a round would receive shocks of varying intensity and when participants won a round they could adjust the shock levels for their alleged opponents. The intensity level of the shock was used as the measure for aggressiveness. Recent adaptations of the CRTT allow participants to set the intensity (usually volume and/or duration) of a noise blast instead of an electric shock (e.g., Bushman, 1995), as they are easier to use and bring up fewer ethical issues (Ferguson & Rueda, 2009). As there is no real opponent, the sequence of wins and losses, as well as the settings "chosen" by the opponent, are typically randomized and preset. Generally, louder and longer noise blasts are considered indicators of higher levels of aggressiveness.

There has been disagreement among aggression researchers about the validity of the CRTT. This debate spans different issues, including alternative motives for high or low settings (e.g., reciprocity or deterrence), demand characteristics of the experimental situation, confusion of competition and aggression, construct validity and external validity, the (psychological) distance between participant and opponent, a lack of alternative (i.e., nonviolent) responses, and the absence of social sanctions for aggression. However, since the present study focuses on issues of standardization, we refer the reader to the available literature that discusses all of the topics outlined above (Ferguson & Rueda, 2009; Giancola & Chermack, 1998; Giancola & Zeichner, 1995a; Ritter & Eslea, 2005; Savage, 2004; Tedeschi & Quigley, 1996, 2000).

Clinical Relevance

Although we are aware that the CRTT has not been developed as a clinical instrument, the test and its psychometric properties are relevant to clinical research for two main reasons:

1. The CRTT (in one form or another) is being used to answer research questions in the area of clinical psychology. This includes investigations of social and cerebral response in criminal psychopaths (Veit et al., 2010); effectiveness of prescription drugs in reducing hostility in panic disorders (Bond, Curran, Bruce, O'Sullivan, & Shine, 1995); and the facilitation of aggression through various substances, such as alcohol (Bond & Lader, 1986; Pihl et al., 1995) or prescription drugs (Bond & Lader, 1988; Weisman, Berman, & Taylor, 1998). In some cases, practical recommendations for clinicians regarding the diagnosis (McCloskey, Berman, Noblett, & Coccaro, 2006) and treatment (Ben-Porath & Taylor, 2002) of patients are made based on results obtained with the CRTT. Given the impact of clinical research on the definition, assessment, diagnosis, and treatment of disorders in clinical practice, the importance of using objective, reliable, and valid measures cannot be overstated. Accordingly, Ferguson and Rueda (2009) warned against using the CRTT for clinical research and making any statements about clinical implications based on findings obtained with the CRTT (particularly versions that use noise blasts instead of electroshocks).

2. Results from nonclinical fields of study in which the CRTT is particularly popular, such as media effects research, are frequently generalized into the clinical realm. Bushman and Gibson (2011), for example, describe the CRTT as "a weapon that could be used [by the participants] to blast their partner" (p. 30). Bushman et al. (2007) found support for "theories proposed by scholars of religious terrorism who hypothesize that exposure to violent scriptures may induce extremists to engage in aggressive actions" (p. 204) in their study on university students' noise blasts after reading a passage from the Bible in which God sanctions violence. They conclude that "to the extent that religious extremists engage in prolonged, selective reading of the scriptures, focusing on violent retribution toward unbelievers instead of the overall message of acceptance and understanding, one might expect to see increased brutality" (p. 205). Some scholars also warn about a public health risk and equate the effects of media violence on aggression to those of substance use, abusive parents, or poverty (Anderson et al., 2010). Anderson, Gentile, and Buckley (2007), for example, consider violent video games as one of several risk factors that may cause aggressive, violent behavior and, in highly extreme and rare cases, even school shootings. Sometimes scholars make direct comparisons with other research areas and, e.g., claim that the effects of media violence are similar to those of smoking on lung cancer (Bushman & Anderson, 2001). Given that such statements are being made in the scientific literature, it comes as little surprise that similar conclusions have been found in the public debate. In a recent court case, a prominent media violence researcher was paid by the defense to serve as an expert witness and he argued that a teenage victim of a brutal multiple murder was the perpetrator of that murder, in part because he played digital games (Rushton, 2013). As the CRTT is used in a majority of experiments in this field (Anderson et al., 2010), and taking into account the lack of standardization and data supporting the validity of CRTT score interpretations (Ferguson & Rueda, 2009), such conclusions or comparisons are particularly inappropriate.

Standardization

Our primary concern is the lack of standardization of both test procedure and data analysis for the CRTT. The test

has been used in many different versions, sometimes even by the same authors (Ferguson, 2011). Due to this lack of standardization, Ferguson, Smith, Miller-Stratton, Fritz, and Heinrich (2008) suggested a standard procedure for the CRTT. This version consists of 25 trials, and the sequence of losses and wins, as well as the computer's settings of volume and duration, is preset and randomized, so that the same pattern exists for every participant. The settings of the computerized opponent are always displayed to the participant after each trial. The average volume and duration settings are treated as separate and equal scores for aggressive behavior, although the authors express their concerns about using duration settings at all in future studies. Until now, however, only a few researchers have adopted the version proposed by Ferguson et al. (2008).

Inconsistencies are found in the procedure of the CRTT, as well as in the ways in which the CRTT data are analyzed by different (and sometimes even the same) authors. While the procedural aspects refer to the setup of the test, i.e., how the raw data are generated, the statistical differences refer to how the data are analyzed. Of course, the procedural decisions also affect the options for statistical analyses. The following section is meant to list the variations of both procedures and statistical analyses of the CRTT that can be found throughout the literature.

Bushman and Baumeister (1998) used the sum of volume and duration settings in the first of 25 trials. They provided the reason for choosing this procedure in a footnote stating that the first trial is the only noise setting that is “unprovoked,” (p. 222) and that in all subsequent trials the participants' settings in the study converged with those of their opponents (i.e., there was no difference between conditions).

Anderson and Dill (2000) calculated separate aggression scores from average volume and duration settings for each possible trial outcome (winning and losing), resulting in a total of four scores for aggressiveness. They expected retaliatory behavior to be higher after losing a trial. In addition, they log-transformed the duration data because they were positively skewed, but reported and interpreted the original mean differences as being significantly different between conditions. Although three out of four parameters yielded nonsignificant differences, they asserted a “convergence of findings” (p. 787) in their discussion.

The use of four separate parameters is also different from the procedure of Lindsay and Anderson (2000), who multiplied volume with log-transformed duration settings. The average over 25 trials of those products was their measure for overall aggression. Carnagey and Anderson (2005) averaged the products of volume and the square root of duration to form a single “aggressive energy score” (p. 887). No reason is given for this other than the claim that this single score supposedly is a valid measure and that duration should be square rooted. By contrast, Anderson and Carnagey (2009) only allowed their participants to set the volume, and they used two indicators for aggression (mean volume and the total number of volume settings of 8 to 10 on a total scale of 1 to 10). The reasons given for using this number of high volume settings were that they were more clearly aggressive, more likely to instigate retaliation, and easier to communicate to nonexperts.

The same inconsistencies in CRTT procedure can also be observed in other publications. Bartholow, Sestir, and Davis (2005) multiplied the average volume and duration settings to form a composite aggressive behavior score. Although Bartholow, Bushman, and Sestir (2006) also used volume and duration settings, they standardized and summed the two parameters instead of multiplying them. Conversely, Sestir and Bartholow (2010) only analyzed average volume, not allowing their participants to set duration at all. Finally, Engelhardt, Bartholow, and Saults (2011) used a one-trial version of the CRTT because they wanted to eliminate reciprocal behavior and concerns over retribution. Unfortunately, they used the term “noise intensity” (p. 541) throughout their article without specifying whether this includes volume, duration, or both, and how this index was calculated.

In a two-phase version of the CRTT used by Bartholow and Anderson (2002), participants played two complete rounds with 25 trials each. During the first phase, only the opponent could set the duration and intensity of the noise blasts. During the second phase, roles were reversed so that participants could retaliate for the punishment they received in the first phase. The authors only considered average volume and the number of high volume settings because they did not find a significant effect on any of the duration measures. The reason they gave is that their participants supposedly ignored the duration option. Anderson et al. (2004) separated the volume settings of Phase 2 into four different scores: first trial and means of Trials 2-9, 10-17, and 18-25. They claimed this was a common procedure with the CRTT because early trials were supposedly more important since they were more likely to be used for retaliation. We do not wish to inappropriately speculate or judge why there was not always a rationale for modifications to the CRTT as there might be numerous reasons (e.g., spatial limitations in academic journals), but we believe that the lack of justifications for such alterations is a fact that warrants mentioning in a methodological critique of the test in question.

Sometimes the option of setting the volume and/or duration to zero as a way to act nonaggressively is provided. However, including zero values in mean calculations could potentially skew some of the aggression scores mentioned above. Including settings of zero as an option also raises further questions, for example, how to handle trials in which participants set only one of the two intensity parameters to zero.

From a methodological point of view, inconsistent procedures and analyses are highly problematic because they infringe upon the objectivity criterion of psychological diagnostic test theory. Simmons, Nelson, and Simonsohn (2011) pointed out that flexibility in data collection, analysis, and reporting in psychological research dramatically

increases actual rates of false-positive findings. Moreover, if there is no standardized procedure for a test and no standardized way to process the raw data into a meaningful value, the question remains whether the unstandardized score really approximates the true value of the construct. Aggression scores that are calculated with different procedural versions of the same test become very difficult to compare. Under the assumption that all these different procedures and analyses are equally capable of measuring the construct of aggressiveness, it is unclear why so many versions exist.

We believe that most of the time modifications to the CRTT have been made in good faith and for valid reasons, for example, to extend previous findings or to answer novel research questions (such as which opponent noise pattern elicits more aggressive responses). Without a doubt, theory-driven modifications of a method such as the CRTT, with the aim of answering specific research questions, can contribute to the understanding of psychological processes and extend the area in which a certain test can be applied. However, many authors do not explain in detail why they decided on a specific test procedure or on the aggression score they calculated from the raw data. In many cases, it is not clear why a particular score should be more suitable than others to address the respective research questions. Sometimes, the decision to focus on one of many possible scores seems to have been made post hoc, not prior to data collection. Our concern is not that changes to the structural aspects of the CRTT will render it useless but that it might tap into different constructs and that the results will have to be interpreted accordingly.

Reactivity

Participants can hear (and see) their opponents' settings and are likely to react to this in their own choice of volume and/or duration. While most researchers use preset randomized settings, they do not use the same pattern in all studies. Without a standardized noise pattern, participants of different studies will most likely receive different noise blasts by their opponent, which might cause variations in their settings.

Even with a standardized preset pattern of wins/losses and noise intensities, it is likely that the alleged opponent's responses have an effect on the behavior and/or motives (e.g., retaliation) of the participants. This can lead to a lack of item homogeneity, as it is possible that not all trials actually measure the same variable (i.e., the same behavior). Researchers using the CRTT have approached this problem with different rationales, e.g., by focusing on the first trial, as it is the only nonreciprocal data point in the test's one-phase version (Bushman & Baumeister, 1998) or because it is the first chance to retaliate in the two-phase version (Anderson et al., 2004). In the latter study, the authors used two different noise patterns (ambiguous and increasing) to show that, in fact, there is an effect on the participants' responses. However, the impact of trial outcome and intensity patterns is usually not controlled for, and to our knowledge, no one has ever systematically checked how much intra- and interindividual variance can be explained just by the opponent's volume and/or duration settings.

Research Questions

Until now, there has been no study that addresses the aforementioned objectivity issues by systematically comparing the different analysis procedures for the CRTT. To address this gap in the literature, we used data from three separate studies that were conducted independently of each other by the authors of this article, and all used the CRTT as their measure of aggression.

The three experiments were selected because they all studied digital games as a potential cause of aggressive behavior. Media violence is a research field in which the CRTT is particularly popular (Anderson et al., 2010). There is a fairly large body of research indicating that this stimulus does cause differentials in CRTT scores. In fact, all but two of the studies that we cited in the section on standardization investigated effects of violence in digital games on postplay aggression. In addition, the authors of this article were involved in these studies and thus had full access to the raw data. Each study was conducted separately and two different research groups were involved in the studies: The second author of this article was in charge of data collection and the original analysis of Study 1 (Mohseni, 2013), the first author did the same for Study 2 (Elson, Breuer, Van Looy, Kneer, & Quandt, 2013), and the third and fourth authors were responsible for Study 3 (Breuer, Scharkow, & Quandt, in press). Originally, the studies were conducted to investigate the effects of media violence and not to evaluate the CRTT.

The results presented in this article come from a secondary analysis of the data of these three studies. We want to stress that this article is not concerned with the original hypotheses and results of each study, and it was not our aim to replicate effects across studies. In this article, we are interested in the issues of standardization and reactivity associated with the CRTT. Hence, we do not want to compare effects across studies but investigate whether different analysis procedures lead to different results within each study.

More specifically, we wanted to answer the following research questions:

RQ1. Do measures of volume and duration converge?

RQ2. Do different methods to calculate CRTT aggression scores yield comparable results?

RQ3. Do the settings of the simulated opponent explain parts of the variance in the settings of the participants?

Study Descriptions

In this section, we briefly describe the methods, materials, samples, and procedures for each of the three studies, as well as their main findings. The following section will only describe the methods of each study with a special focus on the version of the CRTT that was used. Readers interested in the results of the original studies should refer to the online supplemental materials or the individual publications (Breuer et al., in press; Elson et al., 2013; Mohseni, 2013).

Study 1

This study (Mohseni, 2013) addressed the question whether situations of violent helping behavior in digital role-playing games increase aggressiveness and helpfulness outside the game.

Stimulus materials and measures. To answer this question, two independent variables, namely, in-game violence and in-game help, were created by modifying the role-playing game The Elder Scrolls IV: Oblivion (Bethesda Game Studios, 2006). Participants had to solve a quest in the game either by using violence or by stealthily traversing the map (violent vs. nonviolent). The quest was either framed as helping behavior (quest giver had to be saved) or as a treasure hunt. Participants were assigned to one of four conditions: “rescue” (helping with violence), “kill” (violence only), “help” (helping only), and “treasure hunt” (no helping and no violence). Aggressive behavior was measured with Bushman and Baumeister's (1998) version of the CRTT: The Competitive Reaction Task Reward & Punish (Version 3.2.5). Participants were able to set the volume and duration. Volume was calibrated using Bruel & Kjaers Sound Pressure Level Meter Type 2232, resulting in settings from 1 (55 dB) to 10 (95 dB). Participants could also choose a setting of 0 (0 dB). A win/loss pattern (12/13 trials) and the opponent's settings (increasing noise intensities over time) were predefined for all participants. Since Giancola and Zeichner (1995b) found that men are more likely to express aggression in intensity, while women are more likely to use duration, the measure for aggressive behavior in this study was the volume setting in the first round for male participants, and the duration setting in the first round for females.[1] The Aggressive Motives Scale (AMS; Anderson & Murphy, 2003; Anderson et al., 2004) is supposed to measure instrumental aggression (two items) and revengeful aggression (four items) during the CRTT and was used to assess the participants' motives. Helping behavior was assessed as the willingness to assist in another experiment (Greitemeyer & Osswald, 2010).

Procedure and sample. Before entering the lab, participants met a male confederate who was supposedly waiting for the experiment to start. The experimenter led the confederate to his cabin first and told him to wait for a second experimenter. After that, the participant was brought to a cabin and asked to sign an informed consent form. Participants were then taught how to play the CRTT, as recommended by Anderson et al. (2007, p. 65). Directly after that, participants started playing one version of Oblivion (see above). When the quest was solved, the experimenter returned and told the participant that his opponent was ready so that the CRTT may now begin. After the CRTT, participants filled out several questionnaires. At the end of the experiment, participants were thanked, debriefed, and those who did not participate for course credit received a monetary compensation of €10 (about $13).

Participants (N = 216; 188 males and 28 females)[2] were German bachelor and master students from different faculties. The largest proportion of participants majored in cognitive science (12.6%), psychology (7.5%), and law (5.6%). Mean age in the sample was M = 23.48 years (SD = 3.21). Due to language problems and a technical issue, data from two male participants had to be discarded from further analyses, leaving a total of 214 participants distributed over all four game conditions (53 in the “rescue” and “kill” conditions, 54 in the “help” and “treasure hunt” conditions).

Study 2

While Study 1 was more concerned with violence and helping as a means to reach a goal, in the second experiment (Elson et al., 2013) displays of violence and the game speed of a first-person shooter were manipulated to assess their effects on aggressive behavior.

Stimulus material and measures. Two features of the game Unreal Tournament 3 (Epic Games, 2007) were modified: displays of violence and game speed. Game speed was either set to the default value of 100% (normal speed) or to 140% (high speed). In the violent conditions, players wielded a grenade launcher and shooting opponents led to a gory death animation. In the nonviolent conditions, participants used a tennis-ball shooting Nerf gun. Aggressive behavior was measured with the standardized version of the CRTT with volume and

[1] It should be noted, however, that Giancola and Zeichner found this sex difference for electroshocks and not noise blasts. Whether these two stimuli can be equated as aggressive measures remains subject to further research.

[2] In the original study, data of the female subsample were not used for analysis. Therefore, results presented here may differ from results of the original study.

duration settings (Ferguson et al., 2008) using the experiment software Presentation (Neurobehavioral Systems, 2010). However, instead of using dB values, the range of volume was determined in a prestudy in which 15 participants were asked to select a volume level on the Windows sound settings that was “almost unbearable.” Taking that value as the maximum, the volume was scaled down linearly, resulting in settings from 1 to 10. The win/loss pattern (12/13 trials) and opponent settings were randomized for each participant. The mean of Volume X Duration over all trials was used as the measure for aggressive behavior.
Additionally, physiological arousal was measured by four different parameters. Two measures assessed autonomic responses: heart rate (HR) and skin conductance level (SCL). Two others were behavioral indicators: body movement and pressure exerted on mouse and keyboard.
Procedure and sample. Participants were assigned to one of four conditions (nonviolent vs. violent, normal- vs. high-speed). After entering the lab and signing the informed consent form, participants received a short briefing about the game's controls and objectives. After they finished three warm-up rounds, the experimenter started the main playing session that lasted 12 min. Afterward, participants were told that the second part of the experiment was about to begin, in which they would play 25 rounds of a reaction time game against a participant in another laboratory. Instructions were also presented on the computer screen before the first trial. After completing the CRTT, participants were thanked and debriefed. As an incentive, 40 computer games were raffled among the participants after completion of data collection.
Participants (N = 87; 60 males and 27 females) were undergraduate (79.4%) and graduate (4.6%) students from two large German universities and high school graduates (12.6%) who were about to enroll as university students. All participants were recruited via the online recruitment tool Cortex (Elson & Bente, 2009) and 3.4% did not indicate their current occupation. University students were recruited via mailing lists and came from different fields, the largest proportion being communication (20.7%) and psychology (13.8%). Mean age of the participants was M = 26.07 years (SD = 5.87). Due to technical difficulties, data from three participants had to be discarded from further analyses, leaving a total of 84, evenly distributed over all four game conditions (21 each).

Study 3

The third study (Breuer et al., in press) was originally conducted to test the frustration-aggression hypothesis (Berkowitz, 1989) in the domain of video games. Specifically, this study investigated effects of winning/losing and opponent behavior in a co-located multiplayer sports video game on negative emotions and aggression.
Stimulus material and measures. Participants played a match of the soccer video game 2010 FIFA World Cup South Africa (EA Canada, 2010) against a male confederate with substantial practice. The confederate was instructed to either win or lose against the participant and to either be friendly and helpful or to (mildly) trash-talk while playing. The trash-talking was not scripted, as it was meant to be adaptive. The confederate was provided with a list of sample phrases that he was free to combine and adapt according to the course of the game. Sample statements included ironical remarks such as “nice pass,” or snarky comments such as “that was easy.”
The version of the CRTT used in this study differed from those used in Studies 1 and 2 in several aspects. To provide the participants with an appropriate target for their aggression, participants in Study 3 were told that they would play the CRTT against the same person against whom they had played the soccer console game. Instead of an auditory cue, participants had to react to a visual cue displayed on-screen in order to address the issue of the noise blasts being a potential means to reduce the reaction time of the opponent (Adachi & Willoughby, 2011). If they won a trial, participants could choose the duration of blasts, ranging from 1-9 s, using the corresponding number keys on a keyboard. There were a total of 10 trials and the sequence of winning and losing trials (five each), as well as the opponent's settings, were randomized individually for each participant. The volume setting was excluded because Ferguson and Rueda (2009) reported insufficient correlations between duration and intensity in their CRTT validation study, and duration (in seconds) was believed to be a more intuitively comprehensible unit for all participants. A very unpleasant noise level was determined in a small prestudy with five participants. This volume level was held constant in all trials. The duration setting of the first trial was used as the measure for aggressive behavior. Afterward, the CRTT participants also completed a brief questionnaire that included items on negative emotions. Negative emotions were measured by four items from the German translation (Krohne, Egloff, Kohlmann, & Tausch, 1996) of the Positive Affect Negative Affect Schedule (Watson, Clark, & Tellegen, 1988).
Procedure and sample. Upon arrival at the laboratory, each participant gave their informed consent and was asked whether he or she had any experience playing 2010 FIFA World Cup South Africa on the Xbox 360 console. If the participant did not have any experience with the game, he or she played a practice session against an easy computer opponent for 5 min.
Following the practice phase, participants played a 2 X 5 min match against the confederate in one of the four conditions. After the match, the confederate was led into an adjacent room, and participants were told that they would play a reaction time game against him. At the end of the experiment, participants were thanked and those who did not participate for course credit received a monetary compensation of €10 (about $13). All participants

were debriefed via e-mail at the end of the data collection phase to avoid an early uncovering of the role of the confederate and the purpose of the study.
A total of 91 participants signed up for the experiment using the Cortex online recruiting tool (Elson & Bente, 2009). The majority of the sample (80.3%) was bachelor and master students, and most of them were enrolled for communication (49.9%). The data of 15 participants were excluded from further analysis because of language problems, participants having suspicions about the purpose of the study, knowing the confederate, or the game resulting in a draw. The mean age of the remaining 76 participants (48 female; 28 male) was M = 22.60 years (SD = 3.20).

Secondary Analysis of the CRTT Data

In order to answer the first research question (RQ1), asking if volume and duration measures converge, we calculated the correlation between these two variables for Studies 1 and 2. In Study 1, volume and duration measures for each trial showed a medium-sized significant correlation ($r = .49$), p (one-tailed) $< .001$. The correlation of average volume and duration measures for each participant was substantially higher ($r = .82$), p (one-tailed) $< .001$. Study 2 showed a similar medium-sized significant correlation for each trial ($r = .46$), p (one-tailed) $< .001$, while the correlation of average volume and duration measures for each participant was higher ($r = .76$), p (one-tailed) $< .001$.

Convergence of CRTT Scores

Although the correlations per participant were quite high and show that volume and duration measures are related, they might not measure the same construct. The results of our comparisons between different analysis procedures support this notion, as we show in the remainder of this section. In our comparisons, we included most of the analysis procedures for the one-phase CRTT that have been published previously and could be applied to our data sets.
Naturally, it would have been possible to add even more aggression scores to our analyses or to use them in more advanced statistical analyses, such as multilevel models or latent growth models. However, we wanted to test whether the available scoring techniques are comparable and whether they lead to different results in the most basic and, in this case, the most common statistical analyses. Adding yet another method for scoring would only further complicate any attempt at standardization. For our comparisons, we only used one independent variable per study. For Studies 1 and 2, this is violent content; for Study 3, we focused on the outcome of the game (winning vs. losing). These variables were selected because they have been previously identified as causes of aggressive behavior (as reflected by CRTT scores). As stated before, the aim of this analysis is not to examine the convergence of the effects of the independent variables on the CRTT across studies. Instead, we are interested in the variability of results when using different CRTT scores within each study and whether this variability can be replicated in studies that differ in their design.
Study 1. For Study 1, none of the aggression scores calculated were significantly different between the experimental conditions (see Table 1). In-game violence only had a small, but nonsignificant effect on the volume chosen in the first trial. However, it should be noted that there was a high variability in the indices' significance levels (from $p = .070$ to $.934$).
Study 2. The comparisons for Study 2 show a more ambiguous pattern (see Table 2). With the exception of average volume after losing, all measures for aggression based on average intensity (volume and duration) yielded nonsignificant differences between the experimental conditions with p-values ranging from .045 to .668. The range of p-values between all indices (including those not based on averages) was between <.001 and .959. The effect on volume was mostly larger than on duration, indicating that volume and duration are at least not equal in terms of their sensitivity to the effects of stimuli. This difference could not be observed for the first trial settings, however. Effect sizes ranged from .0 to .39, the largest being the number of high volume blasts, $F(1, 80) = 16.01$, $p < .001$, $\omega = .39$, showing that participants who played a violent game used volume settings from 8 to 10 more frequently than those that played a nonviolent game. This suggests that the number of extreme volume scores does not tap into the same construct as mean-based scores.

Table 1

Study 1: Effects of In-Game Violence on Different CRTT Aggression Scores

Aggression score	F(1, 210)	p	ω
Mean volume	1.23	.269	.03
Mean volume after wins	1.10	.295	.02
Mean volume after losses	1.06	.303	.02
Mean duration	0.23	.633	.0
Mean duration after wins	0.01	.934	.0
Mean duration after losses	0.59	.443	.0
Mean volume X duration	0.63	.429	.0
Mean volume X$\sqrt{duration}$	0.83	.364	.0
Mean volume X $\log_e$(duration)	2.31	.130	.08
Total high volume settings	0.13	.719	.0
Total high duration settings	0.1	.730	.0
First trial volume	3.31	.070	.10
First trial duration	0.26	.609	.0

Note. CRTT = Competitive Reaction Time Task.

Table 2

Study 2: Effects of Displayed Violence on Different CRTT Aggression Scores

Aggression score	F(1, 80)	p	ω
Mean volume	3.28	.074	.16
Mean volume after wins	1.46	.230	.07
Mean volume after losses	4.14	.045	.19
Mean duration	0.95	.334	.0
Mean duration after wins	0.19	.668	.0
Mean duration after losses	1.74	.191	.09
Mean volume X duration	2.78	.099	.14
Mean volume X$\sqrt{duration}$	2.77	.100	.14
Mean volume X $\log_e$(duration)	2.17	.144	.12
Sum high volume settings	16.01	<.001	.39
Sum high duration settings	0.17	.685	.0
First trial volume	0.08	.779	.0
First trial duration	0.01	.959	.0

Note. CRTT = Competitive Reaction Time Task.

Mathematically, it seems odd that the significantly larger number of high volume settings did not cause the mean volume to be significantly higher. To explain this puzzling finding, we tested whether there were also differences in the number of low volume
settings (range 1-3). A two-factorial analysis of variance (ANOVA) revealed a significant effect of violent game playing on the total of low volume settings, F(1, 80) = 10.57, p = .002, ω = .32, indicating that playing a violent game also led to the more frequent use of volume settings from 1 to 3. Apparently, high volume and low volume settings canceled themselves out, resulting in nonsignificant mean differences.

Study 3. Since participants were only able to set duration in Study 3, we could use only six out of 13 analysis procedures (see Table 3). While there were small effects of losing on mean duration and mean duration after wins, they were both nonsignificant. The only significant effect of game outcome was on duration settings in the first trial (ω = .20,p = .043). Participants who lost against the confederate on average chose longer durations. The remaining aggression scores yielded small effects ranging from ω = .09 to .15, with p-values between .096 and .212.

Convergence of the different scores. Looking at the comparisons of analysis procedures for each study, the answer to the second research question (RQ2), asking whether different methods to calculate CRTT scores produce comparable results, has to be negative. Effect sizes ranged from ω = .0 to .10 in Study 1, .0 to .39 in Study 2, and .09 to .20 in Study 3. As effect sizes are not the same for all analysis methods, it seems that the calculation of different aggression scores can lead to results that are substantially different from each other; in one case, even diametrically opposed. For example, depending on which aggression score is calculated (and reported) with the data from Study 2, results could provide evidence that playing a violent digital game increases postplay aggressive behavior (number of high volume settings), decreases it (number of low volume settings), or has no effect at all (most other indicators). We also found inconsistencies in Study 3, with one variant yielding a

significant effect, while five others did not. Note that Study 1, having the largest N by far, yielded the most consistent effect sizes. This could indicate an issue with precision in the CRTT, meaning that the test is prone to producing false-positive significant differences when conducted with small to medium samples as is common practice in experimental psychology (see Simmons et al., 2011). However, even in Study 1 there was a large variability in p-values from .070 to .934.

While none of the scores in Study 1 reached the criterion of significance, using the methodological flexibility procedures discussed by Simmons et al. (2011), such as adding or removing a small number of participants or controlling for simple covariates (e.g., age), some of the findings could easily become significant. Not only are those practices relatively common in psychological research (John, Loewenstein, & Prelec, 2012), statistically significant findings also dramatically increase the likelihood of publication, resulting in a peculiar prevalence of p-values just below .05 (Masicampo & Lalande, 2012), particularly in controversial fields with major social implications, such as media violence (Ioannidis, 2005). Of course, we do not wish to imply that aggression research (or other psychological research) is invalid on these grounds, but want to express our concern that the unstandardized use of the CRTT might increase the likelihood of finding (and publishing) significant effects when, in fact, there are none.

Reactivity

To answer the third research question (RQ3), asking whether the settings of the simulated opponent can explain parts of the variance in the settings chosen by the participants, we regressed the participants' volume and duration settings on the results of the previous trial (win/loss, volume and duration of the noise blast received). As win/loss pattern and opponent's intensity settings were identical for all participants in Study 1, we analyzed the intraindividual variation of all respondents across 25 trials for this study by regressing the participants' settings in every trial on the settings of the opponent in the previous trial. On average, the explained variance in the respondents' settings was R^2 = .39 (SD = .25) for volume and R^2 = .32 (SD = .21) for duration. Effectively, this means that about one third of all variance in the CRTT data could be explained solely by the opponent's settings in the previous round. Compared to Studies 1 and 3, the amount of variance explained just by trial outcome and the opponent's settings of the previous round was a lot smaller in Study 2: R^2 = .13 (SD = .09) for participant's volume settings and R^2 = .16 (SD = .12) for duration settings. In Study 3, both outcome and noise duration were randomized in every trial individually for each participant. As the settings of the opponent were only displayed after loss trials, we could only estimate the amount of variance explained by the opponent's settings for these trials. The results were very similar to those of Study 1, with an average of R^2 = .41 (SD = .32) of explained variance in the participants' duration settings (after loss trials).

Table 3

Study 3: Effects of Game Outcome on Different CRTT Aggression Scores

Aggression score	F(1, 72)	p	w
Mean duration	2.84	.096	.15
Mean $\log_e$(duration)	2.79	.099	.15
Mean duration after wins	2.84	.096	.15
Mean duration after losses	1.63	.205	.09
Sum high duration settings	1.58	.212	.09
First trial duration	4.24	.043	.20

To further investigate whether there would be differences in the participants' reactivity depending on the outcome of a previous round, we conducted separate regression analyses for trials after rounds won versus rounds lost (see Table 4) for Studies 1 and 2. Study 3 was excluded from this comparison as the opponent settings were not displayed when a participant won a round. As in our previous analyses, there were large differences in the reactivity between Studies 1 and 2. More important, there was an influence of opponent settings independent of the outcome of the previous round, with slightly inconsistent R^2s ranging from .34 to .47 for Study 1, and .16 to .22 for Study 2.

Of course, the reactivity problem might even be more complex, as these analyses only accounted for the settings of the previous round on the next one. It is certainly possible that the opponent's settings also produced changes in reactivity in subsequent rounds.

Motives

The issue of reactivity also opens up additional questions about what influences the participants' settings in the CRTT. As mentioned above, other studies found differences in responses depending on the characteristics of the noise pattern. However, there have only been a few attempts to identify what causes those reactions, such as

personality traits or emotional states. We believe that motivations to aggress also play a major role here, as these motivations could be influenced by violent media content, personality traits, and the behavior of the opponent. Participants in the first study were asked about their motivations during the CRTT. We used these data to determine the amount of variance in the participants' settings explained by different motives. While the AMS is supposed to measure the motives of instrumental and revengeful aggression, the phrasing of the items is open to alternative interpretations. For instance, the first item of the instrumental aggression scale, “I wanted to impair my opponent's performance in order to win more,” could measure a nonaggressive competition motive, while the second item, “I wanted to control my opponent's level of responses,” could tap into the motive to control or manipulate the opponent. Within the revengeful aggression scale, the two items, “I wanted to pay back my opponent for the noise levels he or she set” and “I wanted to blast him or her harder than he or she blasted me,” do seem to measure a motive for revenge, but the other two, “I wanted to make my opponent mad” and “I wanted to hurt my opponent,” could measure a form of aggression that has nothing to do with revenge. Additionally, a scale intercorrelation of up to r = .49 (Anderson et al., 2004) may indicate that the AMS only measures one single construct.

Table 4

Effects of Opponent's Settings on Participant's Settings by Outcome of the Previous Round

		R^2			
		Study 1		Study 2	
Participant settings	Previous round outcome	M	SD	M	SD
Volume	Win	.39	.25	.21	.17
	Lose	.47	.27	.16	.13
Duration	Win	.34	.22	.22	.18
	Lose	.39	.24	.18	.16

As the factorial structure by Anderson and Murphy (2003) could not be replicated (since the two items of the instrumental aggression subscale correlated negatively), all items were subjected to a principal component analysis (PCA) with orthogonal rotation,[3] which revealed two different factors. The first consisted of three items (impair performance, make mad, hurt), and the second consisted of two (control response, pay back), leaving one item that loaded on both factors (blast harder). The first factor seems to reflect the motive of aggression and competition, while the second seems to reflect retaliation and control. Accordingly, these factors cannot be interpreted as individual motives. They are still useful for predicting aggression scores though, especially since using all items together in a single linear regression model would very likely lead to issues of multicollinearity. Theoretically, the motives could both moderate and mediate the effect of any independent variable on the CRTT scores. As the wording of the AMS items clearly shows, some or even most of the motives have a trait component. There are people who are generally more likely to (re)act aggressively or to retaliate. The overlap with influential personality traits, such as trait aggression, would speak for the motives being moderators. At the same time, the motivations also have state components as they can be influenced by both the treatment (e.g., retaliation in the case of a rude opponent in Study 3) and the CRTT itself (settings chosen by the opponent; see section on reactivity). For example, aggressive motives should be more prominent in aggressive conditions and result in higher CRTT scores, while nonaggressive motives should be more prominent in nonaggressive conditions and result in lower CRTT scores.

Regressing the motives on “mean volume X duration” in Study 1, both factors led to higher mean settings, explaining 47% of the variance (see Table 5).

The first factor also predicted higher settings in the first trial, whereas the second was not significant here. In this case, both factors explained 26% of the variance (see Table 6). Although there was no interaction between the effects of experimental conditions and motives, this does not necessarily mean that the motives were not influenced by any of the independent variables. As our data show, these motives can have a large impact on CRTT scores. However, further research is needed to investigate systematically whether the AMS really measures motives (as they are assessed post hoc), whether other motives could play a role (e.g., conformity or social desirability), and in which way other variables (e.g., personality traits, or experimental stimuli) possibly influence these motives.

[3] Conducting a principal axis factoring instead of a PCA and/or using oblique instead of orthogonal rotation, led to the exact same factor structure (with some items loading lower on their corresponding factor in cases of the PAF, and with small interfactor correlations of $r < .12$ in cases of oblique rotation).

Discussion

Our results show that modifications to the CRTT test procedure result in different assessments of behavior. For example, whether zero is included as an option for participants to set their intensity and whether the pattern of opponent settings is increasing or ambiguous, might lead to very different findings, even if the studies are otherwise similar. While it might seem trivial that different versions of a test yield different results, this is a major problem in the case of the CRTT as it is being used widely, and as there is still no thoroughly validated standardized test procedure.

Methodologically even more problematic, however, are the differences among analyses of the raw data collected with the CRTT. The choice of a calculation method for aggression scores can severely influence the significance, size, or even direction of an effect. Our data could support different hypotheses about the effects of our stimuli, by either deliberately omitting some of the results, or just by picking one aggression score at random.

Furthermore, our findings suggest that volume and duration do not measure the same construct, although they clearly seem to be related. This does not necessarily constitute a problem with the CRTT. In fact, we would consider it a benefit if the CRTT captured different (sub) dimensions of aggressive behavior. However, no attempts to systematically disambiguate the different latent variables supposedly measured by volume and duration have been made thus far. Using the electroshock variant of the CRTT, Verona, Reed, Curtin, and Pole (2007) defined shock intensity as overt and shock duration as covert aggression and found differential use of the two settings between men and women (see also Giancola & Zeichner, 1995b; Ritter & Eslea, 2005). Whether these initial results also apply to the noise blast version of the CRTT, however, has yet to be tested. Thus, sex could be a further variable to be considered when using the CRTT.[4] The main concern of the present study is to compare different CRTT scores within the studies. The proportion of men and women in each study could explain the differences in effect sizes across studies, but not the variability within each one. The results presented in this article support the question raised by Ferguson et al. (2008), asking whether the CRTT should be used in aggression research at all. In order to find a definitive answer to this question, however, further studies evaluating the objectivity, reliability, and validity of the CRTT are necessary.

The practical implications of our results are that one has to consider and interpret findings of studies in which the CRTT is used with extreme caution, particularly when detailed reasons for any modification are not provided. We do not want to suggest that the consumption of drugs, alcohol, and violent media does not induce or increase aggressiveness. In fact, we believe that any research looking into causes of human aggression is of high relevance, as long as the results are reliable. The problems associated with the use of the CRTT, however, can seriously diminish the credibility and significance of any laboratory research on aggression.

Table 5

Study 1: Effects of Motives on Mean Volume X Duration

Factor	B	SE B	β	p
1	8.34	0.639	.65	<.001
2	2.66	0.639	.21	<.001

Note. R^2 = .47 ($p < .001$), $R^2_{adjusted}$ = .47, variance inflation factor = 1.00.

Table 6

Study 1: Effects of Motives on First Trial Volume X Duration

Factor	B	SE B	β	p
1	5.26	0.616	.51	<.001
2	-0.56	0.616	-.05	.366

Note. R^2 = .26 ($p < .001$), $R^2_{adjusted}$ = .25, variance inflation factor = 1.00.

Of course, our research has its own limitations that should be acknowledged. First, only data sets from three studies were available to us, while the number of experiments using the CRTT is much larger. Replicating our comparisons with more and possibly even larger data sets could help to identify a potential bias in the different ways used to process raw CRTT data.[5]

Another limitation is that all three studies used digital games as stimuli to elicit aggression. The consolidation of our findings would require supplemental analyses of data sets from studies that investigate other possible sources of aggression. Ironically, the three studies presented in this article all employed different versions of the CRTT. For Study 1, Version 3.25 of the Competitive Reaction Task Reward & Punish by Bushman and Baumeister

[4] A systematic examination of sex differences could not be made with the three data sets available to us, as the ratios across sex and the study designs differed too much between the studies.

[5] Authors willing to share their CRTT raw data with us for further analysis are welcome to contact the first author of this article.

(1998) was used. Study 2 used the standardized version suggested by Ferguson et al. (2008). Study 3 made modifications to Ferguson's version due to the nature of the research questions and hypotheses, although one might argue that the exclusion of volume was not directly justified by the research interest. The choice to only use duration was mainly made for practical reasons, such as the problem of proper calibration (see section on calibration and scaling below). All three studies also originally used different scores as the measure for aggressive behavior (first trial volume and duration, mean volume and duration). An ideal comparison among the studies would make sure that they all used the same (standardized) version of the CRTT. However, the differences in this set of studies also reflects the current practice of research using the CRTT and suggests that differences in CRTT procedure (e.g., noise pattern) might cause differences in the results.

Despite various problems with the CRTT and its analysis, there are currently no suitable alternatives for a behavioral measure of aggression that can be used in laboratory research with adult participants. Some scholars consider the Hot Sauce Paradigm (HSP; Lieberman, Solomon, Greenberg, & McGregor, 1999) to be a better measure (Adachi & Willoughby, 2011). However, a systematic validation of this test has only begun quite recently (e.g., Beier & Kutzner, 2012).

If we do not want to abandon laboratory research on aggressive behavior altogether, we have to find ways to deal with the problems of available tests. The following two sections are meant to provide some recommendations on how to use the CRTT in future studies and offer suggestions on how it might be improved upon or complemented by other methods. In the first section, we present recommendations for the future use of the CRTT based on the findings of the present study, while the second part is concerned with general advice from previous studies and our own considerations.

Recommendations Based on Our Findings

Based on our findings, we want to provide guidelines for researchers studying aggressive behavior in the laboratory. Following the structure of the article, we offer suggestions on how to address issues of standardization and how to control reactivity problems.

Standardization

At this point it is impossible to say which CRTT variant is the “right one.” We recommend defaulting to the standardized version of the CRTT suggested by Ferguson et al. (2008), simply because it is relatively close to the original CRTT by Taylor (1967), without having the ethical problem of using electroshocks. Of course, without proper validation, any variant is technically as good as the next one. Nonetheless, we still think the version suggested by Ferguson et al. (2008) provides a good starting point, as its authors were the first who explicitly called for standardization and suggested a “standard operating procedure.” If there are reasons for modifications, researchers should provide these and explain the benefits of their changes in detail. In keeping with the suggestions of Simmons et al. (2011), authors should ideally also explain if and how the results of their studies could differ using the standardized version of the CRTT.

With regard to the analysis, there is no definitive answer to the question of how to calculate aggression scores, or whether different scores might measure different types of aggression, as long as none of them have been properly validated. As it seems that volume and duration do not measure the exact same construct, it is advisable to consider them as separate measures for related subdimensions of aggression. If researchers are interested in measuring unprovoked aggression, they should also look at the settings in the first trial. Those studying provoked aggression or retaliation, on the other hand, should focus on all trials except the first one.

We are aware that modifications to established methods (e.g., the Stroop test) for new research questions are not uncommon in psychological science. However, we see two major differences compared to the use of the CRTT: First, the modifications being made are rarely theory-driven. Often, new variants seem to include arbitrary changes that are neither explained nor justified on the grounds of new research questions. Second, many of the variations of the CRTT are generalized to real-world acts of aggression, although there are no data supporting the external validity for any one of them. We believe that before a laboratory procedure is modified to answer particular research questions, there should be compelling evidence for sound psychometric properties of a “base version” of that procedure from which the variations can be derived. This is particularly true when researching a clinical topic that is highly relevant for society, such as aggression.

Reactivity

Should researchers decide to use all trials, they should analyze and report how much of the variance in the participants' settings can be explained by the settings of the opponent in the previous round. We do not necessarily consider reactivity a major problem, as it confirms that, in principle, the CRTT measures some kind of reciprocal behavior by the participant. One should not forget, however, that the opponent noise pattern is pregenerated and does not adapt to participants' settings, meaning that it is arguably not reciprocal and, hence, artificial and not very human-like. The reactivity might also become an issue when investigating the effects of

antecedent stimuli, as they could potentially interact with reactivity effects, or even be superimposed by them. As this is the nature of the CRTT (with the exception of the first trial), we strongly recommend reporting the magnitude of the reactivity in a similar fashion as in this article.
To understand why participants chose their settings, the CRTT should be followed by an assessment of the participants' motivations, such as the AMS (Anderson & Murphy, 2003), although it remains unclear if this measurement of intentions is influenced by the CRTT itself. The AMS should be extended to assess more motives, such as conformity, deescalation, or schadenfreude, as these could also play a role. Studies on provoked/motivated aggression (e.g., the frustration-aggression hypothesis) should provide participants with an appropriate target for their aggression (e.g., another participant or confederate who provokes them). Using identical patterns of wins/losses and opponent settings for all participants is another important measure to address standardization and reactivity issues. We would even go one step further and endorse a standardized pattern for all studies using the CRTT (unless, of course, the pattern is being manipulated as an independent variable), as we consider this a part of the test material that should not introduce an uncontrolled source of variance.

General Recommendations

Nonaggressive Options

If zero settings as a nonaggressive option for participants are included, considering how to properly treat them in the statistical analyses is paramount. However, we remain skeptical whether zero settings can actually be included without skewing mean-based aggression scores in further analyses, and therefore recommend using settings from 1 to 10. Instead of zero settings, following the suggestions of Tedeschi and Felson (1994), we advocate including other nonaggressive alternatives, such as options to interact with the alleged opponent or to revoke any exposure to a potentially aggressive situation (i.e., the option to avoid or abort the CRTT).

Calibration and Scaling

Another general issue of the CRTT's standardization concerns is its calibration and scaling. Typically, noise intensity can be set on a scale from 1 to 10 in steps of 5 dB, ranging from 50 to 95 dB, but sometimes up to 115 dB. Some versions of the test also offer the additional option of setting the volume to zero (0 dB). The decibel scale, however, is a questionable choice for scaling noise intensity for two reasons: first, standardizing dB levels in all types of headphones is highly susceptible to interference due to variances in the sound card producing, the headphone speaker playing, and the volume meter recording the noise. Second, it appears that most scholars using the CRTT consider the available intensity settings from 0 to 10 to reflect linear intervals in the volume and its unpleasantness for humans. However, decibel is a logarithmic unit. For example, the increase in discomfort from 50 to 55 dB is substantially smaller than the one from 90 to 95 dB. Moreover, the “maximum unpleasantness” varies greatly between studies, from 95 dB (a subway train at a distance of 60 m) to 115 dB (a rock concert); the latter actually being four times louder than the former. Without any preceding transformation of the data into a linear scale, the assumption of equidistance between levels of noise and unpleasantness can lead to a misestimation of resulting aggression scores.
Ideally, volume should be calibrated using high quality level meters (e.g., Bruel & Kjaer Type 2232). For calibration, sone should be preferred over decibels, as it is a better approximation of perceived loudness. The typical range of 55-95 dB roughly equals 5-105 sone. If it is not possible to calibrate in sone, decibels can be transformed to approximate perceived loudness as follows: $x = 2^{\Delta L/10}$, where x is the difference in loudness and ΔL is the difference in volume. For example, compared to 50 dB, 55 dB ($\Delta L = 5$) is perceived to be 1.4 times, 60 dB ($\Delta L = 10$) 2.0 times, and 95 dB ($\Delta L = 45$) 22.6 times louder. Calibrating headphones is tricky, as there exists no standardized calibration distance. Normally, decibels are measured with a distance of 1 m (~3.3 feet) from the source of the sound. However, this is of little practical use in the case of headphones. We therefore suggest putting the volume meter as close to the headphone as the human ear would be.

Validity

While some authors found proof for the convergent and discriminant validity of the CRTT (Anderson & Bushman, 1997; Carnagey & Anderson, 2005; Giancola & Chermack, 1998), others did not (Ferguson, 2007; Ferguson & Rueda, 2009; Ferguson et al., 2008; Tedeschi & Quigley, 1996, 2000). This dissent is not so much based on different findings as it is on different convictions about which findings, indicators, or forms of validation are suitable to prove validity. For instance, some authors (e.g., Giancola & Chermack, 1998) believe that because aggressive persons score higher in the CRTT than nonaggressive persons, this would indicate the external validity of the test. According to Tedeschi and Quigley (2000), the reasoning behind this form of validation is that “If aggressive people produce behavior A more often than nonaggressive people outside the laboratory, and aggressive people produce behavior B more often than nonaggressive people inside the laboratory, then A = B”

(p. 133). Specifically, if aggressive people behave violently more often than nonaggressive people in the "real world," and aggressive people give higher noise blasts than nonaggressive people when using the CRTT in the lab, then the noise blast intensity should be an indicator of violent behaviors. However, according to Tedeschi and Quigley, this conclusion constitutes a logical fallacy, which they illustrated with another example: High temperatures cause people to drink more fluids outside a laboratory and to rate others more negatively inside a laboratory. However, this does not allow the conclusion that drinking more fluids is the same kind of behavior as giving negative ratings.

In their extensive literature review of clinical and research aggression instruments, Suris et al. (2004) were unable to find a construct validation study for the CRTT. A recent lab-field comparison from 82 meta-analyses by Mitchell (2012) raises further concerns about the external validity of laboratory aggression research in comparison to other areas of psychological research, particularly industrial-organizational psychology.[6]

In line with Tedeschi and Felson (1994), we are convinced that an intent to harm (physically or otherwise) is a necessary and indispensable precondition for every form of aggressive behavior. Therefore, the main problem of all aggression measures is how to ensure that participants really have the intent to harm. Within the CRTT, plausible nonaggressive explanations for administering strong noise blasts could, e.g., be (a) falling for the cover story or (b) the desire to satisfy the experimenter and, thus, the demand characteristics of the test, especially if there is no nonaggressive option (Ritter & Eslea, 2005). Motive(s) determining the strategy in the CRTT procedure can also lead to very different reactions to the opponent's noise pattern. The underlying strategy or motive essentially determines the response pattern in laboratory measures of aggression. We believe it is paramount to gather evidence for the CRTT's construct validity in order to satisfy the definition of aggression and to ensure that participants actually have the intent to harm their alleged opponents. Asking participants about their intentions right after the CRTT with the AMS (Anderson & Murphy, 2003) has its limitations due to their retrospective nature and possible biases, such as social desirability. Therefore, validation studies in which participants' intentions are uncovered outside of an experimental situation should be given preference. This could be achieved by a study similar to Beier and Kutzner's (2012) validation of the Hot Sauce Paradigm (Lieberman et al., 1999) and with the inclusion of (additional) qualitative methods, such as interviews.

If the CRTT is used in experimental studies that are not (exclusively) designed to evaluate the validity of the test, researchers should include additional (ideally validated) measures of aggressiveness to investigate convergent and discriminant validity. To ensure that differences between experimental groups observed with the CRTT can be interpreted accurately, measurement of aggressive personality traits and pre-post designs can be viable solutions (Valadez & Ferguson, 2012), although pre-post designs could again introduce the problem of sensitization and contamination of the second measurement.

A further validation instrument could be physiological responses to the CRTT. In many studies that investigate causes of aggression, arousal indicators (most often galvanic skin response and heart rate) are frequently recorded during the stimulus exposure but not during the CRTT. Physiological responses could contribute to the test's validity, as more aggressive participants should show increased arousal (Zillmann, 1983). This would also help to uncover which events (losing a round, receiving or setting noise blasts) in the CRTT are particularly agitating, although naturally, physiological measurements do not allow drawing conclusions about an intention to hurt.

To summarize the order in which our recommendations should be considered, we created a flowchart (see Figure 1).

Conclusions

Our study provides empirical evidence that the CRTT suffers from several objectivity issues, in particular a lack of standardization. This can lead to inaccurate statements about the effects of various stimuli used in aggression research, such as violent media or alcohol. However, this article only addresses problems associated with the objectivity of the test. Despite the impact that this has on the significance of laboratory research on aggression, we believe that a satisfying solution to the problem of objectivity is feasible. Based on our own findings and previous methodological work on the CRTT, we can draw three main conclusions:

1. The results of studies that use the CRTT and meta-analyses that include these have to be interpreted with great caution. Moreover, given the questionable external validity of the test, researchers should be careful when they generalize results to situations outside the lab or make inferences about potential long-term effects to the point of public health issues.

[6] Although it should be mentioned that there are domains (e.g., gender-focused comparisons) in which matters seem even worse.

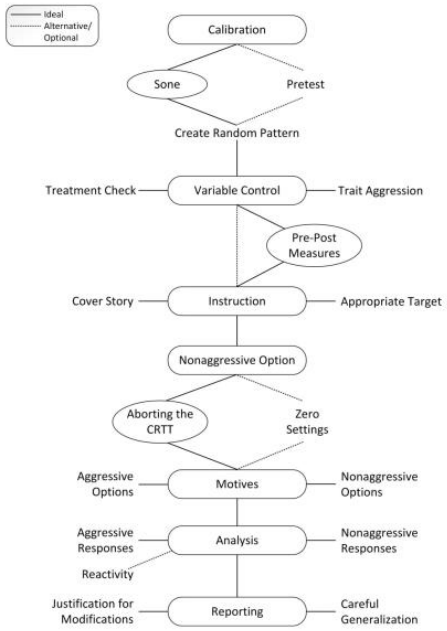

Figure 1. Recommendation flowchart for using the Competitive Reaction Time Task (CRTT).

2. Scholars who still want to use the CRTT to measure aggression must agree on a standardized version of the CRTT (as proposed by Ferguson et al., 2008). All modifications to this standardized version should be made explicitly and justified properly in the corresponding research reports.

3. If a standardized version is established and all issues of objectivity are addressed, the next step is to use this version for thorough validation studies that also investigate the construct validity of the test.

References

Adachi, P. J. C., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? Psychology of Violence, 1, 259-274. doi:10.1037/a0024908

Anderson, C. A., & Bushman, B. J. (1997). External validity of “trivial” experiments: The case of laboratory aggression. Review of General Psychology, 1, 19-41. doi:10.1037/1089-2680.1.1.19

Anderson, C. A., & Carnagey, N. L. (2009). Causal effects of violent sports video games on aggression: Is it competitiveness or violent content? Journal of Experimental Social Psychology, 45, 731-739. doi:10.1016/j.jesp.2009.04.019

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 (Ed.), Advances in experimental social psychology (Vol. 36, pp. 199-249). New York, NY: Elsevier. doi:10.1037/e552522012-008

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. Journal of Personality and Social Psychology, 78, 772-790. doi:10.1037/0022-3514.78.4.772

Anderson, C. A., Gentile, D. A., & Buckley, K. E. (2007). Violent video game effects on children and adolescents: Theory, research, and public policy. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195309836.001.0001

Anderson, C. A., & Murphy, C. R. (2003). Violent video games and aggressive behavior in young women. Aggressive Behavior, 29, 423-429. doi:10.1002/ab.10042

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A.,...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, 151173. doi:10.1037/a0018251

Baron, R. A., & Richardson, D. R. (1994). Human aggression (2nd ed.). New York, NY: Plenum Press.

Bartholow, B. D., & Anderson, C. A. (2002). Effects of violent video games on aggressive behavior: Potential sex differences. Journal of Experimental Social Psychology, 38, 283-290. doi:10.1006/jesp.2001.1502

Bartholow, B. D., Bushman, B. J., & Sestir, M. A. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. Journal of Experimental Social Psychology, 42, 532-539. doi:10.1016/j.jesp.2005.08.006

Bartholow, B. D., Sestir, M. A., & Davis, E. B. (2005). Correlates and consequences of exposure to video game violence: Hostile personality, empathy, and aggressive behavior. Personality and Social Psychology Bulletin, 31, 1573-1586. doi:10.1177/0146167205277205

Beier, R., & Kutzner, F. (2012). Choose a juice! The effect of choice options and intention on aggression in a modified hot-sauce paradigm. Paper presented at the 13th annual meeting of Society for Personality and Social Psychology, San Diego, CA.

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

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. Psychological Bulletin, 106, 59-73. doi:10.1037/0033-2909.106.1.59

Bethesda Game Studios. (2006). The Elder Scrolls IV: Oblivion [Computer game]. Novato, CA: 2K Games.

Bond, A., Curran, H. V., Bruce, M. S., O'Sullivan, G., & Shine, P. (1995). Behavioural aggression in panic disorder after 8 weeks' treatment with alprazolam. Journal of Affective Disorders, 35, 117-123. doi:10.1016/ 0165-0327(95)00053-4

Bond, A., & Lader, M. (1986). The relationship between induced behavioural aggression and mood after the consumption of two doses of alcohol. British Journal of Addiction, 81, 65-75. doi:10.1111/j.1360-0443.1986.tb00296.x

Bond, A., & Lader, M. (1988). Differential effects of oxazepam and lorazepam on aggressive responding. Psychopharmacology, 95, 369-373. doi:10.1007/BF00181949

Breuer, J., Scharkow, M., & Quandt, T. (in press). Sore losers? A reexamination of the frustration-aggression hypothesis for colocated video game play. Psychology of Popular Media Culture.

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

Bushman, B. J., & Anderson, C. A. (2001). Media violence and the American public: Scientific facts versus media misinformation. American Psychologist, 56, 477-489. doi:10.1037/0003-066X.56.6-7.477

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, 219-229. doi:10.1037/0022-3514.75.1.219

Bushman, B. J., & Gibson, B. (2011). Violent video games cause an increase in aggression long after the game has been turned off. Social Psychological and Personality Science, 2, 29-32. doi:10.1177/1948550610379506

Bushman, B. J., Ridge, R. D., Das, E., Busath, G. L., & Key, C. W. (2007). When God sanctions killing: Effect of scriptural violence on aggression. Psychological Science, 18, 204-207. doi:10.1111/j.1467-9280.2007 .01873.x

Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. Psychological Science, 16, 882-889. doi:10.1111/j.1467-9280 .2005.01632.x

EA Canada. (2010). 2010 FIFA World Cup South Africa [Video game]. Redwood City, CA: EA Sports.

Elson, M., & Bente, G. (2009). CORTEX: Computer-aided registration tool for experiments. Cologne, Germany: University of Cologne. Retrieved from http://cortex.uni-koeln.de

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (in press). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. Psychology of Popular Media Culture. Advance online publication. doi:10.1037/ppm0000010

Engelhardt, C. R., Bartholow, B. D., & Saults, J. S. (2011). Violent and nonviolent video games differentially affect physical aggression for individuals high vs. low in dispositional anger. Aggressive Behavior, 37, 539-546. doi:10.1002/ab.20411

Epic Games. (2007). Unreal Tournament 3 [Computer game]. Chicago, IL: Midway Games.

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. Aggression and Violent Behavior, 12, 470-482. doi:10.1016/j.avb.2007.01.001

Ferguson, C. J. (2011). The wild west of assessment. Measuring aggression and violence in video games. In L. Annetta & S. C. Bronack (Eds.), Serious educational game assessment (pp. 43-56). Rotterdam, the Netherlands: Sense. doi:10.1007/978-94-6091-329-7_3

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, 121-137. doi:10.1007/s11292-009-9069-5

Ferguson, C. J., Smith, S. M., 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, 118-132. doi:10.1080/10926770802250678

Giancola, P. R., & Chermack, S. T. (1998). Construct validity of laboratory aggression paradigms: A response to Tedeschi and Quigley (1996). Aggression and Violent Behavior, 3, 237-253. doi:10.1016/S1359-1789(97)00004-9

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

Giancola, P. R., & Zeichner, A. (1995b). An investigation of gender differences in alcohol-related aggression is strongly related. Journal of Studies on Alcohol, 56, 573-579.

Greitemeyer, T., & Osswald, S. (2010). Effects of prosocial video games on prosocial behavior. Journal of Personality and Social Psychology, 98, 211-221. doi:10.1037/a0016997

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2, e124. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23, 524-532. doi:10.1177/0956797611430953

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der “Positive and Negative Affect Schedule” (PANAS) [Studies with a German version of the “Positive and Negative Affect Schedule” (PANAS)]. Diagnostica, 42, 139-156.

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

Lindsay, J. J., & Anderson, C. A. (2000). From antecedent conditions to violent actions: A general affective aggression model. Personality and Social Psychology Bulletin, 26, 533-547. doi:10.1177/0146167200267002

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. The Quarterly Journal of Experimental Psychology, 65, 2271-2279. doi:10.1080/17470218.2012.711335

McCloskey, M. S., Berman, M. E., Noblett, K. L., & Coccaro, E. F. (2006). Intermittent explosive disorder-integrated research diagnostic criteria: Convergent and discriminant validity. Journal of Psychiatric Research, 40, 231-242. doi:10.1016/j.jpsychires.2005.07.004

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. Perspectives on Psychological Science, 7, 109-117. doi:10.1177/1745691611432343

Mohseni, M. R. (2013). Virtuelle Nothilfe. Ein Experiment zum Effekt von virtueller Hilfe, Gewalt und Nothilfe auf Hilfe- und Gewaltverhalten [An experiment on the effect of virtual helping, violence, and emergency assistance on helping and violent behavior.] (Doctoral dissertation). University of Osnabrück, Osnabrück, Germany.

Neurobehavioral Systems. (2010). Presentation (Version 14.5) [Computer software]. Albany, CA: Author. Retrieved from http://www.neurobs.com

Pihl, R. O., Young, S. N., Harden, P., Plotnick, S., Chamberlain, B., & Ervin, F. R. (1995). Acute effect of altered tryptophan levels and alcohol on aggression in normal human males. Psychopharmacology, 119, 353-360. doi:10.1007/BF02245849

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

Rushton, B. (2013). Backdooring it: Defense maneuvers around setback. Illinois Times. Retrieved from http://www.illinoistimes.com

Savage, J. (2004). Does viewing violent media really cause criminal violence? A methodological review. Aggression and Violent Behavior, 10, 99-128. doi:10.1016/j.avb.2003.10.001

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, 934-942. doi:10.1016/j.jesp.2010.06.005

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, 1359-1366. doi:10.1177/0956797611417632

Suris, A., Lind, L., Emmett, G., Borman, P. D., Kashner, M., & Barratt, E. S. (2004). Measures of aggressive behavior: Overview of clinical and research instruments. Aggression and Violent Behavior, 9, 165-227. doi:10.1016/S1359-1789(03)00012-0

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

Tedeschi, J. T., & Felson, R. B. (1994). Violence, aggression, and coercive actions. Washington, DC: American Psychological Association. doi: 10.1037/10160-000

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

Tedeschi, J. T., & Quigley, B. M. (2000). A further comment on the construct validity of laboratory aggression paradigms: A response to Giancola and Chermack. Aggression and Violent Behavior, 5, 127-136. doi:10.1016/S1359-1789(98)00028-7

Valadez, J. J., & Ferguson, C. J. (2012). Just a game after all: Violent video game exposure and time spent playing effects on hostile feelings, depression, and visuospatial cognition. Computers in Human Behavior, 28, 608-616. doi:10.1016/j.chb.2011.11.006

Veit, R., Lotze, M., Sewing, S., Missenhardt, H., Gaber, T., & Birbaumer, N. (2010). Aberrant social and cerebral responding in a competitive reaction time paradigm in criminal psychopaths. NeuroImage, 49, 3365-3372. doi:10.1016/j.neuroimage.2009.11.040

Verona, E., Reed, A., Curtin, J. J., & Pole, M. (2007). Gender differences in emotional and overt/covert aggressive responses to stress. Aggressive Behavior, 33, 261-271. doi:10.1002/ab.20186

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063-1070. doi:10.1037/0022-3514.54.6.1063

Weisman, A. M., Berman, M. E., & Taylor, S. P. (1998). Effects of clorazepate, diazepam, and oxazepam on a laboratory measurement of aggression in men. International Clinical Psychopharmacology, 13, 183-188. doi:10.1097/00004850-199807000-00005

Zillmann, D. (1983). Arousal and aggression. In R. G. Geen & E. Donnerstein (Eds.), Aggression: Theoretical and empirical reviews (Vol. 1, pp. 75-102). New York, NY: Academic Press.