Journal of Experimental Social Psychology 87 (2020) 103905

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores

Judith Koppehele-Gossel[a,*], Lisa Hoffmann[a], Rainer Banse[a], Bertram Gawronski[b]

[a] Department of Psychology, University of Bonn, Germany
[b] Department of Psychology, University of Texas at Austin, TX, United States of America

ARTICLE INFO

☆This paper has been recommended for acceptance by Nicholas Rule

Keywords:
Evaluative priming
Implicit measures
Reliability
Response latencies
Psychometrics

ABSTRACT

Evaluative priming is based on the notion that evaluative classifications of target stimuli are faster (vs. slower) when they are preceded by a prime stimulus of the same (vs. opposite) valence. Although evaluative priming is widely used as an implicit measure of evaluation, there is no common procedure for the treatment of response-latency outliers. To address this limitation, four studies examined common outlier-treatments in terms of (1) the overall size of evaluative priming effects, (2) their internal consistency, and (3) their relation to corresponding explicit measures in the domains of conditioned attitudes (Study 1), political attitudes (Study 2), racial attitudes (Study 3), and ethnic attitudes (Study 4). The algorithm with the best performance uses a priori cutoffs of 300 ms and 1000 ms, treating response times beyond these cutoffs as missing values. Internal consistency was low for all algorithms, indicating limits in the usefulness of evaluative priming for correlational studies.

1. Introduction

The emergence of implicit measures in the mid-1990s had a fundamental impact on psychological science. Different from the traditional reliance on explicit self-reports, implicit measures are based on the idea that psychological attributes (e.g., attitudes, stereotypes, self-esteem, self-concept) can be inferred from people's speed and accuracy in responding to stimuli on highly controlled experimental tasks (for a review, see Gawronski & De Houwer, 2014). The most prominent example is the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998), which has been cited >10000 times in the two decades since its seminal publication. However, despite its popularity, the IAT has also been the target of abundant criticism (for an overview, see Teige-Mocigemba, Klauer, & Sherman, 2010). Although some of this criticism poses a challenge to implicit measures in general (e.g., Arkes & Tetlock, 2004; Blanton & Jaccard, 2006; Gawronski, Morrison, Phills, & Galdi, 2017), many objections against the IAT are task-specific (e.g., Meissner & Rothermund, 2013; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009; Rothermund & Wentura, 2004; Teige-Mocigemba, Klauer, & Rothermund, 2008), highlighting the importance of alternative instruments that do not suffer from the same limitations.

One such alternative is the evaluative priming task (EPT; Fazio, Jackson, Dunton, & Williams, 1995), which has been developed more than a decade before the IAT to investigate the automatic activation of attitudes (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). On a typical trial of the EPT, participants are briefly presented with an attitude object as a prime stimulus, which is followed by a positive or negative word as a target stimulus. Participants' task is to indicate as quickly as possible whether the target word is positive or negative by pressing one of two designated keys. The basic idea underlying the EPT is that quick and accurate responses to the target words should be facilitated when the prime stimulus elicits an evaluative response that is congruent with the valence of the target words. In contrast, quick and accurate responses to the target words should be impaired when the prime stimulus elicits an evaluative response that is incongruent with the valence of the target words (for a meta-analysis, see Herring et al., 2013).

Although the task structure of the EPT resolves many instrument-specific concerns about the IAT, a major limitation is that there is still no consensually accepted procedure for the treatment of outliers in the distributions of response latencies obtained with the EPT (i.e., target responses that are too fast or too slow to capture meaningful influences of the primes). Thus, different from the common use of the same established algorithm in research using the IAT (Greenwald, Nosek, & Banaji, 2003), researchers using the EPT rely on a wide range of procedures to maximize “signal” and minimize “noise” in evaluative priming data. This practice is problematic for at least three reasons. First, different outlier-treatments in research with the EPT can make it difficult to compare findings across studies, which can hamper progress

* Corresponding author at: Department of Psychology, University of Bonn, Kaiser-Karl-Ring 9, 53111 Bonn, Germany.
E-mail address: judith.koppehele@gmail.com (J. Koppehele-Gossel).

https://doi.org/10.1016/j.jesp.2019.103905
Received 26 March 2019; Received in revised form 25 September 2019; Accepted 10 October 2019

in terms of a cumulative science (see Herring et al., 2013). Second, flexibility in the use of outlier-treatments can increase the likelihood of false positives when researchers choose their preferred algorithm in a post-hoc fashion based on whether it produces a predicted outcome (see Simmons, Nelson, & Simonsohn, 2011). Third, although arbitrary post-hoc choices of outlier-treatments can be overcome through consistent use of the same procedure or pre-registration, commitment to a particular algorithm can lead to false negatives if the chosen outlier-treatment has suboptimal psychometric properties (see Fiedler, Kutzner, & Krueger, 2012).

The current work aimed to address these issues by examining the psychometric properties of different outlier-treatments in research with the EPT. The overarching goal was to identify the algorithm with the best psychometric properties, so that it could serve as a standard procedure for future research with the EPT and potential reanalyses of existing data. Toward this end, we compared the ten most frequently used algorithms for outlier-treatments of EPT data in terms of (1) the overall size of evaluative priming effects, (2) their internal consistency, and (3) their relation to corresponding explicit measures. The psychometric properties of the algorithms were analyzed in four studies that investigated evaluative priming effects in the areas of conditioned attitudes (Study 1), political attitudes (Study 2), racial attitudes (Study 3), and ethnic attitudes (Study 4).

1.1. Outlier-treatments

The primary reason why evaluative priming data require pre-processing of outliers is that participants' responses to the targets are sometimes too slow to capture meaningful influences of the primes. First, unusually slow responses can occur when participants do not pay attention to the task, which can negatively affect the psychometric properties of evaluative priming scores. Second, the overall size of evaluative priming effects has been shown to depend on sufficiently short intervals, in that priming effects decrease with increasing delays between prime exposure and responses to the targets (e.g., Hermans, De Houwer, & Eelen, 2001). In addition to the challenges posed by outliers at the upper end of the distribution, responses are sometimes too fast to capture meaningful responses to the targets (e.g., when responses are implausibly fast for a correct identification of the target stimulus and the implementation of the required response). To address these issues, researchers have used different procedures to maximize “signal” and minimize “noise” in the data obtained with the EPT. A shared feature of these procedures is that all of them eliminate response latencies from trials with incorrect responses. In addition, most researchers exclude participants with excessive error rates in their responses to the targets (e.g., participants with error rates outside the continuous error distribution of the sample). Yet, extant algorithms differ in terms of the subsequent treatment of outliers in the distribution of response latencies.

In the current research, we screened the EPT literature for common outlier procedures and investigated the psychometric properties of the ten most frequently used algorithms.[1] In a first step, we used Google Scholar to identify all publications citing Fazio et al.'s (1995) seminal article on the EPT, which resulted in a total of 2957 hits (June 4, 2019). In a second step, the identified publications were screened for the following criteria: (1) published in English; (2) published in a peer-reviewed journal; (3) reports at least one empirical study; (4) reported study included EPT; (5) priming task used words as target stimuli; (6) target stimuli had to be categorized in terms of valence; (7) reaction times served as dependent variable; and (8) responses were recorded via button presses. These criteria resulted in 150 articles comprising 223 individual studies (see https://osf.io/hjm4z/ for details on publications and studies). Of the ten most frequently used procedures in the identified set of studies, one excludes only error trials without further treatment of response latency data; six use a priori cut-off values to eliminate outliers; and three identify outliers based on the actual distribution of response latencies. An overview of the ten procedures is provided in Table 1. In the current studies, we compared these algorithms in terms of (1) the overall size of evaluative priming effects, (2) the internal consistency of evaluative priming scores, and (3) the relation of evaluative priming scores to corresponding explicit measures.

1.2. Reliability estimation

To estimate the reliability of the evaluative priming scores obtained with the identified outlier-treatments, we calculated Cronbach's Alpha values of internal consistency using two parceling procedures. Both procedures divide the full data set into two subsets of equal size. The first procedure divides the trials into two subsets based on whether they were part of the first or the second half of the task (two-block). The second procedure divides the trials into two subsets based on whether they had an odd or an even position number in the overall sequence of trials (odd-even). A disadvantage of the two-block split is that reliability estimates could be artificially suppressed if the reliability of evaluative priming scores changes over the course of the task (e.g., when evaluative priming scores are highly reliable in the first half, but entirely unreliable in the second half). In this case, numerical scores of internal consistency could potentially underestimate the true reliability of evaluative priming scores. This disadvantage is addressed in the odd-even split, because changes over the course of the task should have the same effect on odd and even trials. However, a disadvantage of the odd-even split is that potential changes in the reliability of evaluative priming scores over the course of the task cannot be detected. For these reasons, we used both two-block and odd-even splits to estimate the reliability of the priming scores obtained with the different outlier-treatments. To the extent that the reliability of evaluative priming scores remains constant over the course of the task, the two parceling procedures should produce similar estimates of internal consistency. In contrast, if the reliability of evaluative priming scores changes over the course of the task (e.g., because of fatigue), estimates of internal consistency should be higher for odd-even splits than for two-block splits.

1.3. Sample size and statistical power

For each study, we aimed to recruit 100 participants, which provides a power of 80% in detecting a small evaluative priming effect of $d = 0.28$ and a medium-size correlation of $r = .27$ between evaluative priming scores and corresponding explicit measures. The power estimates for both analyses are in line with meta-analytic effect sizes of evaluative priming effects (Herring et al., 2013) and meta-analytic correlations between implicit and explicit measures (e.g., Cameron, Brown-Iannuzzi, & Payne, 2012; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). In Study 2, we had the opportunity to collect data from a larger sample of 400 participants, which provides a power of 80% in detecting a small evaluative priming effect of $d = 0.14$ and a small correlation of $r = .14$ between evaluative priming scores and corresponding explicit measures. The data for each study were collected in one shot without intermittent statistical analyses. We report all measures, all conditions, and all data exclusions. The materials, raw data, and analysis files for all studies are publicly available at https://osf.io/hjm4z/.

2. Study 1: conditioned attitudes

Evaluative conditioning (EC) is defined as the change in the evaluation of a conditioned stimulus (CS) due its pairing with a positive or

[1] Our search revealed a total of 48 algorithms, but a substantial portion of them were used in only one or two articles or in a very small number of studies. For these reasons, we decided to focus on the ten most frequently used algorithms.

Table 1
Overview of the ten most frequently used procedures for the treatment of reaction time outliers in evaluative priming data.

Algorithm	Description	N_{total}	$n_{logtrans}$
Errors only	Exclude error trials without further treatment of response latency data	32	6
300–1000 ms	Exclude error trials and reaction times lower than 300 ms and higher than 1000 ms	17	2
300–1500 ms	Exclude error trials and reaction times lower than 300 ms and higher than 1500 ms	12	2
0–800 ms	Exclude error trials and reaction times higher than 800 ms	10	0
250 ms – 3 *SD*	Exclude error trials and reaction times lower than 250 ms and higher than 3 standard deviations of mean reaction time of a given participant	10	4
0–1500 ms	Exclude error trials and reaction times higher than 1500 ms	9	6
250–1500 ms	Exclude error trials and reaction times lower than 250 ms and higher than 1500 ms	9	0
300–3000 ms	Exclude error trials and reaction times lower than 300 ms and higher than 3000 ms	8	6
300 ms – 2 *SD*	Exclude error trials and reaction times lower than 300 ms and higher than 2 standard deviations of mean reaction time of a given participant	7	7
±2 SD	Exclude error trials and reaction times lower than 2 standard deviations and higher than 2 standard deviations of mean reaction time of a given participant	7	1

Note. N_{total} refers to the total number of individual studies with a particular algorithm. $n_{logtrans}$ refers to the number of studies applying an additional log-transformation to the data.

negative unconditioned stimulus (US) (De Houwer, 2007; Gast, Gawronski, & De Houwer, 2012). Using an EC procedure that has proven its effectiveness in producing significant EC effects on evaluative priming scores (e.g., Gawronski, Balas, & Creighton, 2014; Gawronski & Mitchell, 2014; Gawronski, Mitchell, & Balas, 2015; Hu, Gawronski, & Balas, 2017a, 2017b), Study 1 investigated the impact of the ten outlier-treatments on the detection of significant EC effects with evaluative priming. Assuming little systematic variation in EC effects across individuals (cf. Vogel, Hütter, & Gebauer, 2019), Study 1 focused primarily on the overall size of EC effects, putting less emphasis on the internal consistency of evaluative priming scores and their correlation with self-reported CS evaluations. To the extent that all participants are influenced by the EC procedure to the same extent, any variation in measurement scores across individuals would be measurement error. As a result, correlations between two parcels of EC effects on evaluative priming (i.e., internal consistency) should be rather low even when the two parcels reliably capture experimental effects of CS-US pairings on evaluative responses to the CS. Moreover, correlations between EC effects on evaluative priming and EC effects on self-reported evaluations should be rather low even when both measures reliably capture experimental effects of CS-US pairings on evaluative responses to the CS. For these reasons, Study 1 used the overall size of EC effects as the primary criterion for the evaluation of outlier-treatments, putting less emphasis on internal consistency and correlations with self-reported evaluations.

2.1. Methods

2.1.1. Participants

One-hundred psychology undergraduates (73 female, 26 male, 1 unknown) at the University of Texas at Austin participated in the study for course credit. Due to a computer malfunction, data from two participants were incomplete. One additional participant showed an error rate on the EPT that was outside the error distribution of the sample (93%). Data from these three participants were excluded from the analyses, leaving us with a final sample of 97 participants. The study was part of a one-hour battery that included one unrelated study in addition to the current one. The current study was always administered as the second one in the battery. The study was approved by the Institutional Review Board of the University of Texas at Austin (IRB #2014-06-0078).

2.1.2. Materials

As CSs, we adapted ten computer-generated images of shapes with different color patterns from Gawronski and Mitchell (2014). Four of these images were paired with a positive picture as the US; four were paired with a negative picture as the US; and two images were not paired with a valenced picture to serve as neutral baseline primes in the EPT. As USs, we used four positive and four negative pictures from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2008). The positive USs showed a girl (Image 2035; mean valence rating = 7.52), sky divers (Image 5621; mean valence rating = 7.57), nature (Image 5760; mean valence rating = 8.05), and a rollercoaster (Image 8492; mean valence rating = 7.21); the negative USs showed an elderly woman (Image 2590; mean valence rating = 3.26), a snake (Image 1050; mean valence rating = 3.46), a cemetery (Image 9001; mean valence rating = 3.10), and an aimed gun (Image 6230; mean valence rating = 2.37).

2.1.3. EC procedure

The EC procedure was adapted from earlier research by Gawronski and colleagues, demonstrating the effectiveness of the procedure in producing significant EC effects on evaluative priming measures (e.g., Gawronski et al., 2014, 2015; Gawronski & Mitchell, 2014; Hu et al., 2017a, 2017b). The procedure included ten presentations of each of the eight CS-US pairs, summing up to a total of 80 trials. Each trial started with a fixation cross that was displayed for 250 ms in the center of the screen. The fixation cross was followed by the CS for 1000 ms, which was replaced by the US for 1000 ms. The inter-trial interval was 1500 ms. The images used as CSs were displayed in a size of 2.00 × 1.43 in.; the pictures used as USs were displayed in a size of 14.22 × 10.67 in. Each CS was always presented with the same US. The particular pairings of CSs and USs were counterbalanced by means of a Latin square. Participants received the following instructions for the EC procedure:

> *The current study investigates visual perception. For this purpose, you will be presented with different kinds of images. Some of the images will be computer-generated drawings; other images will be photographs. Your task is to pay close attention to these images. We will later ask you a number of questions about the images that you have seen. The visual perception task will take approximately 5 min. Please pay close attention to the images throughout the entire task. When you are ready to start, please click "continue."*

2.1.4. Evaluative priming task

After the EC task, participants completed an EPT that included the CSs and the two baseline stimuli as primes and ten positive and ten negative adjectives as targets. The procedural details of the EPT were adapted from earlier EC research by Gawronski and colleagues (e.g., Gawronski et al., 2014, 2015; Gawronski & Mitchell, 2014; Hu et al., 2017a, 2017b). The positive target words were: *pleasant*, *good*, *outstanding*, *beautiful*, *magnificent*, *marvelous*, *excellent*, *appealing*, *delightful*, *nice*; the negative target words were: *unpleasant*, *bad*, *horrible*, *miserable*, *hideous*, *dreadful*, *painful*, *repulsive*, *awful*, *ugly*. Each trial started with a fixation cross that was displayed for 500 ms in the center of the screen. The fixation cross was followed by a prime stimulus, which was replaced by the target word after 200 ms. Participants' task was to press a

right-hand key (*Numpad 5*) as quickly as possible when the target word was positive and a left-hand key (*A*) when the target word was negative. The target words remained on the screen until participants made their response. Incorrect responses were followed by the word *ERROR*! for 1500 ms. The inter-trial interval was 500 ms. Each of the eight CSs and each of the two baseline primes was presented once with each of the ten positive target words and once with each of the ten negative words, summing up to a total of 200 trials. The order of trials was randomized individually for each participant. Participants received the following instructions for the EPT:

> *The following component of this study is a concentration test. For this purpose, you will be presented with positive and negative words. Additionally, you will be presented with pictures that briefly appear before the words are presented. Your main task is to indicate as fast as possible whether the word on the screen is a positive or a negative word. Please press the "A" key when you see a negative word on the screen, and please press the "5" key of the number block when you see a positive word. In order to facilitate faster responses, please keep your main left-hand finger on the A key on the left side of the keyboard, and your main right-hand finger on the 5 key of the number block on the right side of the keyboard. Please try to respond as quickly as possible without making too many errors. Again, please press the "A" key when you see a negative word on the screen, and please press the "5" key when you see a positive word.*

2.1.5. Evaluative rating task

In addition to completing the evaluative priming measure of CS evaluations, participants were asked to rate how pleasant or unpleasant each CS made them feel on 7-point scales ranging from 1 (*very unpleasant*) to 7 (*very pleasant*). Order of the CSs was randomized individually by the computer for each participant. The order of the EPT and the evaluative rating task was counterbalanced across participants.

2.2. Results

Overall, participants showed the correct response on 96% of the trials in the EPT (range: 83% to 100%). Error rates did not differ between the first (3.8%) and the second half (3.9%) of the task, $t(96) = 0.26$, $p = .795$, $d = 0.03$ (see Table 2). Latencies from trials with incorrect responses were excluded from the aggregation of evaluative priming scores. For each of the ten outlier-treatments, we calculated a priming score for responses to positive target words by subtracting the mean reaction time for positive target words preceded by CSs paired with positive USs from the mean reaction time for positive target words preceded by CSs paired with negative USs. Higher scores on this index reflect greater positivity toward CSs paired with positive USs compared to CSs paired with negative USs (see Wentura & Degner, 2010; Wittenbrink, 2007). Correspondingly, a priming score for responses to negative target words was calculated by subtracting the mean reaction time for negative target words preceded by CSs paired with positive USs from the mean reaction time for negative target words preceded by CSs paired with negative USs. Higher scores on this index reflect greater negativity toward CSs paired with positive USs compared to CSs paired with positive USs (see Wentura & Degner, 2010; Wittenbrink, 2007). The latter index was then subtracted from the former index to obtain a priming index of EC effects: Priming Index = [RT(positive targets | negative CS) - RT(positive targets | positive CS)] - [RT(negative targets | negative CS) - RT(negative targets | positive CS)]. Higher scores on this index reflect a greater preference for CSs paired with positive USs over CSs paired with negative USs, with a value of zero serving as a neutral reference point of equal evaluations of the two kinds of CSs.

2.2.1. Overall priming effect

Table 3 shows the means and standard deviations of the evaluative priming indices for the ten outlier-treatments. Table 4 shows the results of one-sample *t*-tests comparing the overall priming indices to the neutral reference point of zero. Statistically significant EC effects were obtained for the 0–800 ms cutoff and the 300–1000 ms cutoff. The size of EC effects was similar for the two outlier-treatments with Cohen's *d*s of 0.26 and 0.27, respectively. Although the other outlier-treatments showed effects in the expected direction, EC effects were considerably smaller and did not reach statistical significance.

2.2.2. Internal consistency

Estimates of internal consistency for odd-even and two-block splits for the ten outlier-treatments are presented in Table 5. Cronbach's Alpha values were very low or negative for both parceling procedures regardless of the outlier-treatment. Overall, there was no evidence that estimates of internal consistency were lower for the two-block split compared to the odd-even split, which speaks against the possibility that the reliability of evaluative priming scores might change over the course of the task.

2.2.3. Relation to explicit measure

An index of EC effects on self-reported evaluations was calculated by subtracting the mean ratings of CSs paired with negative USs from the mean ratings of CSs paired with positive USs. Higher scores on this index reflect a greater preference for CSs paired with positive USs over CSs paired with negative USs, with a value of zero serving as a neutral reference point of equal evaluations of the two kinds of CSs. A one-sample *t*-test revealed a significant EC effect with a difference score significantly greater than zero ($M = 1.22$, $SD = 1.45$), $t(96) = 8.31$, $p < .001$, $d = 0.84$. EC effects on evaluative priming scores showed a significant positive relation with EC effects on self-reported evaluations for the 250 ms-3 *SD* cutoff (see Table 6). The correlations for the other

Table 2
Percentage of excluded trials as a function of outlier-treatment, Studies 1-4.

	Study 1			Study 2			Study 3			Study 4		
	%	Min	Max	%	Min	Max	%	Min	Max	%	Min	Max
Errors 1st half	3.82	0	29	8.43	0	58	5.97	0	31	3.74	0	14
Errors 2nd half	3.92	0	23	4.99	0	31	5.65	0	21	4.99	0	20
0–800 ms	19.11	2	62	16.22	2	71	13.05	1	45	18.70	3	91
0–1500 ms	5.52	0	24	7.71	0	42	6.49	0	20	6.56	1	23
250–1500 ms	5.63	0	30	8.02	0	44	6.65	0	22	6.58	1	23
250 ms – 3 *SD*	5.74	1	27	8.52	1	43	7.56	2	22	6.25	1	19
300–1000 ms	10.87	1	45	11.30	1	56	8.91	0	35	11.79	2	69
300–1500 ms	5.65	0	31	8.23	0	46	6.79	0	23	6.61	1	23
300–3000 ms	4.24	0	27	7.37	0	44	6.27	0	22	4.68	0	18
300 ms – 2 *SD*	7.92	2	29	10.90	2	44	9.81	4	25	8.45	1	21
±2 *SD*	7.81	2	19	10.64	2	37	9.76	4	22	8.43	1	20

Note. Percentages of excluded trials refer to combined exclusions of error trials and response latency outliers. Minimum and maximum values are rounded to whole numbers.

Table 3
Evaluative priming effects as a function of outlier-treatment, Studies 1–4.

	Study 1		Study 2		Study 3		Study 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Errors only	14.42	75.02	25.01	130.60	13.36	116.81	21.60	234.89
0–800 ms	7.59	29.38	18.86	39.69	9.20	28.24	8.32	19.61
0–1500 ms	7.14	52.27	28.63	58.53	3.42	49.72	5.97	44.06
250–1500 ms	7.35	52.02	28.00	56.89	3.27	49.33	5.97	44.28
250 ms – 3 *SD*	6.93	54.84	26.59	57.98	7.16	66.25	0.57	63.84
300–1000 ms	9.30	34.94	21.56	45.63	8.05	36.14	10.34	27.77
300–1500 ms	7.69	51.86	27.70	56.62	2.92	49.51	5.98	44.28
300–3000 ms	13.37	66.25	28.35	71.67	−0.34	62.60	6.89	56.11
300 ms – 2 *SD*	4.57	48.43	24.08	53.00	6.16	59.84	1.03	44.68
±2 *SD*	3.85	48.80	24.45	54.57	6.19	59.30	0.99	44.55

Note. *M* = mean. *SD* = standard deviation. Errors only = only error trials are excluded.

Table 4
Results of one-sample t-tests comparing evaluative priming effects against zero as a function of outlier-treatment, Studies 1–4.

	Study 1			Study 2			Study 3			Study 4		
	t(96)	*p*	*d*	*t*(405)	*p*	*d*	*t*(102)	*p*	*d*	*t*(109)	*p*	*d*
Errors only	1.89	.061	0.19	**3.86**	**<.001**	**0.19**	1.16	.249	0.11	0.97	.337	0.09
0–800 ms	**2.55**	**.013**	**0.26**	**9.58**	**<.001**	**0.48**	**3.31**	**.001**	**0.33**	**4.45**	**<.001**	**0.42**
0–1500 ms	1.35	.182	0.14	**9.86**	**<.001**	**0.49**	0.70	.487	0.07	1.42	.158	0.14
250–1500 ms	1.39	.167	0.14	**9.92**	**<.001**	**0.49**	0.67	.502	0.07	1.41	.160	0.13
250 ms – 3 *SD*	1.25	.216	0.13	**9.24**	**<.001**	**0.46**	1.10	.275	0.11	0.09	.925	0.01
300–1000 ms	**2.62**	**.010**	**0.27**	**9.52**	**<.001**	**0.47**	**2.26**	**.026**	**0.22**	**3.91**	**<.001**	**0.37**
300–1500 ms	1.46	.147	0.15	**9.86**	**<.001**	**0.49**	0.60	.551	0.06	1.42	.159	0.14
300–3000 ms	1.99	.050	0.20	**7.97**	**<.001**	**0.40**	−0.06	.956	−0.01	1.29	.200	0.12
300 ms – 2 *SD*	0.93	.355	0.09	**9.16**	**<.001**	**0.45**	1.05	.298	0.10	0.24	.810	0.02
±2 *SD*	0.78	.439	0.08	**9.03**	**<.001**	**0.45**	1.06	.292	0.10	0.23	.816	0.02

Note. *d* = Cohen's *d*, derived from the division of *t*-statistic by the root of *N*. Errors only = only error trials are excluded. *SD* = standard deviation. Statistically significant priming effects are highlighted in bold font.

outlier-treatments were close to zero, the only exception being the Errors only and the 300–3000 ms algorithm which produced marginally significant positive correlations.

2.3. Discussion

Using the overall size of EC effects on evaluative priming scores as the primary evaluation criterion for the ten outlier-treatments, the results of Study 1 provide positive evidence for the 0–800 ms cutoff and the 300–1000 ms cutoff. Both outlier-treatments produced significant EC effects with roughly similar effect sizes. No significant EC effects were obtained with the other eight outlier-treatments. Consistent with the assumption that there is little systematic variation in EC effects across individuals, internal consistencies were close to zero for all outlier-treatments. However, despite their low internal consistency, EC effects on evaluative priming scores and EC effects on self-reported evaluations were significantly correlated for the 250 ms-3 *SD* algorithm. The latter finding seems somewhat surprising, given that this outlier-treatment showed negatively signed internal consistencies close to zero.[2] Although the obtained correlation with self-reported evaluations may reflect meaningful individual differences in EC effects (Vogel et al., 2019), such a conclusion seems at odds with the finding that two parcels of EC effects on evaluative priming were unrelated regardless of the parceling procedure. Thus, even if there are meaningful individual differences in EC effects (Vogel et al., 2019), the low internal consistencies of EC effects suggest that evaluative priming does not capture these differences in a reliable manner. This contradiction implies the possibility that the obtained correlation between EC effects on evaluative priming scores and EC effects on self-reported evaluations are false positives (e.g., driven by outliers) instead of reflecting genuine individual differences in EC effects. Study 2 aimed to address this ambiguity by comparing the impact of the ten outlier-treatments on evaluative priming scores in a domain with large individual differences: political attitudes.

3. Study 2: political attitudes

Study 2 investigated the impact of the ten outlier-treatments on the detection of individual differences in political attitudes. Toward this end, American participants completed an EPT designed to measure evaluations of Hillary Clinton and Donald Trump shortly after the 2016 U.S. Presidential Election. Different from the emphasis on the overall size of EC effects on evaluative priming scores in Study 1, there was no a priori reason to expect a significant evaluative priming effect in Study 2. To the extent that preferences for Hillary Clinton versus Donald Trump are symmetrically distributed around the neutral midpoint, the overall size of evaluative priming effects becomes rather meaningless as a criterion for the evaluation of the outlier-treatments. Yet, given that people systematically differ in their political attitudes, the ten outlier-treatments can be evaluated in terms of their ability to capture individual differences in political preferences. The latter aspect should be reflected in high internal consistencies of evaluative priming scores reflecting preferences for one over the other candidate. Moreover, given

[2] Negative reliability estimates likely result from reversed priming scores in one the two test halves. Because priming scores are computed identically for both test halves, the negative estimates cannot result from wrong "item coding" (as it is usually the case for questionnaire items). Because priming scores are based on reaction-time differences, it is possible that all true variance is discarded by the difference computation when the true effects are very similar: A – B = (true variance score A + error A) – (true variance score B + error B). If the result of this difference mainly represents errorA and errorB, it is possible that the difference results in low (or even negative) reliability estimates.

Table 5
Cronbach's alpha values as a function of outlier-treatment and parceling procedure, Studies 1–4.

		Outlier procedure									
Study (*N*)	reliability estimate	Errors only	0 - 800 ms	0 - 1500 ms	250 – 1500 ms	250 ms - 3 *SD*	300 - 1000 ms	300 - 1500 ms	300 - 3000 ms	300 ms - 2 *SD*	±2 *SD*
1 (97)	Two-block	.11	−.07	.10	.10	−.08	−.29	.10	<.01	.11	.11
		[−.24, .46]	[−.50, .35]	[−.26, .46]	[−.26, .46]	[−.50, .34]	[−.80, .22]	[−.25, .46]	[−.39, .40]	[−.24, .46]	[−.24, .46]
	Odd-even	−.34	−.37	−.11	.03	−.28	−.03	.02	−.34	−.13	−.21
		[−.79, .10]	[−.91, .17]	[−.55, .33]	[−.35, .41]	[−.76, .20]	[−.44, .38]	[−.36, .40]	[−.82, .14]	[−.54, .28]	[−.67, .24]
2 (406)	Two-block	.25	.50	.37	.31	.26	.39	.31	.33	.35	.38
		[.14, .36]	[.41, .60]	[.25, .49]	[.18, .44]	[.12, .40]	[.27, .51]	[.17, .44]	[.20, .46]	[.22, .47]	[.26, .50]
	Odd-even	.15	.47	.44	.41	.23	.42	.39	.29	.33	.38
		[.01, .29]	[.36, .57]	[.33, .55]	[.30, .53]	[.08, .38]	[.31, .53]	[.27, .51]	[.15, .43]	[.20, .46]	[.26, .50]
3 (103)	Two-block	.45	.07	.19	.23	.38	.16	.26	.33	.49	.42
		[.26, .64]	[−.30, .43]	[−.11, .49]	[−.05, .51]	[.18, .59]	[−.16, .48]	[−.02, .53]	[.09, .57]	[.30, .68]	[.20, .63]
	Odd-even	−.32	.20	.45	.47	.64	.39	.50	.43	.78	.75
		[−.68, .05]	[−.11, .51]	[.24, .66]	[.26, .67]	[.51, .78]	[.15, .62]	[.30, .69]	[.22, .65]	[.69, .86]	[.66, .85]
4 (110)	Two-block (all trials)	−.09	.07	.50	.46	.33	.26	.47	.37	.31	.31
		[−.28, .10]	[−.28, .41]	[.32, .69]	[.26, .66]	[.13, .54]	[−.02, .54]	[.27, .66]	[.14, .60]	[.06, .56]	[.06, .56]
	Odd-even (all trials)	.06	−.24	.17	.19	.65	−.10	.19	.20	.33	.33
		[−.06, .18]	[−.70, .22]	[−.14, .48]	[−.11, .49]	[.52, .78]	[−.50, .31]	[−.11, .49]	[−.10, .50]	[.09, .57]	[.09, .57]
	Two-block (1 − 200)	.52	−.32	.21	.21	.44	.03	.21	−.35	.11	.11
		[.35, .70]	[−.81, .18]	[−.09, .50]	[−.09, .50]	[.23, .64]	[−.33, .39]	[−.09, .50]	[−.85, .16]	[−.22, .43]	[−.22, .43]
	Odd-even (1–200)	.25	−.64	.27	.27	.64	−.25	.27	−.08	.05	.05
		[−.03, .53]	[−1.26, −.03]	[.00, .54]	[.00, .54]	[.50, .77]	[−.70, .21]	[.00, .54]	[−.49, .32]	[−.27, .38]	[−.28, .38]

Notes. For negative values, the determinant of the covariance matrix was described as zero or close to zero. For these values, statistics based on the inverted matrix cannot be calculated and their values are displayed as system-defined missing values. The values in squared brackets represent the bootstrapped (n = 1000) upper and lower 95%-confidence boundaries of the respective reliability estimate.

Table 6
Correlations between evaluative priming scores and explicit measures as a function of outlier-treatments, Studies 1–4.

Explicit measure	Study 1		Study 2		Study 3		Study 4			
	Evaluative conditioning effects on explicit evaluations		Explicit preference for Clinton over Trump		Explicit preference for Whites over Blacks		Motivation to act without prejudice		Subtle and blatant prejudice against Turks	
	r(95)	*p*	*r*(404)	*p*	*r*(101)	*p*	*r*(108)	*p*	*r*(108)	*p*
Errors only	.19	.069	**.21**	**<.001**	.07	.482	<.01	.963	.18	.054
0–800 ms	.06	.555	**.48**	**<.001**	.09	.368	−.03	.752	.06	.568
0–1500 ms	.07	.522	**.41**	**<.001**	.04	.698	−.01	.937	.10	.319
250–1500 ms	.05	.614	**.39**	**<.001**	.05	.652	−.01	.936	.09	.329
250 ms – 3 *SD*	**.21**	**.035**	**.35**	**<.001**	.08	.398	−.12	.221	.16	.102
300–1000 ms	.05	.658	**.43**	**<.001**	.16	.105	−.10	.357	.13	.174
300–1500 ms	.04	.666	**.39**	**<.001**	.05	.649	−.01	.904	.10	.317
300–3000 ms	.18	.082	**.31**	**<.001**	−.02	.832	−.01	.894	.15	.114
300 ms – 2 *SD*	.07	.511	**.38**	**<.001**	.13	.205	−.04	.681	.11	.243
±2 *SD*	.09	.374	**.39**	**<.001**	.13	.182	−.04	.715	.11	.245

Note. Errors only = only error trials are excluded. *SD* = standard deviation.
Significant correlations are highlighted in bold font.

that political attitudes is one of the few domains in which implicit and explicit measures show very high correlations (*r*s between .60 and .70; see Nosek, Graham, & Hawkins, 2010), outlier-treatments can also be evaluated in terms of their impact on correlations with corresponding explicit measures. Whereas more reliable assessment of political preferences via evaluative priming should be reflected in higher correlations with self-reported political preferences, less reliable assessment should be reflected in lower correlations with self-reported political preferences. For these reasons, Study 2 used internal consistency and correlations with self-reported political preferences as the primary criteria for the evaluation of outlier-treatments, putting less emphasis on the emergence and size of a significant evaluative priming effect.

3.1. Methods

3.1.1. Participants

Four-hundred-and-ten psychology undergraduates (272 female, 137 male, 1 unknown) at the University of Texas at Austin participated in the study for course credit. Four participants showed error rates that were outside the error distribution of the sample (>41%). These participants were excluded from further analyses, leaving us with a final sample of 406 participants. The study was part of a one-hour battery that included two unrelated studies in addition to the current one. The current study was always administered as the third one in the battery. The study was approved by the Institutional Review Board of the University of Texas at Austin (IRB # 2016-07-0024).

3.1.2. Evaluative priming task

Participants completed an EPT that included five images of Donald Trump and five images of Hillary Clinton as primes (adapted from Gawronski et al., 2017), and ten positive and ten negative adjectives as targets. The target words, task instructions, and procedural details were identical to the EPT in Study 1. Each of the ten primes was presented once with each of the ten positive target words and once with each of the ten negative words, summing up to a total of 200 trials.

3.1.3. Evaluative rating task

After completion of the EPT, participants were asked to rate their feelings toward Donald Trump and Hillary Clinton on three 7-point scales. One item used the end-point labels *very negative* (1) and *very positive* (7); one item used the end-point labels *very unpleasant* (1) and *very pleasant* (7); and one item used the end-point labels *very bad* (1) and *very good* (7). The 5 images of a given politician were displayed on the right side of the screen when participants provided their ratings of the respective candidate.

3.2. Results

Overall, participants showed the correct response on 93% of the trials in the EPT (range: 68% to 100%). Participants made significantly more errors in the first compared with the second half of the priming task (8.4% vs. 5.0%), $t(405) = 10.29$, $p < .001$, $d = 0.51$ (see Table 2). Latencies from trials with incorrect responses were excluded from the aggregation of evaluative priming scores. For each of the ten outlier-treatments, we calculated a priming score for responses to positive target words by subtracting the mean reaction time for positive target words preceded by Clinton from the mean reaction time for positive target words preceded by Trump. Higher scores on this index reflect greater positivity toward Clinton compared to Trump (see Wentura & Degner, 2010; Wittenbrink, 2007). Correspondingly, a priming score for responses to negative target words was calculated by subtracting the mean reaction time for negative target words preceded by Clinton from the mean reaction time for negative target words preceded by Trump. Higher scores on this index reflect greater negativity toward Clinton compared to Trump (see Wentura & Degner, 2010; Wittenbrink, 2007). The latter index was then subtracted from the former index to obtain a priming index of preferences for Clinton over Trump: Priming Index = [RT(positive targets | Trump) - RT(positive targets | Clinton)] - [RT(negative targets | Trump) - RT(negative targets | Clinton)].

3.2.1. Overall priming effect

Table 3 shows the means and standard deviations of the evaluative priming indices for the ten outlier-treatments. Table 4 shows the results of one-sample *t*-tests comparing the overall priming indices to the neutral reference point of zero. Statistically significant priming effects were obtained for all outlier-treatments. The direction of the priming effect suggests that, on average, participants in the sample showed a significant preference for Clinton over Trump. The effect size of this preference was similar for most outlier-treatments, with Cohen's *d*s between 0.40 and 0.49. The only exception was the Errors only algorithm with a Cohen's *d* of 0.19.

3.2.2. Internal consistency

Estimates of internal consistency for odd-even and two-block splits for the ten outlier-treatments are presented in Table 5. Cronbach's Alpha values were higher compared to the ones obtained in Study 1. The highest values were obtained for the 0–800 ms cutoff. Yet, regardless of outlier-treatment and parceling procedure, estimates were still lower compared to the ones typically shown by the IAT and the AMP (see Gawronski & De Houwer, 2014), with modal values around .40. There was again no evidence that estimates of internal consistency were systematically lower for the two-block split compared to the odd-

even split, which speaks against the possibility that the reliability of evaluative priming scores might change over the course of the task.

3.2.3. Relation to explicit measure

An index of self-reported preferences for Clinton over Trump was calculated by subtracting the mean individual ratings of Trump from the mean individual ratings of Clinton. A one-sample *t*-test revealed a difference score significantly greater than zero, indicating that, on average, participants in the sample showed a significant preference for Clinton over Trump ($M = 2.13$, $SD = 2.67$), $t(405) = 16.05$, $p < .001$, $d = 0.80$. Evaluative priming scores of candidate preferences showed significant positive correlations with self-reported preference scores for all outlier-treatments (see Table 6), ranging from .21 to .48. The lowest correlation was found for the Errors only algorithm; the highest correlation was found for the 0–800 ms cutoff.

3.3. Discussion

Using correlations with self-reported candidate preferences as one central criterion, Study 2 provided positive evidence for the ten outlier-treatments in detecting meaningful individual differences in candidate preferences via evaluative priming. The highest correlations were found for the 0–800 ms cutoff followed by the 300–1000 ms cutoff. The lowest correlation was obtained for the Errors only algorithm. In terms of internal consistency, the outlier-treatment with the best performance was the 0–800 ms cutoff, although estimates of internal consistency were still lower compared to the ones typically shown by the IAT and the AMP (see Gawronski & De Houwer, 2014). The 0–1500 ms cutoff and the 300–1000 ms cutoff showed estimates of internal consistency that were only slightly lower compared to the 0–800 ms cutoff. The lowest internal consistency was obtained for the Errors only algorithm.

Unexpectedly, all data sets produced a significant evaluative priming effect, indicating that, on average, participants in the sample showed a preference for Hillary Clinton over Donald Trump. The same was true for self-reported preferences, indicating that the distribution of political preferences was not symmetric around the neutral midpoint. Because we had no a priori reason to expect such an asymmetry in political preferences, we deem the direction and size of evaluative priming effects as informative about characteristics of the sample rather than features of the outlier-treatments.

4. Study 3: racial attitudes

Study 3 investigated the impact of the ten outlier-treatments in a domain that permits evaluations in terms of all three criteria: racial attitudes. First, there is evidence that implicit racial bias is pervasive across demographic groups (Nosek et al., 2007), suggesting that evaluative priming scores of racial bias should be significantly greater than zero at the sample level. Second, there is evidence for large individual differences in implicit racial bias (Nosek et al., 2007), suggesting that evaluative priming scores of racial bias should have high internal consistencies in capturing these individual differences. Third, measures of implicit and explicit bias have been claimed to capture distinct, but related constructs (Nosek et al., 2007), suggesting that evaluative priming scores of racial bias should be significantly correlated with racial bias on self-report measures, even when these correlations may be moderate overall (for a meta-analysis, see Hofmann et al., 2005). Based on these considerations, Study 3 used all three criteria to evaluate the impact of the ten outlier-treatments on the measurement of racial bias with the EPT.

4.1. Methods

4.1.1. Participants

One-hundred-and-six psychology undergraduates at the University of Texas at Austin (78 female, 25 male, 3 unknown) participated in the study for course credit. Due to computer malfunctions, data from three participants were lost, which left us with a final sample of 103 participants. The study was part of a one-hour battery that included two unrelated studies in addition to the current one. The current study was always administered as the last one in the battery. The study was approved by the Institutional Review Board of the University of Texas at Austin (IRB # 2016-07-0024).

4.1.2. Materials

As prime stimuli, we used head-and-shoulder color photographs of five white men and five black men adapted from Gawronski, Peters, Brochu, and Strack (2008). As target words, we used ten positive nouns (*paradise*, *summer*, *harmony*, *freedom*, *honesty*, *pleasure*, *sunrise*, *love*, *peace*, *vacation*) and ten negative nouns (*cockroach*, *poison*, *vomit*, *bomb*, *virus*, *disaster*, *terror*, *rotten*, *accident*, *pollution*) adapted from Gawronski, Cunningham, LeBel, and Deutsch (2010). As a neutral control prime, we used an image of a gray square of the same size as the images of the ten face primes.

4.1.3. Evaluative priming task

The procedural details of the EPT were identical to Studies 1 and 2. Each of the ten face primes and the neutral control prime were presented once with each of the 20 target words, summing up to a total of 220 trials. Participants received the following instructions for the EPT:

> *In the next part of this study, you will be presented with positive and negative words. In addition, you will be presented with images of faces that briefly appear before the words are presented. Your task is to indicate as quickly as possible if the word on the screen is positive or negative. Please press the "A" key when you see a negative word, and please press the "5" key of the number block when you see a positive word. In order to facilitate faster responses to the words, please keep your main left-hand finger on the "A" key and your main right-hand finger on the "5" key. Please concentrate on the words and try to ignore the faces. And please try to respond as quickly as possible without making too many errors. Again, please press the "A" key when you see a negative word and the "5" key when you see a positive word.*

4.1.4. Evaluative rating task

To maximize conceptual correspondence between the measures of implicit and explicit bias (see Gawronski, 2019), the evaluative rating task measured self-reported evaluations of the faces presented in the EPT. Toward this end, participants were asked to rate the positivity versus negativity of their immediate gut reaction toward each of the ten faces used as prime stimuli on 7-point scales ranging from 1 (*very negative*) to 7 (*very positive*). Order of the faces was randomized individually for each participant.

4.2. Results

Overall, participants showed the correct response on 94% of the trials in the EPT (range: 81% to 100%). There was no significant difference between error rates in the first (6%) and the second (5.7%) half of the EPT, $t(102) = 0.78$, $p = .438$, $d = 0.08$ (see Table 2). Latencies from trials with incorrect responses were excluded from the aggregation of evaluative priming scores. For each of the ten outlier-treatments, we calculated a priming score for responses to positive target words by subtracting the mean reaction time for positive target words preceded by white faces from the mean reaction time for positive target words preceded by black faces. Higher scores on this index reflect greater positivity toward white faces compared to black faces (see Wentura & Degner, 2010; Wittenbrink, 2007). Correspondingly, a priming score for responses to negative target words was calculated by subtracting the mean reaction time for negative target words preceded by white faces from the mean reaction time for negative target words preceded by black faces. Higher scores on this index reflect greater negativity

toward white faces compared to black faces (see Wentura & Degner, 2010; Wittenbrink, 2007). The latter index was then subtracted from the former index to obtain a priming index of preference for white over black faces: Priming Index = [RT(positive targets | black) - RT(positive targets | white)] - [RT(negative targets | black) - RT(negative targets | white)].

4.2.1. Overall priming effect

Table 3 shows the means and standard deviations of the evaluative priming indices for the ten outlier-treatments. Table 4 shows the results of one-sample *t*-tests comparing the overall priming indices to the neutral reference point of zero. Statistically significant priming effects were obtained only for the 0–800 ms cutoff and the 300–1000 ms cutoff, showing a significant preference for white over black faces at the sample level (Cohen's *d*s of 0.33 versus 0.22, respectively). Evaluative priming effects obtained with the other outlier-treatments did not reach statistical significance.

4.2.2. Internal consistency

Estimates of internal consistency for odd-even and two-block splits for the ten outlier-treatments are presented in Table 5. Cronbach's Alphas were mixed, ranging from zero (or negative values) to .78. The lowest values were found for the Errors only algorithm and the 0–800 cutoff; the highest values were found for the 300 ms-2 *SD* cutoff and the ±2 *SD* cutoff with Cronbach's Alphas of .49 and .42 for the two-block split and Cronbach's Alphas of .78 and .75 for the odd-even split. Different from the first two studies, estimates of internal consistency tended to be lower for the two-block split compared to the odd-even split, suggesting that the reliability of evaluative priming scores might change over the course of the task.

4.2.3. Relation to explicit measures

An index of self-reported preference for white over black faces was calculated by subtracting the mean ratings of white faces from the mean ratings of black faces. A one-sample *t*-test comparing self-reported preference scores against zero did not reveal a significant preference for one group over the other at the sample level ($M = 0.17$, $SD = 1.07$), $t(102) = 1.59$, $p = .115$, $d = 0.16$. Evaluative priming scores were not significantly related to self-reported preference scores regardless of the outlier-treatment (see Table 6). The lowest correlation was found for the 300–3000 ms cutoff with $r = -.02$; the highest correlation was found for the 300–1000 ms cutoff with $r = .16$.

4.3. Discussion

Using all three criteria to evaluate the impact of the outlier-treatments on the measurement of implicit racial bias with the EPT, Study 3 obtained the best performance for the 300–1000 ms cutoff. Although estimates of internal consistency were substantially lower with this outlier-treatment compared to the ones typically shown by the IAT and the AMP (see Gawronski & De Houwer, 2014), evaluative priming scores of implicit racial bias were significantly greater than zero and showed the highest correlation to explicit racial bias among the ten outlier-treatments. A significant evaluative priming effect of implicit racial bias was also found for the 0–800 ms cutoff, but the internal consistency of evaluative priming scores and the correlation with explicit racial bias were lower for this algorithm. The highest estimates of internal consistency were found for the 300 ms-2 *SD* cutoff and the ±2 SD cutoff, but these outlier-treatments showed no significant evaluative priming effect at the sample level and correlations with explicit racial bias were somewhat smaller compared to 300–1000 ms cutoff.

5. Study 4: ethnic attitudes

Study 4 shifted the focus from racial to ethnic bias, investigating the impact of the ten outlier-treatments on the measurement of preferences for Germans over Turks in a sample of German participants. In addition to the different target groups, Study 4 differed from Study 3 in three aspects that permitted a more thorough investigation of the outlier-treatments. First, we increased the number of trials in the EPT to investigate whether the low internal consistencies obtained in the previous three studies could be increased by increasing the number of observations. Second, instead of using a measure of explicit bias capturing self-reported evaluations of the faces in the EPT, Study 4 used an established scale to measure prejudice against Turkish people (Pettigrew & Meertens, 1995). Third, because correlations between implicit and explicit bias have been shown to depend on the motivation to control prejudice (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski, Geschke, & Banse, 2003; Payne, 2001), Study 4 additionally included an established scale to measure individual differences in the motivation to act without prejudice (Banse & Gawronski, 2003). Expanding on evidence that correlations between implicit and explicit bias are higher for participants with low motivation to control compared to participants with high motivation to control, Study 4 investigated whether the ten outlier-treatments influence the emergence and size of this interaction pattern.

5.1. Methods

5.1.1. Participants

One-hundred-and-ten students at the University of Bonn in Germany (67 female, 43 male) participated in the study. Eighty-four psychology students received course credit for their participation; the remaining 26 participants volunteered out of interest without compensation. The study protocol was in accordance with the Declaration of Helsinki.

5.1.2. Materials

As prime stimuli, we used ten typical German first names (male: *Lukas*, *Phillip*, *Jonas*, *Felix*, *Paul*; female: *Clara*, *Bettina*, *Maria*, *Julia*, *Anna*) and ten typical Turkish first names (male: *Murat*, *Mustafa*, *Ahmed*, *Erkan*, *Onur*; female: *Fatma*, *Yasemin*, *Bahar*, *Begüm*, *Ebru*). The numbers of the letters were the same for the German and the Turkish names and only included letters that exist in both languages. As neutral baseline primes, we used ten German nouns: *Gabel* (*fork*), *Anlass* (*occasion*), *Bereich* (*area*), *Halle* (*hall*), *Stuhl* (*chair*), *Tisch* (*table*), *Wand* (*wall*), *Kreis* (*circle*), *Löffel* (*spoon*), and *Fenster* (*window*). As targets words, we used German translations of the ten positive and ten negative nouns in Study 3.

5.1.3. Evaluative priming task

Each trial of the EPT began with the presentation of a prime stimulus for 300 ms, followed by a positive or negative target word. Participants were instructed to press a left-hand key (*X*) for negative words and a right-hand key (*M*) for positive words. Incorrect responses were followed by a red *X* presented in the center of the screen for 200 ms. The inter-trial interval was 400 ms. Participants first completed six practice trials. The practice trials were followed by 13 test blocks of 40 trials each, summing up to 520 test trials. The order of the primes and targets was randomized by the computer for each participant. Participants received the following instructions for the EPT (translated from German):

> *In the following task, words should be classified into categories. Please react as quickly as possible to the presented words but also try to make as little mistakes as possible* (*occasional mistakes are okay*). *You will be presented with either negative or positive words, which you have to categorize. Please press the key 'X' for negative words and the key 'M' for positive words. Before each word that should be categorized, you will briefly see a different word or letter sequence. Your task is to ignore this stimulus. Only react to the word that is clearly visible. As soon as you press 'start', the first stimuli you will have to categorize will appear. The other stimuli will follow right after the first one. As a reminder, please*

press 'X' for negative words and 'M' for positive words.

After each block, participants received the following reminder of the instructions:

Now, a new section of the same task starts. Please press 'start' when you are ready to continue. As a reminder: Please press 'X' for negative words and 'M' for positive words.

5.1.4. Self-report measures

After the EPT, participants completed two self-report measures: (1) the Motivation to Act Without Prejudice Scale (MAWP; Banse & Gawronski, 2003) which is a modified German version of the Motivation to Control Prejudiced Reactions Scale (Dunton & Fazio, 1997), and (2) the German version of Pettigrew and Meertens' (1995) Subtle and Blatant Prejudice Scale (SBP; Zick, 1997). Responses on the MAWP Scale were measured with 5-point rating scales ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Responses on the SBP Scale were measured with 4-point rating scales that included different labels for each item.

5.2. Results

Overall, participants showed the correct response on 96% of the trials in the EPT (range: 83% to 100%). Participants made significantly more errors in the second (5.0%) compared to the first half (3.7%) of the EPT, $t(109) = 6.30$, $p < .001$, $d = 0.60$ (see Table 2). Latencies from trials with incorrect responses were excluded from the aggregation of evaluative priming scores. For each of the outlier-treatments, we calculated a priming score for responses to positive target words by subtracting the mean reaction time for positive target words preceded by German names from the mean reaction time for positive target words preceded by Turkish names. Higher scores on this index reflect greater positivity toward Germans compared to Turks (see Wentura & Degner, 2010; Wittenbrink, 2007). Correspondingly, a priming score for responses to negative target words was calculated by subtracting the mean reaction time for negative target words preceded by German names from the mean reaction time for negative target words preceded by Turkish names. Higher scores on this index reflect greater negativity toward Germans compared to Turks (see Wentura & Degner, 2010; Wittenbrink, 2007). The latter index was then subtracted from the former index to obtain a priming index of preference for Germans over Turks: Priming Index = [RT(positive targets | Turkish) - RT(positive targets | German)] - [RT(negative targets | Turkish) - RT(negative targets | German)].

5.2.1. Overall priming effect

Table 3 shows the means and standard deviations of the evaluative priming indices for the ten outlier-treatments. Table 4 shows the results of one-sample *t*-tests comparing the overall priming indices to the neutral reference point of zero. Statistically significant priming effects were obtained for the 0–800 ms cutoff and the 300–1000 ms cutoff, showing a significant preference for Germans over Turks. The size of evaluative priming effects was similar for the two outlier-treatments, with Cohen's *d*s of 0.42 and 0.37, respectively. Although the other outlier-treatments showed effects in the expected direction, priming effects were considerably smaller and did not reach statistical significance.

5.2.2. Internal consistency

Estimates of internal consistency for odd-even and two-block splits for the ten outlier-treatments are presented in Table 5. Similar to the results of Study 3, Cronbach's Alphas were mixed with estimates ranging from zero (or negative values) to .65. The lowest values were found for the 0–800 cutoff and the Error only algorithm; the highest values were found for the 250 ms-3 *SD* cutoff with a Cronbach's Alpha of .33 for the two-block split and .65 for the odd-even split. As in Studies 1 and 2, there was no evidence that estimates of internal consistency were systematically lower for the two-block split compared to the odd-even split, which speaks against the possibility that the reliability of evaluative priming scores might change over the course of the task. Moreover, although estimates of internal consistency tended to be somewhat smaller when only the first 200 trials were used to calculate evaluative priming scores, the advantage of using all 520 trials was negligible and inconsistent across parceling procedures and outlier-treatments. Thus, using a greater number of trials did not improve the rather low internal consistency of evaluative priming scores.

5.2.3. Relations to explicit measures

The self-report measures showed satisfactory internal consistencies with a Cronbach's Alpha value of .79 for the MAWP Scale and a Cronbach's Alpha of .81 for the SBP Scale. MAWP scores showed a significant negative correlation with SBP scores, $r(108) = -.30$, $p = .001$, indicating that higher motivation to act without prejudice was associated with lower prejudice against Turks on the SBP Scale. There was no significant correlation between evaluative priming effects and MAWP scores as well as between evaluative priming effects and SBP scores regardless of the outlier-treatment (see Table 6). The highest correlations between evaluative priming effects and SBP were found for the Errors only algorithm, the 300–1000 ms cutoff, the 250 ms-3 *SD* cutoff, and the 300–3000 ms cutoff.

To test for a potential interaction between implicit bias and motivation to act without prejudice in the prediction of explicit bias, we conducted a series of multiple regressions in which prejudice against Turkish people on the SBP Scale was regressed onto standardized priming effects, standardized MAWP scores, and the interaction of the two predictors. The interaction was not statistically significant regardless of the outlier-treatment (see Table 7).

5.3. Discussion

Using all three criteria to evaluate the impact of the ten outlier-treatments on the measurement of implicit bias with the EPT, Study 4 replicated the superior performance of the 300–1000 ms cutoff obtained in Study 3. Although estimates of internal consistency were again substantially lower with this outlier-treatment compared to the ones typically shown by the IAT and the AMP (see Gawronski & De Houwer, 2014), evaluative priming scores were significantly greater than zero and showed a positive (albeit non-significant) correlation to explicit bias. A significant evaluative priming effect of implicit bias was also found for the 0–800 ms, but this algorithm showed inferior performance in terms of internal consistency and correlation to explicit bias. The highest estimates of internal consistency were found for the 250 ms-3 *SD* cutoff; the highest correlation to explicit bias was found for the Errors only algorithm. However, neither of these two algorithms showed a significant evaluative priming effect at the sample level.

In addition to three primary criteria, Study 4 also included an established scale to measure individual differences in the motivation to act without prejudice (Banse & Gawronski, 2003). Expanding on evidence that correlations between implicit and explicit bias are higher for participants with low motivation to control compared to participants with high motivation to control (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne, 2001), we tested whether the ten outlier-treatments influence the emergence and size of this interaction pattern. Counter to the results of previous studies, none of the outlier-treatments produced a significant interaction between evaluative priming scores and motivation to control in the prediction of explicit bias.

6. General discussion

Table 8 provides a performance summary of the ten outlier-

treatments in terms of the three primary evaluation criteria. The only two outlier-treatments that produced a significant evaluative priming effect across all four studies were the 0–800 ms cutoff and the 300–1000 ms cutoff. Although we had no a priori reason to expect a significant preference for Hillary Clinton over Donald Trump in the sample of Study 2, the findings of the other three studies provide clear evidence for the superiority of the two outlier-treatments in detecting evaluative priming effects. The results were less clear-cut for the internal consistency of evaluative priming scores. Although six of the ten algorithms revealed moderate estimates of internal consistency in some cases (i.e., 0–800 ms cutoff, 0–1500 ms, 250 ms-3 *SD*, 300–1000 ms, 300 ms-2 *SD*, ±2 *SD*), only the 250 ms-3 *SD* cutoff showed superior performance in more than one study (i.e., Studies 3 and 4) and estimates were unsatisfactory in all other cases. Given that the 250 ms-3 *SD* cutoff did not reveal a significant evaluative priming effect in any of the three studies that used the size of evaluative priming effects as a criterion (Studies 1, 3, 4), the obtained estimates may reflect random variation rather than systematic characteristics of a particular outlier-treatment. Finally, the 250 ms-3 *SD* seemed superior for producing meaningful correlations with corresponding explicit measures in two of the four studies. However, one of these correlations emerged in Study 1 where we had no a priori reason to expect systematic individual differences in conditioned attitudes across participants, and thus no basis to expect a significant correlation between evaluative priming scores and corresponding explicit measures. When this case was treated as non-diagnostic, there was no clear-cut difference between the ten outlier-treatments. All outlie-treatments produced a significant correlation with explicit measures in the domain of political attitudes, but not in the domain of racial and ethnic attitudes. Nevertheless, there was a small advantage of the 300–1000 ms cutoff, which showed non-significant correlations that were somewhat higher compared to the ones of most other algorithms in Studies 3 and 4 (in addition to the significant correlation in Study 2). Based on this pattern of results, the 300–1000 ms cutoff showed the best performance overall, although strong caveats seem in order about the low internal consistencies of evaluative priming scores obtained with this outlier-treatment. Yet, in defense of the 300–1000 ms cutoff, it is worth noting that low internal consistency was a challenge for all algorithms, suggesting that low internal consistency might be a problem of evaluative priming in general rather than a feature of particular outlier-treatments (see Gawronski & De Houwer, 2014). Based on these findings, we recommend the 300–1000 ms cutoff for future research with the EPT and potential reanalyses of existing data.

6.1. Overall priming effects

The current findings suggest a clear superiority of the 0–800 ms cutoff and the 300–1000 ms cutoff in the detection of significant evaluative priming effects. To understand the potential reasons for this superiority, it is worth noting that both procedures use an upper limit that is considerably lower compared to the ones in other outlier-treatments based on a priori cutoff values (e.g., 300–3000 ms). The same is true for the a posteriori cutoffs in algorithms that identify outliers based on the actual distribution of response latencies (e.g., 250 ms-3 *SD*, 300 ms-2 *SD*, ±2 *SD*). In the current studies, distribution-based algorithms suggested outlier cutoffs at the upper end that were considerable higher compared to the ones in the 0–800 ms cutoff and the 300–1000 ms cutoff. Thus, in addition to demonstrating the superiority of the two outlier-treatments, the current findings suggest a necessity of sufficiently short intervals between the presentation of the primes and participants' responses to the targets. This conclusion is different from the argument that automatic effects of the primes on responses to the targets require a sufficiently short interval between the onset of the primes and the onset of the targets (i.e., short stimulus onset asynchrony). Even if the interval between prime presentation and *target presentation* is relatively short, evaluative priming effects tend to be weaker (or eliminated) when the delay between prime presentation and *target response* is too long (see Wentura & Degner, 2010). The current findings suggest that effects of prime valence on target evaluations dissipate 800 to 1000 ms after the prime is replaced with the target stimulus—much faster than presumed by outlier-treatments with higher cutoffs. Thus, outlier-treatments that consider these issues will likely show superior performance in the detection of evaluative priming effects.

6.2. Internal consistency

Echoing earlier concerns about the low internal consistency of evaluative priming scores (Gawronski & De Houwer, 2014), the current findings suggest that the low estimates obtained in earlier studies reflect a problem of evaluative priming in general rather than a feature of suboptimal outlier-treatments. Aside from a small number of cases where estimates of internal consistency reached moderate levels (see Table 8), estimates in the current studies were rather low overall and in many cases close to zero (see Table 5). Even increasing the number of trials from 200 to >500 in Study 4 did not help to increase internal consistencies to a level that seems acceptable from a psychometric point of view (see Table 5). Although low internal consistency may not necessarily question the usefulness of evaluative priming for experimental research on attitude formation and change (but see LeBel & Paunonen, 2011), it does pose a major challenge to research using correlational designs (e.g., prediction of behavior with evaluative priming; for reviews, see Friese, Hofmann, & Schmitt, 2008; Perugini, Richetin, & Zogmeister, 2010). The latter type of research presupposes that evaluative priming scores reliably capture individual differences in evaluative responses, which seems questionable when the rank order of participants on one half of the trials is only weakly related (or unrelated) to the rank order obtained on the other half of the trials (i.e., when internal consistency is low). In such cases, differences in evaluative priming scores across participants mostly reflect unsystematic measurement error rather than systematic differences between participants, which undermines the detection of systematic relations to other measures. Moreover, because the internal consistency of a given measure sets an upper limit for potential correlations with other measures, questions could be raised about the possibility of false positives when correlations obtained with evaluative priming scores exceed the upper limit suggested by the internal consistency of these scores.[3]

6.3. Correlations with explicit measures

The significance of low internal consistency is also reflected in the rather low correlations between evaluative priming scores and corresponding explicit measures in the current studies. Consistent with the argument that the internal consistency of evaluative priming scores sets an upper limit for their correlations with explicit measures (see above), correlations were rather small overall. The only exceptions were the correlations obtained in Study 2, which also revealed slightly higher estimates of internal consistency for several outlier-treatments. Although correlations between implicit and explicit measures depend on multiple factors over and above internal consistency (Gawronski & Brannon, 2019), average estimates of internal consistency across studies and outlier-treatments showed a significant positive relation to the overall size of correlation coefficients with $r = .32$. In other words, the

[3] A potential response to this concern is that (a) evaluative priming scores are based on difference scores and (b) the assumptions of classical test-theory may not apply to difference scores. Although we share the latter concern, it is worth noting that, although all implicit measures are based on difference scores, some of them (e.g., IAT, AMP) have shown internal consistencies that meet the psychometric standards of classical test-theory (see Gawronski & De Houwer, 2014).

Table 7
Results of multiple regression analyses using evaluative priming scores, motivation to act without prejudice, and their interaction to predict subtle and blatant prejudice, Study 4.

		Outlier procedure													
		Exclusion of errors only		0–800 ms		0–1500 ms		250–1500 ms		250 ms – 3 *SD*		300–1000 ms		300–1500 ms	
DV	Subtle and blatant prejudices	B (SE)	β	B (SE)	β	B (SE)	β	B (SE)	β	B (SE)	β	B (SE)	β	B (SE)	β
Step 1	Constant	1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)	
	zPE	0.05* (0.03)	0.19	0.01 (0.03)	0.05	0.03 (0.03)	0.09	0.03 (0.03)	0.09	0.04 (0.03)	0.12	0.03 (0.03)	0.10	0.03 (0.03)	0.09
	zMTAWP	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30	−0.08** (0.03)	−0.29	−0.08** (0.03)	−0.29	−0.09** (0.03)	−0.30
Step 2	Constant	1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)	
	zPE	0.05 (0.03)	0.18	0.02 (0.03)	0.06	0.03 (0.03)	0.11	0.03 (0.03)	0.11	0.04 (0.03)	0.12	0.03 (0.03)	0.10	0.03 (0.03)	0.11
	zMTAWP	−0.08** (0.03)	−0.29	−0.09** (0.03)	−0.31	−0.09** (0.03)	−0.31	−0.09** (0.03)	−0.31	−0.08** (0.03)	−0.29	−0.08** (0.03)	−0.29	−0.09** (0.03)	−0.31
	Interaction zPE × zMTAWP	0.07 (0.08)	0.08	−0.01 (0.03)	−0.04	0.02 (0.03)	0.07	0.02 (0.03)	0.07	<0.01 (0.03)	<0.01	0.02 (0.03)	0.05	0.02 (0.03)	0.07
		R^2 = .13 for step 1; ΔR^2 < .01 for step 2 (p = .415)		R^2 = .09 for step 1; ΔR^2 < .01 for step 2 (p = .684)		R^2 = .10 for step 1; ΔR^2 < .01 for step 2 (p = .459)		R^2 = .10 for step 1; ΔR^2 < .01 for step 2 (p = .453)		R^2 = .11 for step 1; ΔR^2 < .01 for step 2 (p = .980)		R^2 = .10 for step 1, ΔR^2 < .01 for step 2 (p = .572)		R^2 = .10 for step 1; ΔR^2 < .01 for step 2 (p = .475)	

		Outlier procedure				
		300–3000 ms		300 ms – 2 *SD*		±2 SD	
DV	Subtle and blatant prejudices	B (SE)	β	B (SE)	β	B (SE)	β
Step 1	Constant	1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)	
	zPE	0.04 (0.03)	0.15	0.03 (0.03)	0.10	0.03 (0.03)	0.10
	zMTAWP	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30
Step 2	Constant	1.89*** (0.03)		1.89*** (0.03)		1.89*** (0.03)	
	zPE	0.06* (0.03)	0.21	0.03 (0.03)	0.10	0.03 (0.03)	0.10
	zMTAWP	−0.09** (0.03)	−0.32	−0.09** (0.03)	−0.30	−0.09** (0.03)	−0.30
	Interaction zPE × zMTAWP	0.03 (0.02)	0.14	−0.01 (0.03)	−0.02	−0.01 (0.03)	−0.02
		R^2 = .11 for step 1; ΔR^2 = .02 for step 2 (p = .163)		R^2 = .10 for step 1; ΔR^2 < .01 for step 2 (p = .815)		R^2 = .10 for step 1; ΔR^2 < .01 for step 2 (p = .849)	

Notes. DV = dependent variable. PE = *z*-standardized priming effect. zMTAWP = *z*-standardized motivation to act without prejudice score.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 8
Performance summary of outlier-treatments in terms of three evaluation criteria (overall priming effect, internal consistency, relation to explicit measures), Studies 1–4.

	Errors only	0 – 800 ms	0 – 1500 ms	250 – 1500 ms	250 ms – 3 *SD*	300 – 1000 ms	300 – 1500 ms	300 – 3000 ms	300 ms – 2 *SD*	± 2 *SD*
Overall priming effect										
Study 1 - conditioned attitudes: $t(96)$		2.55*				**2.62***				
Study 2 - political attitudes: $t(405)$	(3.86***)	(9.58***)	(9.86***)	(9.92***)	(9.24***)	**(9.52***)**	(9.86***)	(7.97***)	(9.16***)	(9.03***)
Study 3 - racial attitudes: $t(102)$		3.31**				**2.26***				
Study 4 - ethnic attitudes: $t(109)$		4.45***				**3.91*****				
Internal consistency										
Study 1 - conditioned attitudes										
Study 2 - political attitudes		Moderate[a]	Moderate[a]			**Moderate[a]**				
Study 3 - racial attitudes					Moderate[a]				Moderate[a]	Moderate[a]
Study 4 - ethnic attitudes					Moderate[a]					
Correlation to explicit measures										
Study 1 - conditioned attitudes: $r(95)$					(.21*)					
Study 2 - political attitudes: $r(404)$	.21***	.48***	.41***	.39***	.35***	**.43*****	.39***	.31***	.38***	.39***
Study 3 - racial attitudes: $r(101)$						**.16[b]**				
Study 4 - ethnic attitudes: $r(108)$	.18[b]									

Note. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. Significant results in parentheses indicate outcomes in domains without a priori reasons for a significant effect. [a]Estimates of internal consistency are marked as *moderate* when the average estimate (two-block, odd-even) is above .40. [b]Highest non-significant correlation with corresponding explicit measures. Empty cells represent non-significant priming effects, averaged reliability estimates (two-block, odd-even) below .40 and non-significant, low correlations, respectively. The values of the outlier-treatment with the best overall performance are highlighted in bold font.

higher the internal consistency of evaluative priming scores, the higher was their correlation with corresponding explicit measures.

The low estimates of internal consistency may also explain why we were unable to replicate previous findings showing that motivation to control moderates the relation between implicit and explicit bias. Expanding on evidence that correlations between implicit and explicit bias are higher for participants with low motivation to control compared to participants with high motivation to control (e.g., Banse & Gawronski, 2003; Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne, 2001), Study 4 tested whether the ten outlier-treatments influence the emergence and size of this interaction pattern. Counter to the results of previous studies, none of the outlier-treatments produced a significant interaction between evaluative priming scores and motivation to control in the prediction of explicit bias scores. Although we cannot rule out that (1) the effect size of the hypothesized interaction is too small to be detected with the current sample (see Maxwell, Lau, & Howard, 2015) or (2) the hypothesized interaction depends on other factors that were not measured in the current research (see Gawronski et al., 2008), low internal consistency of evaluative priming scores might be another reason why we were unable to replicate the interactive effect of implicit bias and motivation to control in the prediction of explicit bias (see LeBel & Paunonen, 2011). The latter possibility echoes the above concerns that low internal consistency of evaluative priming scores can be detrimental for correlational research relating individual differences in evaluative priming scores to individual differences on other measures. However, we would like to emphasize that these concerns apply specifically to research using individual-difference designs and do not necessarily question the usefulness of the EPT for research using experimental designs.

6.4. Log-transformation of reaction times

Many studies in the EPT literature apply a log-transformation to the reaction time data in addition to the elimination of outliers. A common rationale for this procedure is to reduce potential skew in the distribution of data (see Fazio, 1990), which would violate pre-conditions for the application of various data analytic procedures (e.g., ANOVA). In the current studies, 20 of the 40 data subsets were skewed, with the most pronounced skew emerging for the Errors only procedure (see Supplementary Materials, Table S1). Thus, to explore whether log-transformation changes the overall pattern of results, we repeated all of the reported analyses with an additional log-transformation of the data (Supplementary Materials, Tables S3 to S6). Although log-transformation improved the performance of some algorithms (most notably for the Errors only procedure and the 300–3000 ms cutoff; see Supplementary Materials, Table S4), the 300–1000 ms cutoff still showed the best performance overall, regardless of whether response times were log-transformed or not. Because log-transformation led to worse outcomes in several cases regardless of whether the data were skewed or not, we also checked if log-transformation was always effective in normalizing the data (see Supplementary Materials, Table S2). Our analyses revealed that log-transformation effectively eliminated skew for 13 of the 20 skewed data sets. The remaining 7 data sets were still skewed after transformation and 4 previously normally distributed data sets were skewed after log-transformation.

Together, the findings of our supplementary analyses suggest that log-transformation can help to eliminate skew in EPT data. However, although elimination of skew helped to improve the performance of some algorithms (e.g., Errors only), these improvements did not qualify the superior performance of the 300–1000 ms cutoff. Because (1) there was no significant skew in the untransformed data sets with the 300–1000 cutoff, (2) log-transformation did not substantially improve the performance of the 300–1000 cutoff, and (3) the physical meaning of log-transformed time units is theoretically unclear, we recommend using the 300–1000 cutoff without additional log-transformation.

6.5. Caveat

Our conclusions are based on studies with North American and European samples, offering preliminary support for the generality of our findings across populations from different parts of the world. However, all four samples comprised undergraduate students, raising questions about the generality of our findings across populations with different demographic characteristics. Similarly, all four studies were conducted in highly controlled lab settings, raising questions about the generalizability of our findings to studies with other research settings, such as online and field studies. Although conclusions along these lines should be treated as speculative, there is preliminary evidence that the 300–1000 ms cutoff also shows superior performance in online studies with demographically diverse, non-academic samples (e.g., Van Dessel, Gawronski, Smith, & De Houwer, 2017). Yet, whether the 300–1000 ms cutoff proves to be superior in studies with populations that might have difficulties responding to stimuli in less than a second (e.g., elderly participants, clinical patients) is an important question that needs to be addressed in future research.

7. Conclusion

The current research examined the psychometric properties of different outlier-treatments in research with the EPT. The overarching goal was to identify the algorithm with the best psychometric properties, so that it could serve as a standard procedure for future research with the EPT and potential reanalyses of existing data. Toward this end, we compared the ten most frequently used algorithms in the EPT literature in terms of (1) the overall size of evaluative priming effects, (2) their internal consistency, and (3) their relation to corresponding explicit measures. Outlier-treatments were compared in the domains of conditioned attitudes (Study 1), political attitudes (Study 2), racial attitudes (Study 3), and ethnic attitudes (Study 4). The algorithm with the best performance used a priori cutoffs of 300 ms at the lower end and 1000 ms at the upper end, treating response times beyond these cutoffs (and response times from errors) as missing values. Although this algorithm showed superior performance in the detection of significant priming effects at the sample level and in terms of correlations between evaluative priming scores and explicit measures, estimates of internal consistency were rather low and unsatisfactory from a psychometric point of view. Because low internal consistency was a challenge for all algorithms, it presumably reflects a problem of evaluative priming in general rather than a feature of particular outlier-treatments. Thus, although consistent use of the 300–1000 ms cutoff may help to (1) increase the comparability of empirical findings, (2) reduce the likelihood of false positives due to arbitrary choices of outlier-treatments based on predicted outcomes, and (3) reduce the likelihood of false negatives due to outlier-treatments with suboptimal psychometric properties, the current findings highlight limits in the usefulness of the EPT for correlational studies, which require high internal consistency of evaluative priming scores.

Open practices

The four experiments of this article earned Open Materials and Open Data badges. All raw data and analysis syntax are publicly available at https://osf.io/hjm4z/

Acknowledgements

We thank Kristof Keidel, Merlin Monzel, and Finn Rathgeber for their invaluable help with the literature search.

Appendix A. Supplementary Materials

Supplementary Materials to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103905.

References

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15*, 257–278.
Banse, R., & Gawronski, B. (2003). Die Skala Motivation zu vorurteilsfreiem Verhalten: Skaleneigenschaften und Validierung. *Diagnostica, 49*, 4–13.
Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27–41.
Cameron, C. D., Brown-Iannuzzi, J., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behaviors and explicit attitudes. *Personality and Social Psychology Review, 16*, 330–350.
De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10*, 230–241.
Degner, J., & Wentura, D. (2008). The extrinsic affective Simon task as an instrument for indirect assessment of prejudice. *European Journal of Social Psychology, 38*, 1033–1043.
Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316–326.
Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick, & M. S. Clark (Eds.). *Research methods in personality and social psychology* (pp. 74–97). Newbury Park, CA: Sage.
Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.
Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.
Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661–669.
Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*, 285–338.
Gast, A., Gawronski, B., & De Houwer, J. (2012). Evaluative conditioning: Recent developments and future directions. *Learning and Motivation, 43*, 79–88.
Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*, 574–595.
Gawronski, B., Balas, R., & Creighton, L. A. (2014). Can the formation of conditioned attitudes be intentionally controlled? *Personality and Social Psychology Bulletin, 40*, 419–432.
Gawronski, B., & Brannon, S. M. (2019). Attitudes and the implicit-explicit dualism. In D. Albarracín, & B. T. Johnson (Eds.). *The handbook of attitudes. Volume 1: Basic principles* (pp. 158–196). (2nd ed.). New York: Routledge.
Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion, 24*, 1008–1025.
Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (pp. 283–310). (2nd ed.). New York: Cambridge University Press.
Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology, 33*, 573–589.
Gawronski, B., & Mitchell, D. G. V. (2014). Simultaneous conditioning of valence and arousal. *Cognition and Emotion, 28*, 577–595.
Gawronski, B., Mitchell, D. G. V., & Balas, R. (2015). Is evaluative conditioning really uncontrollable? A comparative test of three emotion-focused strategies to prevent the acquisition of conditioned preferences. *Emotion, 15*, 556–568.
Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43*, 300–312.
Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34*, 648–665.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.
Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.
Hermans, D., De Houwer, J., & Eelen, P. (2001). A time course analysis of the affective priming effect. *Cognition and Emotion, 15*, 143–165.
Herring, D. R., White, K. R., Jabeen, Hinojos, M., Terrazas, G., Reyes, S., M., Taylor, J. H., & Crites, S. L. (2013). On the automatic activation of attitudes: A quarter century of evaluative priming research. Psychological Bulletin, 139, 1062–1089.
Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369–1385.
Hu, X., Gawronski, B., & Balas, R. (2017a). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*, 17–32.
Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8*, 858–866.
Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system*

(IAPS): Affective ratings of pictures and instruction manual. Technical report A-7. Gainesville, FL: University of Florida.
LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*, 570–583.
Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*, 487–498.
Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology, 104*, 45–69.
Nosek, B. A., Graham, J., & Hawkins, C. B. (2010). Implicit political cognition. In B. Gawronski, & B. K. Payne (Eds.). *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 548–564). New York: Guilford Press.
Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36–88.
Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.
Perugini, M., Richetin, J., & Zogmeister, C. (2010). Prediction of behavior. In B. Gawronski, & B. K. Payne (Eds.). *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York: Guilford Press.
Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology, 25*, 57–75.
Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the implicit association test: The recoding-free implicit association test (IAT-RF). *The Quarterly Journal of Experimental Psychology, 62*, 84–98.
Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139–165.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility and data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.
Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment, 24*, 237–245.
Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to Implicit Association Tests and related tasks. In B. Gawronski, & B. K. Payne (Eds.). *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117–139). New York: Guilford Press.
Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology, 69*, 23–32.
Vogel, T., Hütter, M., & Gebauer, J. E. (2019). Is evaluative conditioning moderated by big five personality traits? *Social Psychological and Personality Science, 10*, 94–102.
Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski, & B. K. Payne (Eds.). *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). New York: Guilford Press.
Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink, & N. Schwarz (Eds.). *Implicit measures of attitudes* (pp. 17–58). New York: Guilford Press.
Zick, A. (1997). *Vorurteile und Rassismus. Eine sozialpsychologische Analyse [Prejudice and racism: A social psychological analysis]*. Münster, Germany: Waxmann.