Measurement Invariance and Item Response Theory Analysis of the Taylor Aggression Paradigm

Emily N. Lasko*, David S. Chester

Department of Psychology, Virginia Commonwealth University, USA

In press at *Assessment*

*CORRESPONDING AUTHOR INFO

Emily N. Lasko

916 W. Franklin St.

Richmond, VA

laskoen@mymail.vcu.edu

Abstract

The Taylor Aggression Paradigm (TAP) is a widely used laboratory aggression task, yet item response theory (IRT) analyses of this task are nonexistent. To estimate these aspects of the TAP, we combined data from nine laboratory studies that employed the 25-trial version of the TAP (combined *N* = 1,856). One-factor and four-factor solutions for the TAP data exhibited evidence of measurement invariance across gender (men versus women) and experimental provocation (negative versus positive social feedback), as well as negligible instances of differential item functioning. As such, psychometric properties of the TAP were invariant across binary representations of gender and experimental provocation. Further, trials following low and high provocation were the least informative and those following moderate provocation were the most informative. Scoring approaches to the TAP may benefit from giving greater weight to trials following moderate provocation. Overall, we find great utility in applying IRT approaches to behavioral laboratory tasks.

Keywords: Taylor Aggression Paradigm, measurement invariance, item response theory, psychometrics, provocation

Introduction

Human aggression is a costly and complex phenomenon. Laboratory tasks that accurately and reliably assess aggression are needed to fully understand this behavior. The Taylor Aggression Paradigm (TAP; Taylor, 1967) and variants thereof have emerged as the primary approach to laboratory aggression assessment. Despite the large-scale adoption of this paradigm, several of its psychometric qualities remain largely unexamined. In what follows, we examined the measurement invariance of the TAP across men and women and across experimental provocation conditions. We then used item response theory analyses to identify the invariance and informativeness of individual trials of the TAP.

The Taylor Aggression Paradigm

The Taylor Aggression Paradigm (TAP; Taylor, 1967) is one of the most widely used measures of aggressive behavior. In its most basic form, participants compete in a reaction-time game against an opponent and whoever loses each competition trial receives an aversive stimulus. Typically, the aversive stimuli take the form of electric shocks to the skin or a blast of harsh noise over a set of headphones. Aggression is quantified by the severity of the aversive stimulus that participants select for their opponent to receive if their opponent loses the competition and the participant wins. Most often, the TAP is administered as a computer program called the Competitive Reaction-Time Task and participants compete against a fictitious opponent who enacts a pre-programmed schedule of outcomes (i.e., wins and losses) and provocation (e.g., volume settings for noise blasts; Bushman & Baumeister, 1998). Participants are allowed to see their opponent’s settings as this acts as a within-person provocation manipulation. The intention behind this manipulation is that participants will be provoked into behaving aggressively if they see that their opponent tried to administer a high level of noise. This provocation schedule varies within-participants such that, typically, the earlier trials of the task are the ‘high provocation’ trials (i.e., noise levels 8-10) and the provocation levels of the following trials are randomized.

Whether aggression can be accurately and reliably assessed in contrived laboratory settings such as those inherent to the TAP has been a point of considerable debate (e.g., Tedeschi & Quigley, 1996). However, the TAP has exhibited substantial evidence of construct, convergent, and external validity (Anderson & Bushman, 1997; Chester & Lasko, 2019; Giancola & Parrott, 2008; Giancola & Zeichner, 1995; Hyatt, Zeichner, & Miller, 2019; King & Russell, 2019). Valid measures must also exhibit other important yet underestimated psychometric properties, such as measurement invariance (Flake, Pek, & Hehman, 2017; Hussey & Hughes, in press).

Measurement Invariance

An important property of any psychological measure of a latent construct is *measurement invariance*, which refers to the estimate of whether a latent construct is assessed in the same manner across different groups (Coulacoglou, & Saklofske, 2017; Millsap, 2011). If the invariance of a measure remains un-estimated (an unfortunately common practice: Hussey & Hughes, in press), then investigators cannot know whether a measure can be accurately employed across a diverse array of sampling populations or whether its validity and factor structure is specific to one or more of those populations (Flake, Pek, & Hehman, 2017).

Measurement invariance is empirically estimated via confirmatory factor analysis.

In this modality, there are four increasingly strict levels of measurement invariance (Bialosiewicz, Murphy, & Berry, 2013; Coulacoglou & Saklofske, 2017). First, *configural invariance* tests whether a measure's latent factor structure replicates across groups. For example, men might exhibit a two-factor structure and women might exhibit a three-factor structure for a given measure. Second, *metric invariance* estimates whether members of different groups respond to each of the measure's individual items in a similar way (i.e., they have similar factor loadings on each item). Third, *scalar invariance* then estimates whether the groups have similar average intercepts for each item. Fourth and finally, *strict invariance* estimates whether error variance is consistent across groups. Scalar and strict forms of measurement invariance are categorized together as 'strong' invariance and are often difficult to observe. Indeed, 'strong' invariance is a highly constrained approach that assumes that measurement error is equivalent across groups, which is rarely achieved. When non-invariance is observed, additional analyses are needed to identify the specific sources of the between-group variability.

Item Response Theory and Differential Item Functioning

The measurement invariance approaches summarized above are derived from classic test theory. Although these factor analytic approaches are able to examine the measurement invariance of overall measures, there are several factors that render them less suited to identify specific items that might contribute to non-invariance as compared to other methods such as Item Response Theory (IRT), which offers additional information and advantages that the CFA approach does not. IRT analyses examine the extent to which individual items from psychological measures capture the latent constructs they are intended to assess (Embretson, & Reise, 2013). The process of applying the sequence of equality constraints in CFA can detect *overall group differences* in the measure itself, however if the measure exhibits non-invariance at any level, a series of additional analyses (e.g., removing/adding individual constraints until partial invariance is achieved) are needed to determine the specific sources of non-invariance. Conversely, IRT approaches were designed to examine the functioning of a measure's individual items and therefore yields more information per item compared to CFA. As such, if non-invariance is observed the IRT approach is ideal to dig deeper into the specific sources of that non-invariance. Of particular relevance to the testing of measurement invariance are IRT analyses that examine differential item functioning (DIF). DIF occurs when a given item captures the target latent construct differently between two or more subgroups (see Methods for more detail). Yet what subgroups would be most important to examine for signs of measurement invariance and DIF?

Invariance Across Gender Identities

The grouping variable that is often invoked in differentially influencing aggression is gender. Self-identified men tend to be more physically aggressive than self-identified women (Archer, 2004), which is reflected in higher TAP scores observed for men (as compared to women; Zeichner, Parrott, & Frey, 2003). These gender differences in laboratory aggression may be due to real differences in these groups' aggressive behavior or they may simply be due to the TAP actually exhibiting non-invariant psychometric properties for men compared to women. For example, it is possible that men and women approach the task in different ways or the task may simply be scaled differently for men versus women. Indeed, many contextual factors that are simulated by the TAP (e.g., opponent provocation) moderate the effect of gender on aggression

(Arriaga & Aguiar, 2019; Bettencourt & Miller, 1996; Björkqvist, 2018). The crucial role of gender in the study of aggression necessitates that the measurement invariance of the TAP, by gender identity, is investigated.

Invariance Across Experimental Manipulations of Provocation

Although the TAP includes built-in provocation in the form of noise blasts delivered by the participant's opponent, many studies also include experimental manipulations prior to the TAP that are intended to further provoke aggressive behavior. Examples of these added provocations include manipulations of being socially excluded (versus included; e.g., Chester & DeWall, 2017) and receiving feedback from another person that is insulting (versus complimentary; e.g., Chester & Lasko, 2019). These provocation manipulations reliably heighten aggression on the TAP (Chester & Lasko, 2019). However, these mean differences between provoked and unprovoked participants may be due, in part, to the effect of the provocation manipulation on the psychometric properties of the TAP itself and not the latent aggression construct. For example, an already-provoked participant might be more reactive to the provocation inherent to the TAP, creating a different behavioral profile than participants who were initially unprovoked. Measurement invariance analyses are able to examine this possibility.

Present Research

The TAP has been the subject of many efforts to ascertain its psychometric validity. Yet examinations of this paradigm's measurement invariance across key grouping variables are still lacking. In what follows, we combined nine existing datasets on which we performed measurement invariance analyses using both confirmatory factor analyses and IRT according to a two-part preregistered plan (part one: https://osf.io/t59gu; part two: https://osf.io/j3nmt). We predicted that the TAP would exhibit measurement invariance across gender and experimental provocation groups, given the substantial evidence that the task exhibits considerable validity when these groups are combined (e.g., Chester & Lasko, 2019).

Methods

Participants

The data used in the present analyses were from 1,886 undergraduate participants across nine separate studies. All participants were recruited from an introductory psychology subject pool and completed the study for course credit. The following exclusion criteria applied to all nine studies (except where noted): 1) participants must have been at least 18 years old, 2) participants could not have a hearing disorder or other hearing sensitivity. Thirty participants were manually removed from this original dataset for either failing to indicate their gender or indicating a gender other than female or male, resulting in a final sample size of 1,856. Participant demographic information from each study is presented in Table 1. In Studies 1 and 9, 17-year-old students were permitted to participate via a parental consent waiver.

Table 1

Separated and Combined Descriptive Statistics of Participant Demographics from Each Study.

Study	*n*	Females	Males	Age *M (SD*)	Age Range	% Provoked
1	367	250	109	18.65 (0.98)	17 – 26	49%
2	176	121	54	19.23 (3.64)	18 – 62	51%
3	112	71	41	19.52 (2.27)	18 – 36	100%
4	218	165	51	19.14 (2.80)	18 – 42	0%
5	155	127	28	18.83 (1.03)	18 – 26	100%
6	167	118	45	19.04 (1.72)	18 – 30	51%
7	211	132	73	20.02 (4.58)	18 – 55	50%
8	193	126	62	19.11 (2.30)	18 – 40	52%
9	287	197	86	18.83 (1.23)	17 – 32	50%
Total	1886	1307	549	19.10 (2.50)	18 – 62	51%

Materials

Taylor Aggression Paradigm. In each of the nine studies, participants completed a computerized version of the 25-trial Taylor Aggression Paradigm (TAP; Bushman & Baumeister 1998; Taylor, 1967). Participants were instructed that they would be competing in a reaction-time game against a same-sex stranger. In reality, participants were playing against the computer with pre-programmed wins and losses. In each of the task's 25 trials, participants competed to press a button faster and a loud noise blast was delivered to the slower player. Participants chose the volume and duration of the noise their opponent would hear if they lost. Similarly, participants were told that their opponent would choose the volume and duration of the noise that they (the participant) would hear. Volume levels ranged from Level 1 (60dB) to Level 10 (105 dB), in addition to a non-aggressive option (Level 0). Duration also ranges from Level 0 (0 seconds) to Level 10 (5 seconds), increasing duration length by half-second increments. Wins and losses were randomized within participants according to the task's default settings (see Figure 2; Bushman & Baumeister, 1998) and this pattern of randomization was held constant across participants.

Procedure

Study 1[1]. Participants reported their demographics and were then randomly assigned to either be rejected or not via a widely used social rejection paradigm called Cyberball (Williams, Cheung, & Choi, 2000). As social rejection is a form of provocation, this task served as the provocation manipulation. Cyberball took the form of a virtual ball-tossing game, which participants ostensibly played with two other same-sex students. In the rejection condition, participants were assigned to receive only three ball tosses at the beginning and then none while their partners tossed the ball to each other (Chester, Lynam, Milich, & DeWall, 2017). After the provocation manipulation, all participants completed the TAP against one of their Cyberball partners.

Study 2[2]. Participants reported their demographics and then were randomly

[1] Data from Study 1 have been previously published:
Chester, D. S., & DeWall, C. N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. Journal of Personality and Social Psychology, 112(3), 413-430.
Hyatt, C. S., Chester, D. S., Zeichner, A., & Miller, J. D. (2020). Facet-level analysis of the relations between personality and laboratory aggression. Aggressive Behavior, 46(3), 266-277.

[2] Data from Studies 2, 4, and 5 have been previously published:

assigned to either be provoked or not via an essay feedback paradigm in which participants were instructed to write a brief essay about an important time in their life and then exchanged essays with an ostensible partner in a different room who would give them feedback on the essay. Participants were randomly assigned to receive either negative feedback (8/35 points, “One of the WORST essays I’ve EVER read!”) or positive (33/35 points, “Great essay!”) feedback (Bushman & Baumeister, 1998; Chester & DeWall, 2017). After the provocation manipulation, participants completed the TAP against their essay feedback partner.

Study 3. Participants reported their demographics and then were all provoked via the same essay feedback provocation paradigm as Study 2. After being provoked, participants completed the TAP against their essay feedback partner.

Study 4[2]. Participants reported their demographics and then completed the TAP. No provocation manipulation was employed in this study.

Study 5[2]. Participants reported their demographics, completed a battery of personality questionnaires, and then were all provoked via the same essay feedback provocation paradigm as previous studies. This manipulation followed the same procedure as the previous essay feedback studies with the exception that participants wrote about a time they were angry rather than about an important time in their life

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.

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.

Chester, D. S., Merwin, L. M., & DeWall, C. N. (2015). Maladaptive perfectionism’s link to aggression and self-harm: Emotion regulation as a mechanism. *Aggressive Behavior, 41*(5), 443-454.

(Chester, Merwin, & DeWall, 2015). After being provoked, participants completed the TAP.

Study 6[3]. Participants reported their demographics and were then randomly assigned to either be provoked or not via the same Cyberball task as Study 1. After the provocation manipulation, all participants then completed the TAP.

Study 7[3]. Participants were randomly assigned to either be provoked or not by the same essay feedback provocation paradigm as Study 2 and then completed the TAP.

Study 8[3]. This study was nearly identical to Study 7 with the exception that the TAP was counterbalanced with two other behavioral aggression tasks.

Study 9. This study procedure was identical to Study 6.

Data Analysis

Confirmatory factor analysis. We fit a one-factor confirmatory factor analysis (CFA) with maximum likelihood estimation using the *lavaan* package (version 0.6-5; Yves, 2012) for R statistical software (version 3.6; R Core Team, 2019). Of the total 1,856 participants, 1770 were used in the CFA due to listwise deletion of missing

[3] Data from Studies 6, 7, and 8 have been previously published:
Chester, D. S., & DeWall, C. N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology, 112*(3), 413-430.
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.
Chester, D. S. (2019). Beyond the aggregate score: Using multilevel modeling to examine trajectories of laboratory-measured aggression. *Aggressive Behavior, 45*(5), 498-506.
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.
Chester, D. S. (2019). Beyond the aggregate score: Using multilevel modeling to examine trajectories of laboratory-measured aggression. *Aggressive Behavior, 45*(5), 498-506.

observations. We conducted Little's MCAR test using the BaylorEdPsych v0.5 package for R statistical software. The results were non-significant, X2 (81) = 55.01, *p* = .99, and fewer than 5% of observations were missing (4%), therefore no further steps were taken to address missing data. The CFA examined the fit of a model in which all 50 TAP items (25 trials x 2 settings per trial) were set to load onto a single latent "aggression" factor. One randomly chosen item's factor loading was set to 1 to allow for intercept estimation. We decided on this initial single-factor structure because this is one of the most commonly used scoring strategies for the TAP (i.e., a single average across all trials; Chester & Lasko, 2019).

To test the measurement invariance of this factor model, we first ran an unconstrained CFA and then applied increasingly strict equality constraints onto the remaining 49 factor loading parameters. We then compared the model fit of each /constrained model (with parameters set to equal between men and women or provoked and unprovoked participants) to the model before it. In the first constrained model, only the factor loadings were set to equal (i.e., metric invariance). In the second constrained model, both the loadings and intercepts were set to equal across groups (i.e., scalar invariance). In the final constrained model, the loadings, intercepts, and residuals were set to equal across groups (i.e., strict invariance). The fit indices we used to compare these models were X^2, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and the Tucker Lewis Index (TLI).

Item response theory - differential item functioning. To determine which of the 50 TAP items were the individual sources of the non-invariance, we conducted differential item functioning (DIF) analyses using the *mirt* package (version 1.31;

Chalmers, 2012) for R statistical software, following the two-step DIF analytic procedures outlined by Chalmers, Counsell, and Flora (2016). In the first step, we tested all 50 TAP items simultaneously for potential DIF using the one-factor CFA models to impose equality constraints on each individual item's factor loadings. We then categorized each item as either a 'test item' that potentially exhibited DIF (if the X^2 invariance test for that item was statistically significant [i.e., $p < .05$]), or as an 'anchor item' that did not potentially exhibit DIF (the X^2 invariance test for that item was not statistically significant [i.e., $p > .05$]). In the second step of the analysis, we re-ran the CFAs that applied equality constraints to each test item's slopes and intercepts. If the X^2 invariance test for any test item was statistically significant [i.e., $p < .05$], we deemed that item as exhibiting DIF. Both of these steps are necessary to obtain accurate parameter estimates.

***Item response theory - differential test functioning*.** To determine the magnitude of any DIF effects on the validity of the overall TAP, we also conducted differential test functioning (DTF) analyses (Chalmers, Counsell, & Flora, 2016). Instead of focusing on individual items, DTF estimates the impact of DIF on a measure's aggregated score. To do so, these analyses compute two statistics to describe whether two groups, given equivalent levels of the latent trait (e.g., aggression), differ significantly on their expected TAP scores. The *signed DTF (sDTF)* statistic reflects overall measurement bias, across all items, in favor of one group over another at the omnibus (i.e., test) level. The *unsigned DTF (uDTF)* statistic represents the degree to which the expected TAP scores differ between two groups at varying levels of the latent trait. The latter is commonly represented visually via expected score curves, which plot a given range of expected TAP scores as a function of varying levels of the latent trait, separately for each group. The uDTF reflects the degree to which the curves for each group overlap with each other. If one or both of these DTF statistics are statistically significant, then this indicates nontrivial DTF due to the non-invariant items identified in the DIF analyses.

Results

Confirmatory Factor Analyses

The fit of the single-factor model was unexpectedly poor, $X^2(1, 175) = 21{,}934.44$, $p < .001$, Root Mean Square of Approximation (*RMSEA*) = .10, Standardized Root Mean Square Residual (*SRMR*) = .07, Tucker-Lewis Index (*TLI*) = .66, and Comparative Fit Index (*CFI*) = .67. Standardized factor loadings for all TAP items are displayed in Supplemental Table 1.

Measurement invariance by gender. According to our pre-registered criteria, the single factor CFA model’s fit to the TAP data exhibited configural and metric invariance, but not scalar or strict invariance by gender (0 = male, 1 = female; Table 2).

Measurement invariance by provocation. The single factor CFA model’s fit to the TAP data again exhibited configural and metric invariance, but not scalar or strict invariance by provocation condition (1 = provoked, 0 = unprovoked; Table 3).

Table 2

Model Fit Statistics for Each of the Gender Invariance Models Using the One-Factor Structure.

Model	χ^2(*df*)	*RMSEA*	*CFI*	*SRMR*	*TLI*	Model comp.	Δχ^2(Δdf)	Δ*CFI*	Δ*RMSEA*	Δ*SRMR*
M1:Configural	24,005.63 (2,350)	.10	.66	.07	.65	-	-	-	-	-
M2: Metric	24,004.84 (2,399)	.10	.66	.07	.66	M1	00.79 (49)	.00	.00	.00
M3: Scalar	24,241.66 (2,448)	.10	.66	.07	.66	M2	236.82 (49)	.00	.00	.00
M4: Strict	25,522.08 (2,498)	.10	.64	.07	.65	M3	1,280.42 (50)	-.02	.00	.00

Note. df = degrees of freedom, *RMSEA* = Root Mean Square of Approximation, *CFI* = Comparative Fit Index, *SRMR* = Standardized Root Mean Squared Residual, *TLI* = Tucker-Lewis Index, Model comp. = model being compared to (e.g., comparing model 1 [M1] to model 2 [M2]).

Table 3

Model Fit Statistics for Each of the Provocation Invariance Models Using the One-Factor Structure.

Model	χ^2(*df*)	*RMSEA*	*CFI*	*SRMR*	*TLI*	Model comp.	$\Delta\chi^2$(Δdf)	Δ*CFI*	Δ*RMSEA*	Δ*SRMR*
M1: Configural	24,309.42 (2,350)	.10	.66	.07	.65	-	-	-	-	-
M2: Metric	24,354.08 (2,399)	.10	.66	.07	.65	M1	44.66 (49)	.00	.00	.00
M3: Scalar	24,423.81 (2,448)	.10	.66	.07	.66	M2	69.74 (49)	.00	.00	.00
M4: Strict	24,508.97 (2,498)	.10	.66	.07	.67	M3	85.16 (50)	.00	.00	.00

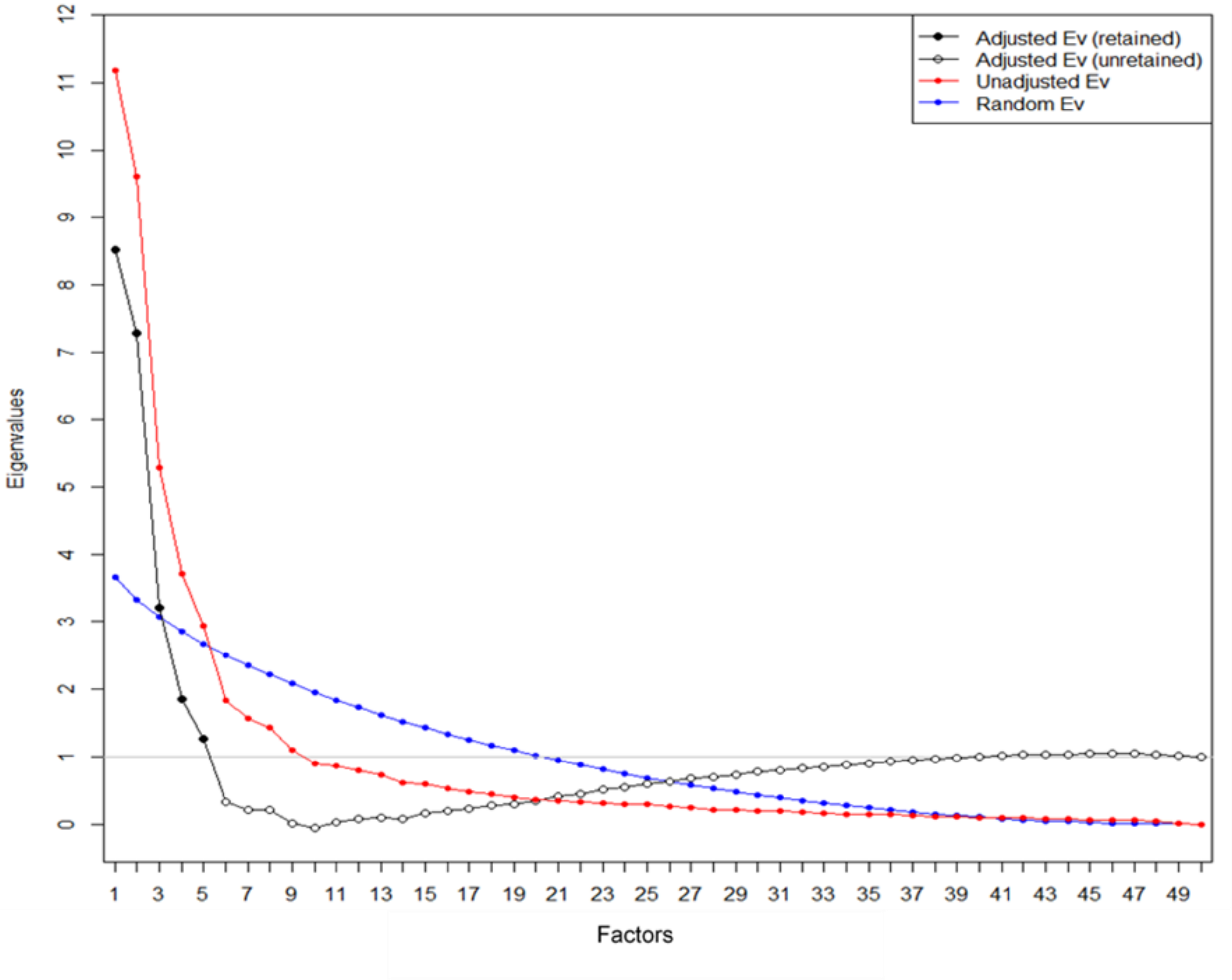
Note. *df* = degrees of freedom, *RMSEA* = Root Mean Square of Approximation, *CFI* = Comparative Fit Index, *SRMR* = Standardized Root Mean Squared Residual, *TLI* = Tucker-Lewis Index, Model comp. = model being compared to (e.g., comparing model 1 [M1] to model 2 [M2]).

Exploratory Factor Analyses

Due to the poor model fit of the initial single-factor CFA, we conducted an exploratory factor analysis (EFA) on all 50 TAP items to identify a more appropriate underlying structure of the TAP data using the *stats* (version 3.6.0; R Core Team, 2019) and *nFactors* (version 1.9.12; Raiche, 2010) packages for R statistical software. The factor solution was examined using both varimax and promax rotations of the factor loading matrix, which did not produce meaningfully different factor solutions. For the sake of simplicity, we therefore only present EFA results that used varimax rotation. Based on a parallel analysis using the *nFactors* (version 1.9.12; Raiche, 2010) package for R statistical software, five factors from the EFA were initially retained, which explained 55% of the variance (Figure 1; Table 4).

Figure 1

Scree Plot Depicting the Eigenvalues of Each Factor from the Parallel Analysis. The Initial Factors from the Actual Data (in red) are Presented Alongside Those Derived From Random Noise (in Blue), Resulting in a Final, Adjusted Five Factor Solution (in Black).



Note. Ev = eigenvalues.

Table 4

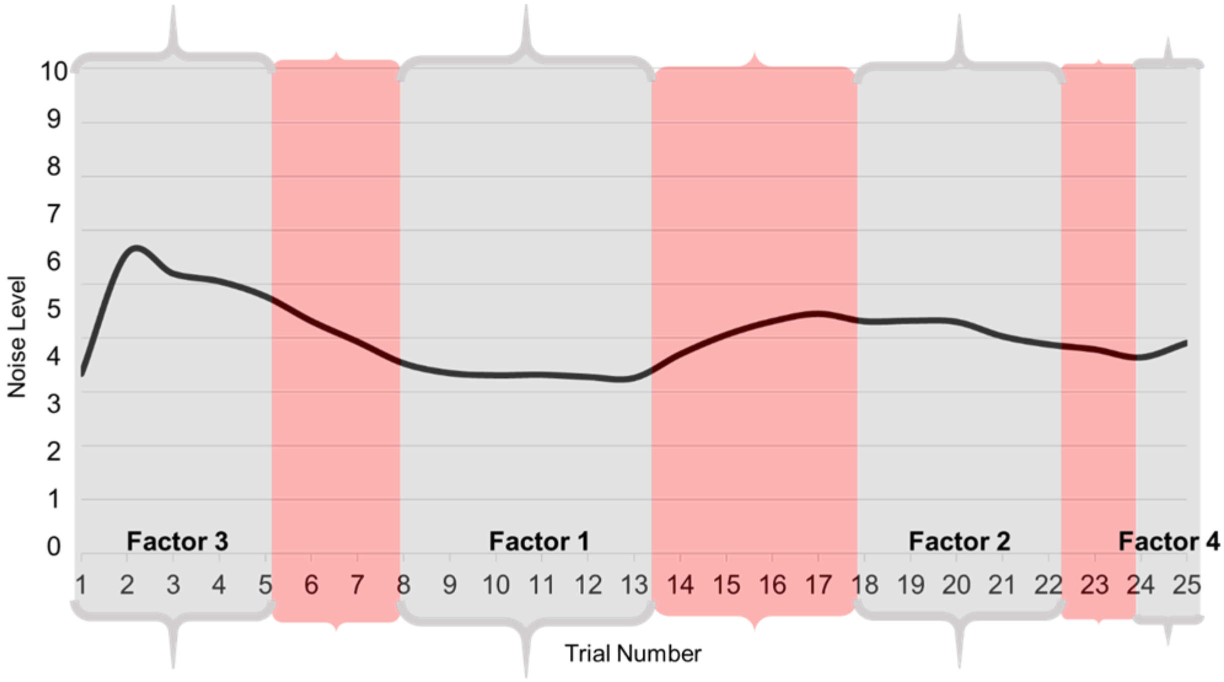
Results from the Horn's Parallel Analysis of all 50 Taylor Aggression Paradigm Items.

Factor	Adjusted Eigenvalue	Unadjusted Eigenvalue	Proportion of Variance Explained
1	8.52	11.19	.17
2	7.29	9.61	.15
3	3.21	5.28	.13
4	1.84	3.71	.07
5	1.26	2.94	.04

We retained 34 of the 50 TAP items, each of which exhibited a factor loading exceeding |.40|. Sixteen TAP items were removed because they exhibited factor loadings below this threshold or because they exhibited cross-factor loadings within |.20|. Removing these items left the fifth factor with no items that exhibited sufficient factor loadings. Therefore, this fifth factor was eliminated and a four-factor solution was adopted (Figure 2).

Figure 2

Trials of the Taylor Aggression Paradigm Categorized by the Factor They Loaded Onto (in Gray), Alongside Trials With Problematic Cross-Factor Loadings (in red). The Black Line Reflects Opponent Provocation Levels on Each Trial (Averaged Across the Opponent's Duration and Volume Settings).



Note. Factor 1 included trials the volume and duration of trials 8 – 13. Factor 2 included the volume and duration of trials 18 – 22. Factor 3 included the volume of trials 1 – 5 and the duration of trials 2 – 4. Factor 4 included the volume and duration of trials 24 – 25.

Confirmatory factor analyses revealed that the four-factor solution derived from the EFA showed modest fit to the data, $X^2(521) = 6{,}278.29$, *RMSEA* = .08, *SRMR* = .05, *CFI* = .86, *TLI* = .85. However, this model fit was substantially improved compared to our original, single-factor model, $\Delta X^2(654) = 15{,}656.10$, $p < .001$. Standardized factor loadings for all TAP items are displayed in Supplemental Table 2.

Measurement invariance by gender. As with the single-factor model, the four-factor model’s fit to the 34 TAP items exhibited configural and metric invariance, but not scalar or strict invariance (Table 5).

Measurement invariance by provocation. As in the single-factor model, the four-factor model’s fit to the TAP data exhibited configural and metric invariance, but not scalar or strict invariance (Table 6).

Table 5

Model Fit Statistics for Each Gender Invariance Model Using the Four-Factor Structure.

Model	χ^2(df)	*RMSEA*	*CFI*	*SRMR*	*TLI*	Model comp.	Δ χ^2(Δdf)	Δ*CFI*	Δ*RMSEA*	Δ*SRMR*
M1: Configural	7,097.21 (1,042)	.08	.85	.05	.84	-	-	-	-	-
M2: Metric	7,132.16 (1,072)	.08	.85	.05	.85	M1	34.97 (30)	.00	.00	.00
M3: Scalar	7,188.97 (1,102)	.08	.85	.05	.85	M2	56.82 (30)	.00	.00	.00
M4: Strict	8,082.61 (1,136)	.08	.83	.05	.83	M3	893.64 (34)	-.02	.00	.00

Note. df = degrees of freedom, *RMSEA* = Root Mean Square of Approximation, *CFI* = Comparative Fit Index, *SRMR* = Standardized Root Mean Squared Residual, *TLI* = Tucker-Lewis Index, Model comp. = model being compared to (e.g., comparing model 1 [M1] to model 2 [M2]).

Table 6

Model Fit Statistics for Each Provocation Invariance Model Using the Four-Factor Structure.

Model	χ^2(df)	*RMSEA*	*CFI*	*SRMR*	*TFI*	Model comp.	$\Delta\chi^2$(Δdf)	*ΔCFI*	*ΔRMSEA*	*ΔSRMR*
M1: Configural	7,097.21 (1,042)	.08	.85	.05	.84	-	-	-	-	-
M2: Metric	7,132.16 (1,072)	.08	.85	.05	.85	M1	20.80 (30)	.00	.00	.00
M3: Scalar	7,188.97 (1,102)	.08	.85	.05	.85	M2	47.09 (30)	.00	.00	.00
M4: Strict	7,471.10 (1,136)	.08	.85	.05	.85	M3	124.63(34)	.00	.00	.00

Note. *df* = degrees of freedom, *RMSEA* = Root Mean Square of Approximation, *CFI* = Comparative Fit Index, *SRMR* = Standardized Root Mean Squared Residual, *TLI* = Tucker-Lewis Index, Model comp. = model being compared to (e.g., comparing model 1 [M1] to model 2 [M2])

Exploratory Differential Item Functioning Analyses

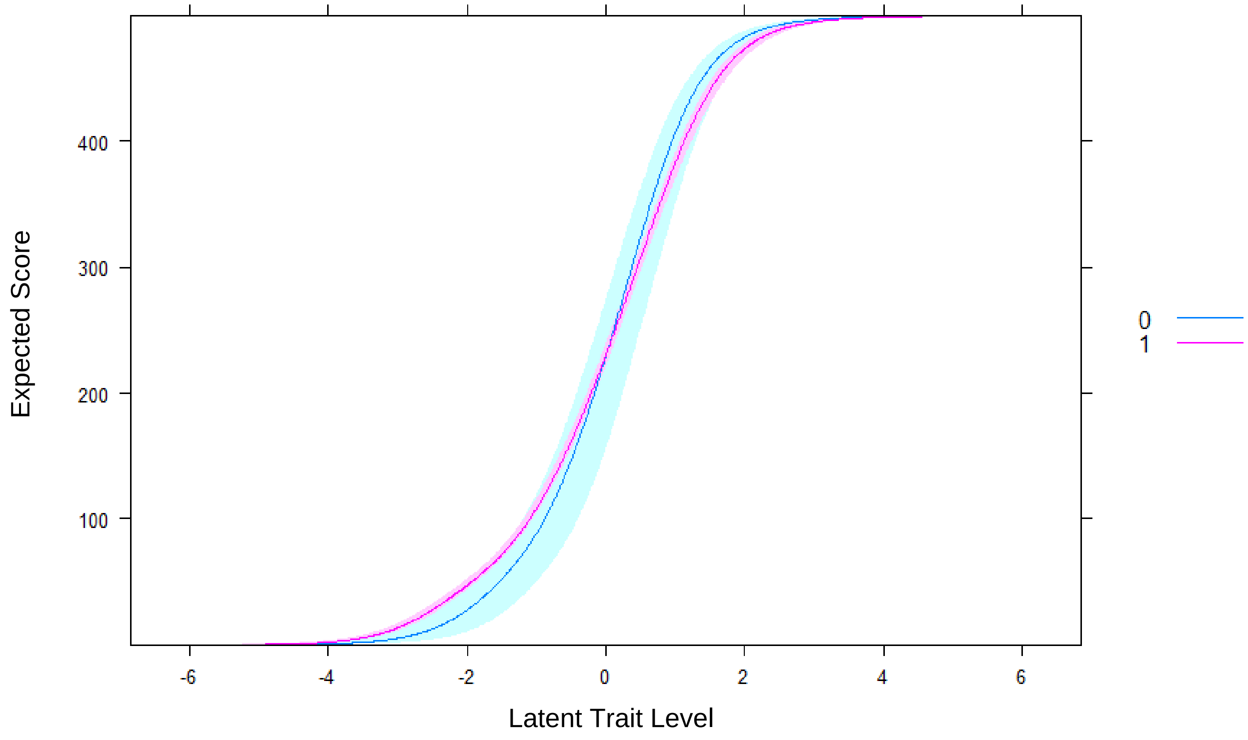
To determine which of the 50 TAP items were the individual sources of the non-invariance we observed in our prior analyses, we conducted differential item functioning (DIF) analyses.

DIF by gender. In the first step of the DIF analyses, only one of the 50 TAP items (i.e., the duration of TAP trial 6) initially exhibited potential DIF between men and women, *AIC* = -5.49, *BIC* = -0.01, $X^2(1) = 7.49$, *p* = .006 (all 50 initial DIF test results are presented in Supplemental Table 3; all 50 item information plots, separated by gender, are depicted in Supplemental Document 1 [https://osf.io/a37h8/]). The second phase of the DIF analysis revealed that this item no longer showed significant DIF, *AIC* = 1.22, *BIC* = 6.69, X^2 (1) = 0.78, *p* = .377.

Follow-up Differential Test Function (DTF) analyses suggested that the effect of the one item that initially exhibited DIF was negligible on the overall TAP, *sDTF* = -1.25 (*95% CI* = -17.36, 12.73), *uDTF* = 7.51 (*95% CI* = 4.94, 18.79), omnibus *p* = .866. These inferential results were reflected in the largely overlapping expected score plots of men and women (Figure 3).

Figure 3

Expected TAP Scores as a Function of Trait Levels of Aggression Between Men (0) and Women (1).



DIF by provocation. Unlike the gender-based DIF analyses, 47 of the 50 TAP items initially showed potential DIF between provoked and unprovoked participants (see Supplemental Table 4 for DIF results for each of the 50 TAP items and see Supplemental Document 2 for individual item information plots). However, the second phase of the DIF analysis revealed that only seven items ultimately showed significant DIF (Table 7; see Supplemental Document 2 for individual expected score plots [https://osf.io/n8vyc/]). Six of these seven DIF items were localized to the last eight trials of the TAP (i.e., trials 18-25).

Follow-up DTF analyses suggested that the effect of the items that exhibited DIF was negligible on the overall TAP, *sDTF* = -1.22 (*95% CI* = 0.16, 0.79), *uDTF* = 3.89 (*95% CI* = 1.39, 6.98), omnibus *p* = .204. These inferential results were reflected in the largely overlapping expected score plots of provoked and unprovoked participants (Figure 4).

Table 7

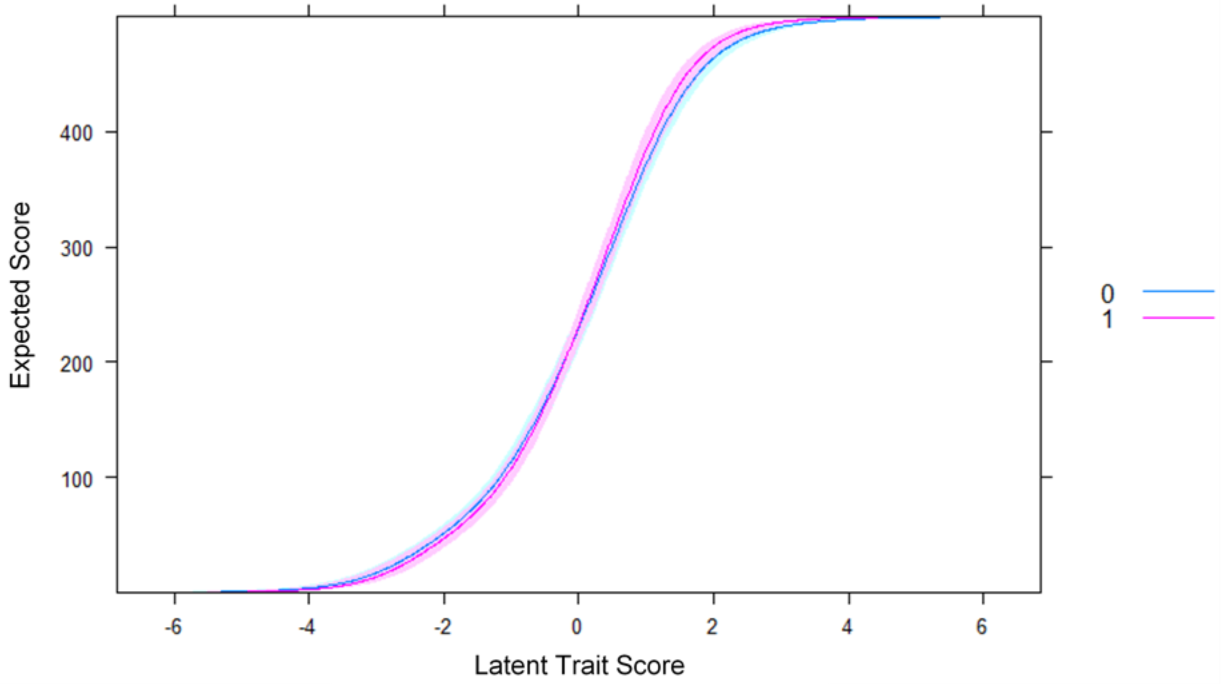
Fit statistics and significance tests for Taylor Aggression Paradigm items that showed Differential Item Functioning between provoked and unprovoked participants.

TAP Response	*AIC*	*BIC*	$X^2(1)$	*p*
Trial 10 (Intensity)	-3.83	1.65	5.83	.026
Trial 20 (Intensity)	-4.26	1.21	6.26	.012
Trial 24 (Intensity)	-2.14	3.33	4.14	.042
Trial 25 (Intensity)	-1.85	3.36	3.85	.050
Trial 18 (Duration)	-4.01	1.47	6.01	.014
Trial 22 (Duration)	-2.37	3.10	4.27	.046
Trial 24 (Duration)	-4.22	1.25	6.22	.013

Note. TAP = Taylor Aggression Paradigm, *AIC* = Akaike Information Criterion, *BIC* = Bayesian Information

Figure 4

Expected TAP Scores as a Function of Trait Levels of Aggression Between Provoked (1) and Unprovoked (0) Participants.

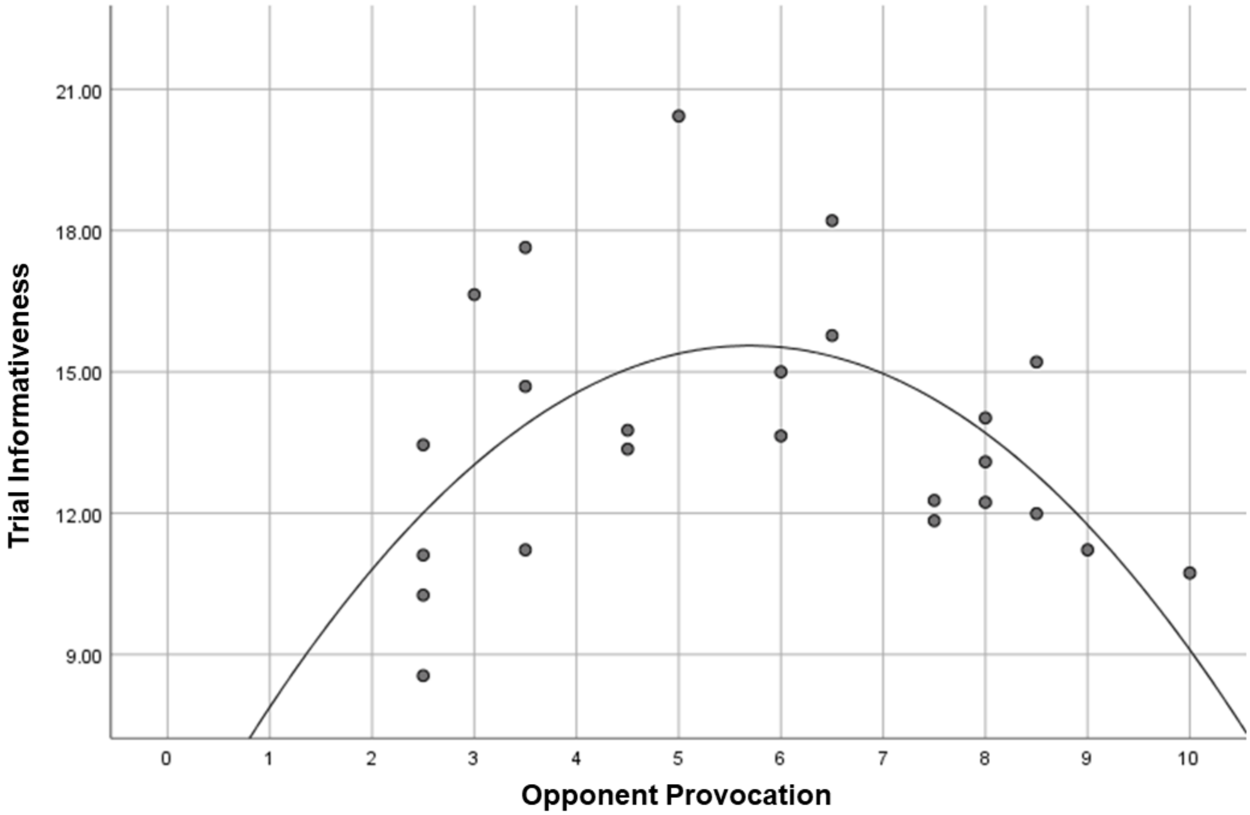


Item Informativeness by Provocation (Exploratory)

We observed an interesting pattern in the item information curves that we computed as part of our DIF analyses, in which many trials exhibited curves that portrayed poor informativeness (e.g., Supplemental Figure S5). Less informative items appeared to be (A) at the beginning or end of the task and (B) preceded by high or low opponent provocation on the previous trial. As part of subsequent exploratory analyses, we extracted area-under-the-curve (AUC) values from each trial's item information curve (collapsing across duration and volume settings; and excluding the first TAP trial because it was not preceded by any noise blasts from the opponent). Plotting these 24 AUC values against the TAP opponent's provocation level from the previous trial (averaging the pre-set duration and volume settings along their 0-10 continuum within each trial) revealed a curvilinear relationship between these two variables (Figure 5). Indeed, the informativeness of each trial was lower for trials that followed low and high provocation and were optimal at moderate levels, peaking at the midpoint of provocation (i.e., 5).

Figure 5

The Curvilinear Association Between the Informativeness of TAP Trials 2-24 and Their Opponent's Provocation Level on the Previous Trial. The curve reflects the informativeness of individual items *(i.e., trials) of the TAP.*



Discussion

The TAP, and the many variants thereof, is an important measure of aggression in the laboratory. Yet, the TAP's ability to measure aggression in a similar way across men and women and across experimentally provoked and unprovoked participants has remained uninvestigated. In this investigation, we estimated this unknown psychometric quality by combining TAP data from over 2,000 participants, testing the hypotheses that this paradigm would exhibit measurement invariance across gender and experimental provocation groups.

Factor Structure of the TAP

The initial one-factor solution to the TAP data exhibited poor model fit. This is somewhat in line with our previous research demonstrating that a multi-factor structure may fit the data better than a single factor (Chester & Lasko, 2019). Subsequently, an exploratory factor analysis returned a four-factor solution. Unexpectedly, the model fit of this four-factor model was also relatively poor (although still an improvement upon the one-factor model). These four factors did not appear to map well onto any structural aspects of the TAP that investigators have previously used to delineate different metrics of aggression (e.g., the first unprovoked trial versus subsequent provoked trials; Lawrence & Hutchinson, 2013). It remains uncertain why the data structure aggregated into these four factors. The provocation levels among the factors did differ, however these differences were modest ones (i.e., provocation point difference of ~1) and thus we believe are insufficient to warrant a distinction between 'low', 'moderate', and 'high' provocation factors. Nonetheless, these interpretations are inherently subjective; as such, readers and other scholars are welcome to reinterpret our results to reach their own conclusions. More research is needed to test the replicability of the four-factor structure we observed. If it is replicated, the underlying reasons for this data structure and the implications it may have for improved quantification strategies for the TAP. Although the first TAP trial did not load as strongly onto the first factor as the other trials, this cannot be interpreted as conclusive evidence that it represents a separate construct as the difference in the factor loadings was a modest one at best. Rather, our findings overall suggest that researchers' treatment of the first, unprovoked TAP trial as a measure of a different construct than later trials may in fact be unwarranted. However, this interpretation remains subjective and additional research in this area is needed to confirm these conclusions.

Factor analyses also showed that the duration and volume settings from the same trial almost always loaded onto the same factor, suggesting that these two metrics are not meaningfully distinct from one another. Volume and duration settings also exhibited similar levels of DIF. Thus, the duration and intensity settings for the TAP's aversive stimuli may largely be redundant and quantification strategies that aggregate across these two aggression modalities are employing a valid means of increasing the internal consistency of the TAP by combining indices of the same construct (Chester & Lasko, 2019). However, the use of both volume and duration indices requires participants to make twice as many responses (i.e., decisions) as they would if only one index were used. Given the well-established literature on decision fatigue (e.g., Pignatiello, Martin, & Hickman, 2020), this use of redundant settings on each trial inflicts meaningful costs on participants that may bias their responses. Therefore, future versions of the TAP might benefit from using only the duration or volume setting as a means to reduce participant burden without undermining the reliability or validity of the TAP.

Both the one-factor and four-factor solutions failed to exhibit adequate model fit to the TAP data, which may undermine the validity of our previous prescription to take a one-factor approach to scoring the TAP (Chester & Lasko, 2019). Even the four-factor solution, the ‘best’ fitting model as determined by exploratory factor analyses, did not exhibit good model fit. This inability to find adequate model fit to the data is reflective of a broader psychometric trend, in which the confirmatory model fit of even the most widely-used and well-validated questionnaires’ (e.g., the Big Five Inventory) factor structures often fail to reach adequacy in independent datasets (Hussey & Hughes, in press). It may be that the experimental elements of the task (e.g., varying levels of provocation, wins and losses) or the nested structure of the data (i.e., volume and duration items nested within trials) undermined the emergence of any coherent factor solution. Perhaps a hierarchical factor structure is more appropriate, necessitating a bi-factor model. Future work is needed to determine the underlying reasons for this poor model fit, such as systematically varying the experimental elements of the task and estimating their effects on the factor structure or accounting for the TAP’s nested data structure (e.g., via exploratory structural equation modeling). Doing so would not only improve our psychometric understanding of the TAP, but would serve as an important example for future work on the structural validity of behavioral laboratory tasks. In the end, we are unable to determine the underlying reason for these poor model fit estimates and urge caution in interpreting our findings in light of these issues.

Invariance by Gender

We found that the TAP exhibited both configural and metric invariance across men and women, which supports our hypotheses and the validity of this task across both groups. However, we did not observe evidence of ‘strong’ invariance (i.e., scalar and strict invariance). These latter forms of invariance are often difficult for scholars to obtain, given their harsh assumptions and standards (Asparouhov & Muthén, 2014; Davidov, Muthen, & Schmidt, 2018). Although not unrealistic or unachievable in all cases, strict invariance is a high threshold. Indeed, only 4% of a sample of 15 commonly used measures exhibited such equivalent factor structure, factor loadings, *and* item-level intercepts across groups (i.e., strict invariance; Hussey & Hughes, in press). It may be particularly difficult to obtain this level of invariance with larger samples as it becomes easier to detect smaller between-group differences (i.e., non-invariance). This may have been particularly true for the large sample that we employed, considering the results of our IRT analyses. Specifically, our DIF and DTF analyses suggested that the influence of any non-invariance across gender groups had little influence on the overall measure. As such, the preponderance of the evidence supports the claim that the TAP is a valid aggression measure across these gender identity groups and even suggest that men's and women's aggression may not be as different as many assume.

Invariance by Provocation

As with gender groups, experimentally-provoked and -unprovoked groups of participants exhibited configural and metric invariance, though not scalar or strict invariance. For reasons outlined above, we take these findings to be generally supportive of our prediction that provoked and unprovoked groups would exhibit invariance on the TAP. We further found that any non-invariance likely arose from nine items that largely occurred in later trials of the TAP, though these non-invariant items had a negligible influence on the overall TAP's validity. It remains unclear why later trials would be more likely to exhibit non-invariance, perhaps experimental provocation may exacerbate fatigue effects that present towards the end of the task. Accordingly, it may be that the invariance of the TAP across experimental provocation conditions might be improved by removing later trials, though further work is needed to empirically validate such possibilities. Alternatively, it is possible that greater provocation (i.e., louder noise blast settings by the opponent) near the end of the task, compounded by the provocation received prior to the task, created a ceiling effect that contributed to the non-invariance we observed for these trials. Nonetheless, these findings overall suggest that experimental provocation manipulations that precede the TAP do not invalidate this paradigm and that the TAP is a valid aggression measure across both unprovoked and provoked conditions.

Informativeness of TAP Trials Based on Prior Provocation

The provocation that is inherent to the task is intended to influence participants' aggressive behavior. Yet, our analyses revealed it may have another, unseen influence. The informativeness of each trial of the TAP (i.e., the extent to which each trial reveals meaningful information about the latent aggression construct), exists as a curvilinear function of provocation. Trials following relatively high or low provocation are less informative than those following moderate levels of provocation. This is likely due, in part, to ceiling and floor effects that arise from the human tendency towards reciprocation. For instance, if the opponent selected a 10, most participants then retaliated towards the upper ceiling of the aggression response range. Conversely, opponents who selected low levels of provocation tended to elicit a reciprocal response towards the floor of the aggression response range. Whereas after a moderate amount of provocation, participants were free to respond broadly in either direction of the response scale. Further, ambiguous levels of provocation allowed for individual differences in the tendency towards retaliatory escalation or conciliatory de-escalation to express themselves. Our findings suggest that the common practice of focusing only on TAP items that follow low or high provocation may be psychometrically unsound for samples comprised of undergraduate student populations. Within this population, a high-provocation-focused or low-provocation-focused psychometric approach may obscure true underlying effects that are only detectable among the more informative trials of the TAP. Trials that involve high or low provocation may have more utility among clinical populations (e.g., borderline personality disorder, antisocial personality disorder) or aggressive forensic populations. More research is needed to determine how best to optimize quantification strategies for the TAP that take advantage of the differential informativeness of the trials, especially across population types (Hyatt, Chester, Zeichner, & Miller, 2019).

Limitations and Future Directions

The chief limitation of this project was that participants were all undergraduate students. Although college students are not devoid of aggressive tendencies, there are likely many differences between this population and broader swaths of humanity. Future research can test whether our findings will replicate in more representative samples or in clinical or forensic groups characterized by heightened aggression. Further, the measurement invariance of the TAP across different cultures remains unknown. A valuable enterprise going forward will be to administer this measure across many cultures and test its psychometric invariance.

We also took a binary categorical approach to gender identity, which is wellknown to exist along a non-binary continuum. We were forced into this position by the available data, which only asked participants to report whether they identified as male or female. Future research must examine the invariance of the TAP across non-binary identities and by modeling female and male gender identities as continuous spectra. It may also prove valuable to investigate invariance across biological sex categories and examine whether findings with this variable show agreement with or diverge from participants' gender identities.

Additionally, we only assessed the sound blast version of the TAP, not other versions of the TAP that employ different modalities of aversive stimuli (e.g., shocks), due to the nature of the existing data we analyzed. Future research should thoroughly examine the psychometric properties of other aggression modalities, which should be subjected to the same methodological scrutiny.

Conclusions

Reducing aggression requires understanding and understanding aggression requires that it is accurately measured. The TAP is a premier aggression measure that enables experimenters to examine the personal and situational factors that make people more or less violent. Using a well-powered and preregistered approach, we demonstrated that the TAP exhibits measurement invariance across men and women and across experimentally provoked and unprovoked participants. We hope these results lend confidence to the use of the TAP to identify meaningful between-group differences in aggression that are not an artifact of poor psychometric properties. More broadly, we hope that investigators will continue to assess and improve the validity of laboratory aggression measures, in hopes that doing so will promote the ultimate reduction of harmful behaviors.

References

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, *8*(4), 291-322.

Anderson, C. A., & Bushman, B. J. (1997). External validity of “trivial” experiments: The case of laboratory aggression. *Review of General Psychology*, *1*(1), 19–41.

Arriaga, P., & Aguiar, C. (2019). Gender differences in aggression: The role of displaying facial emotional cues in a competitive situation. *Scandinavian Journal of Psychology*, *60*(5), 421-429.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*(4), 495-508.

Bettencourt, B., & Miller, N. (1996). Gender differences in aggression as a function of provocation: a meta-analysis. *Psychological Bulletin*, *119*(3), 422.

Bialosiewicz, S., Murphy, K., & Berry, T. (2013). An introduction to measurement invariance testing: Resource packet for participants. *American Evaluation Association*, 1-37.

Björkqvist, K. (2018). Gender differences in aggression. *Current Opinion in Psychology*, *19*, 39-42.

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.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114-140.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chester, D. S., & DeWall, C. N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology*, *112*(3), 413-430.

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.

Chester, D. S., Lynam, D. R., Milich, R., & DeWall, C. N. (2017). Social rejection magnifies impulsive behavior among individuals with greater negative urgency: An experimental test of urgency theory. *Journal of Experimental Psychology: General*, *146*(7), 962.

Chester, D. S., Merwin, L. M., & DeWall, C. N. (2015). Maladaptive perfectionism's link to aggression and self-harm: Emotion regulation as a mechanism. *Aggressive Behavior*, *41*(5), 443-454.

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications*. Academic Press.

Davidov, E., Muthen, B., & Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods & Research*, *47*(4), 631-636.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and*

Personality Science, *8*(4), 370-378.

Giancola, P. R., & Parrott, D. J. (2008). Further evidence for the validity of the Taylor Aggression Paradigm. *Aggressive Behavior*, *34*(2), 214–229.

Giancola, P. R., & Zeichner, A. (1995). Construct validity of a competitive reaction-time aggression paradigm. *Aggressive Behavior*, *21*(3), 199–204.

Hussey, I., & Hughes, S. (in press). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*.

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.

Hyatt, C., Zeichner, A., & Miller, J. (2019). Laboratory aggression and personality traits: A meta-analytic review. *Psychology of Violence*, *9*(6), 675-689.

King, A. R., & Russell, T. D. (2019). Lifetime Acts of Violence Assessment (LAVA) predictors of laboratory aggression. *Aggressive Behavior*, *45*(5), 477-488.

Lawrence, C., & Hutchinson, L. (2013). The influence of individual differences in sensitivity to provocations on provoked aggression. *Aggressive Behavior*, *39*(3), 212-221.

Marsh, H., Guo, J., Parker, P., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*(3), 524-545.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge

Pignatiello, G. A., Martin, R. J., & Hickman Jr, R. L. (2020). Decision fatigue: A conceptual analysis. *Journal of Health Psychology, 25*(1), 123-135.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90.

R Core Team. (2019). R: A language and environment for statistical computing [Software]. Available from http://www.R-project.org/

Raiche, G., & Magis, D. (2010). nFactors: An R package for parallel analysis and non-graphical solutions to the Cattell scree test (R Package Version 2.3.3) [Software]. Available from http://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Shaw, M., Cloos, L., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. Canadian Psychology.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46.

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.

Tedeschi, J. T., & Quigley, B. M. (1996). Limitations of laboratory paradigms for studying aggression. *Aggression and Violent Behavior*, *1*(2), 163–177.

Thissen, D., Steinberg, L., Millsap, R. E., & Maydeu-Olivares, A. (2009). *The Sage handbook of quantitative methods in psychology*. Sage.

West, S. J., Hyatt, C. S., Miller, J. D., & Chester, D. S. (in press). p-Curve analysis of the Taylor Aggression Paradigm: Estimating evidentiary value and statistical power across 50 years of research. *Aggressive Behavior*.

Zeichner, A., Parrott, D. J., & Frey, F. C. (2003). Gender differences in laboratory aggression under response choice conditions. *Aggressive Behavior*, *29*(2), 95-106.

Table S1

Unstandardized factor loadings for the single factor confirmatory factor analysis.

Taylor Aggression Paradigm Trial	Loading		Standard Error	
	Volume	Duration	Volume	Duration
1^	1.00	0.92	0.00	0.05
2*	1.64	1.67	0.07	0.07
3*	1.63	1.58	0.07	0.07
4*	1.61	1.58	0.07	0.07
5^	1.67	1.63	0.07	0.07
6	1.64	1.62	0.07	0.07
7	1.57	1.52	0.06	0.06
8*	1.52	1.44	0.06	0.06
9*	1.48	1.42	0.06	0.06
10*	1.41	1.34	0.06	0.06
11*	1.44	1.35	0.06	0.06
12*	1.41	1.30	0.06	0.06
13*	1.41	1.31	0.06	0.06
14	1.44	1.35	0.06	0.06
15	1.55	1.49	0.07	0.07
16	1.61	1.56	0.07	0.07
17	1.60	1.52	0.07	0.07
18*	1.65	1.55	0.07	0.07
19*	1.59	1.57	0.07	0.07
20*	1.58	1.52	0.07	0.07
21*	1.56	1.48	0.07	0.07
22*	1.57	1.54	0.07	0.07
23	1.56	1.44	0.07	0.06
24*	1.46	1.35	0.07	0.06
25*	1.41	1.36	0.07	0.07

**Indicates trials (volume and duration) retained*

^Indicates trials where only the volume was retained

Table S2

Factor loadings from the Exploratory Factor Analysis.

Taylor Aggression Paradigm Trial	Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
	Volume	Duration	Volume	Duration	Volume	Duration	Volume	Duration	Volume	Duration
1	0.28	0.29	0.22	0.20	0.47	0.43	0.15	0.10	0.07	0.09
2	0.14	0.13	0.18	0.16	0.78	0.78	0.21	0.21	0.17	0.16
3	0.24	0.24	0.24	0.22	0.72	0.70	0.15	0.11	0.12	0.09
4	0.28	0.26	0.25	0.24	0.67	0.65	0.12	0.09	0.16	0.14
5	0.40	0.40	0.32	0.28	0.60	0.58	0.06	0.06	0.09	0.07
6	0.49	0.50	0.33	0.30	0.50	0.47	0.07	0.07	0.07	0.08
7	0.56	0.55	0.34	0.29	0.38	0.38	0.08	0.05	0.08	0.08
8	0.57	0.58	0.33	0.32	0.32	0.27	0.13	0.08	0.03	0.03
9	0.65	0.62	0.22	0.25	0.25	0.21	0.18	0.14	0.07	0.08
10	0.60	0.61	0.22	0.22	0.16	0.13	0.25	0.19	0.13	0.10
11	0.61	0.60	0.16	0.17	0.22	0.19	0.27	0.19	0.13	0.11
12	0.61	0.60	0.12	0.12	0.21	0.17	0.29	0.22	0.14	0.17
13	0.57	0.54	0.13	0.15	0.25	0.17	0.25	0.21	0.17	0.17
14	0.46	0.47	0.19	0.20	0.26	0.19	0.24	0.19	0.34	0.32
15	0.40	0.42	0.30	0.30	0.27	0.22	0.19	0.17	0.41	0.42
16	0.26	0.30	0.43	0.38	0.30	0.25	0.15	0.15	0.50	0.50
17	0.18	0.19	0.53	0.51	0.28	0.22	0.10	0.09	0.43	0.47
18	0.23	0.21	0.59	0.60	0.29	0.21	0.08	0.09	0.30	0.27
19	0.21	0.21	0.63	0.62	0.25	0.23	0.11	0.14	0.21	0.19
20	0.17	0.18	0.69	0.67	0.24	0.21	0.13	0.15	0.14	0.11
21	0.24	0.23	0.67	0.68	0.22	0.18	0.21	0.20	0.05	0.03
22	0.29	0.32	0.59	0.56	0.23	0.20	0.33	0.32	0.05	0.02
23	0.34	0.31	0.46	0.47	0.20	0.18	0.47	0.45	0.04	-0.01
24	0.31	0.29	0.31	0.31	0.18	0.14	0.59	0.59	0.06	0.06
25	0.31	0.28	0.14	0.15	0.16	0.13	0.65	0.63	0.15	0.17

Table S3

Fit statistics and significance tests for the differential item functioning (DIF) analyses based on gender. Items flagged for potential DIF are indicated in bold.

	AIC		*BIC*		*X²(1)*		*p*	
TAP Trial	Volume	Duration	Volume	Duration	Volume	Duration	Volume	Duration
1	1.40	1.99	6.88	7.46	0.60	0.01	.440	.906
2	1.95	1.85	7.42	7.33	0.05	0.15	.819	.703
3	0.38	0.75	5.87	6.23	1.61	1.25	.205	.264
4	1.88	1.46	7.35	6.93	0.12	0.56	.725	.461
5	-0.81	1.99	4.66	7.46	2.81	0.02	.094	.904
6	0.16	**-5.49**	5.65	**-0.01**	1.83	**7.49**	.176	**.006**
7	1.77	1.51	7.24	6.98	0.23	0.49	.629	.483
8	1.81	0.55	7.28	6.03	0.20	1.45	.660	.229
9	0.91	1.79	6.38	7.27	1.09	0.21	.296	.650
10	0.94	0.70	6.42	6.17	1.05	1.30	.305	.254
11	0.07	1.82	5.54	7.29	1.93	0.18	.164	.672
12	1.62	1.74	7.10	7.21	0.38	0.26	.538	.611
13	1.11	-1.39	6.58	4.10	0.90	3.39	.345	.066
14	0.84	1.84	6.31	7.32	1.16	0.15	.281	.695
15	1.92	-0.62	7.39	4.86	0.08	2.62	.779	.106
16	1.65	0.92	7.12	6.40	0.35	1.08	.555	.300
17	1.55	1.89	7.02	7.36	0.45	0.11	.501	.738
18	-0.44	1.94	5.04	7.41	2.44	0.06	.119	.805
19	1.99	1.83	7.46	7.30	0.01	0.17	.914	.681
20	1.18	1.42	6.65	6.89	0.82	0.59	.364	.444
21	0.61	1.95	6.09	7.42	1.39	0.05	.239	.817
22	1.91	-0.15	7.38	5.33	0.10	2.15	.758	.143
23	1.94	0.71	7.41	6.18	0.06	1.29	.807	.256
24	1.68	1.92	7.16	7.39	0.32	0.08	.573	.776
25	1.83	0.29	7.30	5.76	0.17	1.72	.681	.191

Note. TAP = Taylor Aggression Paradigm, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

Table S4

Fit statistics and significance tests for the differential item functioning (DIF) analyses based on provocation condition. Items flagged for potential DIF are indicated in bold.

TAP Trial	*AIC*		*BIC*		*$X^2(1)$*		*p*	
	Volume	Duration	Volume	Duration	Volume	Duration	Volume	Duration
1	**-9.33**	**-10.53**	**-3.85**	**-5.05**	**11.33**	**12.53**	**.00**	**.00**
2	**-9.76**	**-8.19**	**-4.29**	**-2.72**	**11.76**	**10.19**	**.00**	**.00**
3	**-3.32**	**-3.93**	**2.15**	**1.54**	**5.32**	**5.93**	**.02**	**.01**
4	1.08	1.67	6.56	7.14	0.92	0.33	.34	.57
5	-1.99	**-4.11**	3.48	**1.36**	3.99	**6.11**	.05	**.01**
6	**-11.87**	**-14.95**	**-6.40**	**-9.48**	**13.87**	**16.95**	**.00**	**.00**
7	**-10.49**	**-11.57**	**-5.02**	**-6.10**	**12.49**	**13.57**	**.00**	**.00**
8	**-4.16**	-0.30	**1.31**	5.17	**6.16**	2.30	**.01**	.13
9	**-4.45**	**-16.22**	**1.03**	**-10.74**	**6.45**	**18.22**	**.01**	**.00**
10	**-19.54**	**-11.64**	**-14.06**	**-6.17**	**21.54**	**13.64**	**.00**	**.00**
11	**-4.40**	**-13.82**	**1.07**	**-8.35**	**6.40**	**15.82**	**.01**	**.00**
12	**-2.98**	**-8.17**	**2.49**	**-2.69**	**4.98**	**10.17**	**.03**	**.00**
13	**-2.70**	**-5.71**	**2.77**	**-0.24**	**4.70**	**7.71**	**.03**	**.01**
14	**-5.17**	**-7.11**	**0.31**	**-1.64**	**7.17**	**9.11**	**.01**	**.00**
15	**-10.99**	**-15.15**	**-5.52**	**-9.68**	**12.99**	**17.15**	**.00**	**.00**
16	**-9.42**	**-13.43**	**-3.94**	**-7.96**	**11.42**	**15.43**	**.00**	**.00**
17	**-8.55**	**-14.22**	**-3.08**	**-8.74**	**10.55**	**16.22**	**.00**	**.00**
18	**-9.46**	**-19.71**	**-3.99**	**-14.24**	**11.46**	**21.71**	**.00**	**.00**
19	**-9.18**	**-10.67**	**-3.70**	**-5.20**	**11.18**	**12.67**	**.00**	**.00**
20	**-20.07**	**-13.37**	**-14.59**	**-7.89**	**22.07**	**15.37**	**.00**	**.00**
21	**-12.19**	**-14.54**	**-6.72**	**-9.07**	**14.19**	**16.54**	**.00**	**.00**
22	**-9.35**	**-16.86**	**-3.88**	**-11.38**	**11.35**	**18.86**	**.00**	**.00**
23	**-16.16**	**-14.99**	**-10.69**	**-9.52**	**18.16**	**16.99**	**.00**	**.00**
24	**-15.51**	**-18.90**	**-10.04**	**-13.42**	**17.51**	**20.90**	**.00**	**.00**
25	**-13.87**	**-6.87**	**-8.39**	**-1.40**	**15.87**	**8.87**	**.00**	**.00**

Note. TAP = Taylor Aggression Paradigm, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.