The shape of belief: Developing a mousetracking-based relational implicit measure

Jamie Cummins & Jan De Houwer

Ghent University, Belgium

Word count: 5000

Author note

JC and JDH, Department of Experimental Clinical and Health Psychology, Ghent University. This research was conducted with the support of Grant BOF16/MET_V/002 to Jan De Houwer. Correspondence should be sent to jamie.cummins@ugent.be.

Abstract

The Propositional Evaluation Paradigm (PEP; Müller and Rothermund, 2019) has recently shown promise as a *relational implicit measure* (i.e., an implicit measure which can specify how stimuli are related). Whereas the standard PEP measures response times, mousetracking is becoming increasingly-popular for quantifying response competition, with distinct advantages beyond response times. Across four preregistered experiments (N = 737), we interface the utility of the PEP method with the unique benefits of mousetracking by developing a mousetracking PEP (MT-PEP). The MT-PEP very effectively captured group-level beliefs across domains (Experiments 1-4). It produced larger effects (Experiment 3), exhibited superior predictive validity (Experiment 3), and better split-half reliability (Experiments 3-4) than the standard PEP. Both PEPs appear to be intentionally controllable, particularly the MT-PEP (Experiments 3-4). Nevertheless, the MT-PEP shows strong potential in capturing relational information, and may be considered implicit in the sense of capturing fast and unaware (but not unintentional) responding.

Keywords: Propositional Evaluation Paradigm, mousetracking, implicit measures, relational implicit measures, automaticity

The shape of belief: Developing a mousetracking-based relational implicit measure

Implicit measures.

In spite of their status as go-to procedures for assessing mental processes under automaticity conditions (De Houwer et al., 2020; Van Dessel et al., 2020; Gawronski & Sritharan, 2010), most implicit measures (e.g., the Implicit Association Test; IAT, Greenwald et al., 1998; the Affect Misattribution Procedure; AMP, Payne et al., 2005) cannot specify relational information in their procedures. The IAT is capable of telling us that two concepts are related (e.g., that men are more related to doctors than women) but not *how* those stimuli are related. Imagine two participants exhibit such a Men-Doctors IAT effect of magnitude X. Participant 1's effect may be driven by the belief that there *are* more male doctors than female doctors. Participant 2's effect, on the other hand, may be driven by the belief that there *should be* more male doctors than female doctors. These different kinds of implicit beliefs can have very different impacts on behavior (De Houwer et al., 2020; Heider et al., 2015; Remue et al., 2014).

Improving implicit measures

Given these limitations, *relational implicit measures* have been developed to capture such relational information (De Houwer et al., 2015; Cummins & De Houwer, 2019). One promising measure is the Propositional Evaluation Paradigm (PEP; Müller & Rothermund, 2019). Briefly, the PEP harnesses (in)consistencies between presented sentences and required responses (e.g., being required to respond "false" after seeing a sentence that the participant agrees with) in order to assess beliefs. The PEP has already shown promise in measuring gender stereotypes and anti-immigrant racism (Müller & Rothermund, 2019).

Implicit measures can also be improved in ways beyond relational information. In particular, implicit measures typically use response times (RTs) as their dependent variable. However, other outcomes measures can provide advantages beyond RTs. Mousetracking, for

example, allows researchers to view the unfolding of responding across a trial, whereas RTs cannot provide this information (Freeman, 2018; Hehman et al., 2015). Indeed, mousetracking has already been incorporated into the IAT in two previous studies (Smeding et al., 2016; Yu et al., 2012). In particular, Yu and colleagues demonstrated that both response options in the IAT were partially and simultaneously activated during responding on the incompatible block of the IAT. Such a finding gives a critical insight into the dynamics of behavior within the IAT, which could not be observed through RTs.

The current study.

Given the promise of the PEP and the advantages of mousetracking, we aimed to develop and validate a mousetracking PEP (MT-PEP) as a measure of automatic beliefs. Experiment 1 tested whether the MT-PEP was capable of producing group-level effects using factually true or false sentences. Experiment 2 examined whether effects in the MT-PEP reflected automatic processes by using sentences which differed in automatic and deliberated truth values. Experiment 3 tested whether the MT-PEP (and RT-PEP) were sensitive to beliefs about immigrants, and also assessed the fakeability of the two measures. Experiment 4 examined whether modifying the PEPs could reduce the impact of faking.

Experiment 1

In addition to testing whether the MT-PEP can produce effects using factually true and false stimuli, we investigated whether MT-PEP effects were moderated by the proportion of 'catch' trials presented. Catch trials require participants to respond to the truth-value of the sentence stimuli rather than the true/false probes. Due to its preliminary nature, this first experiment was not preregistered (all subsequent experiments were preregistered).

Method

Full details of the procedure and materials for all experiments can be found in the Online Supplementary Materials (OSM). All materials, data, processing and analysis scripts

(and preregistrations Experiments 2-4) can be found on the Open Science Framework (https://osf.io/f4mvr/). For brevity, we describe exploratory analyses only when they are directly relevant to the focus of the manuscript; the results of additional exploratory analyses are presented in the OSM. Data were collected from participants at Ghent University (Experiments 1-2) or Prolific Academic (Experiments 3-4).

Participants. 50 participants provided complete data (34 female, 14 male, 2 no gender given) with a mean age of 21.94 years (SD = 2.30). This sample size provides 90% power to detect a medium Cohen's d effect size (d = 0.47) in our paired-samples t-tests.

Materials. *MT-PEP*. The MT-PEP consisted of 240 trials using two trial types: 'probe' trials and 'catch' trials. Participants were required to respond 'true' or 'false' on each trial by moving the mouse from a starting position (in the bottom-centre of the screen) to either the top-left or top-right of the screen. All trials consisted of a sentence presented word-by-word followed by a prompt. On probe trials, this was either the word 'TRUE' or 'FALSE'. In these trials, participants were required to respond based on the probe word and to *ignore the preceding sentence.* We refer to trials where the required response and the truth value of the sentence were the same as "consistent" trials, and trials where the truth value of the sentences differed from the required response, as "inconsistent" trials. On catch trials, participants were presented with the prompt "?? TRUE OR FALSE ??", and were required to respond "true" ("false") if the sentence was true (false). If the participant responded incorrectly on either trial type they were presented with error feedback. Between participants, the proportion of probe trials to catch trials was varied. Half of the participants completed a 10% catch-trial MT-PEP, which the other half completed a 30% catch-trial MT-PEP. Stimuli within the PEP in this experiment consisted of 20 factually true and false statements about the relative size of different stimuli (e.g., "A continent is bigger than a computer"; see OSM for full information).

Procedure. Participants provided informed consent and demographic information, then completed the 30% or 10% MT-PEP. Participants then answered exploratory questions (see OSM).

Results

We analysed two mousetracking measures: area-under-curve (AUC) and maximum deviation (MD). AUC refers to the area between an optimal response trajectory (i.e., a straight line from the starting position to the response option) and participants' time-normalized average trajectories. A larger AUC score indicates greater deviation of the mouse towards the alternative response option, which indicates an automatic tendency towards this alternative (Hehman et al., 2015). The MD also makes reference to this idealized trajectory, but instead refers to the largest distance between this idealized trajectory and time-normalized response trajectories. Again, larger MD scores indicate greater deviation towards the alternative response.

Hypothesis Testing. We first investigated whether AUC scores differed between trial types. Using a paired samples t-test, we found a significant effect of truth consistency on AUCs in the expected direction; AUCs were larger for inconsistent compared to consistent trials, $t(49) = 7.21$, $p < .001$, Cohen's d = 1.02, 95% CI [0.68, 1.37]. This pattern was also reflected in MD scores, $t(49) = 7.59$, $p < .001$, Cohen's d = 1.07, 95% CI [0.73, 1.43], and RTs, $t(49) = 5.51$, $p < .001$, Cohen's d = 0.78, 95% CI [0.46, 1.10]. We next tested whether MT-PEP effects differed between proportions of catch trials using 3 mixed within-between ANOVAs. We found the expected interaction between truth consistency and proportion of catch trials for AUCs ($F(1, 94) = 5.74$, $p = .019$; see Figure 1) and MDs ($F(1, 94) = 4.82$, $p = .031$), but not RTs ($F(1, 94) = 0.43$, $p = .51$).

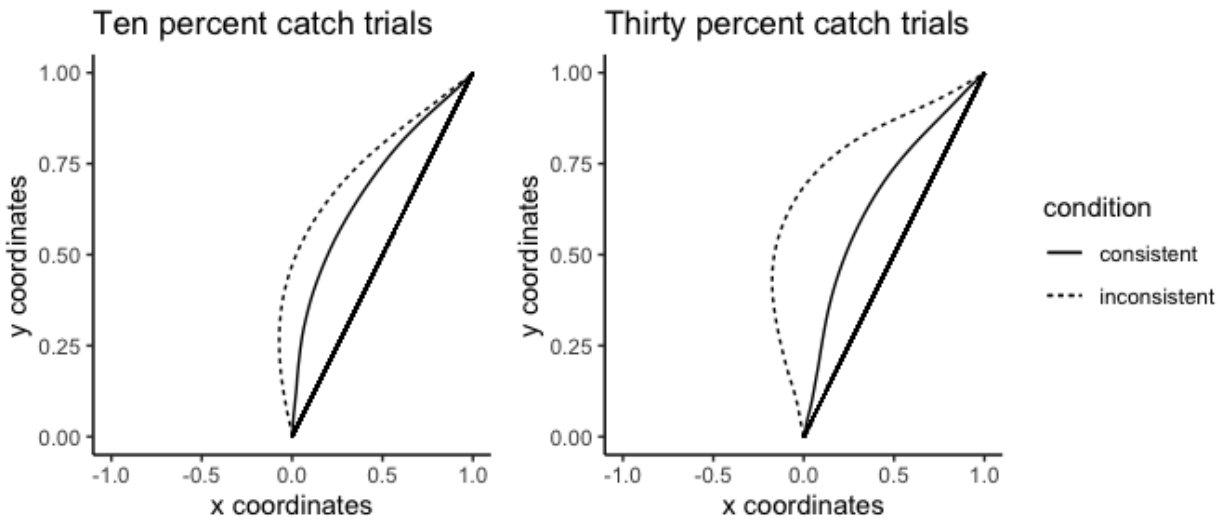

Figure 1. Time-normalized response trajectories for both consistent and inconsistent trials in both versions of the MT-PEP. Responses have been remapped for illustrative purposes, such that the "correct" response option is always on the right-hand-side of the plot. The dark black line represents the optimal response trajectory, with larger deviations from this line indicating greater attraction towards the alternative response option.

Discussion

These results provide initial support for the MT-PEP. Participants in both the 30% and 10% condition showed larger AUC and MD scores (and RTs) on consistent compared to inconsistent trials. AUC and MD effect sizes also increased as a function of greater proportions of catch trials.

Experiment 2

Truth evaluations can be either well-considered (the concentrated integration of multiple pieces of information) or immediate (an initial truth evaluation based on limited information). In some cases, immediate and well-considered truth evaluations can be different, as is the case when categorising atypical category exemplars (e.g., "a whale is a mammal" is immediately evaluated as false, but after consideration is evaluated as true; Dale et al., 2007). If PEP effects reflect well-considered truth evaluations, they should not differ between typical ("a rabbit is a mammal") and atypical (e.g., "a whale is a mammal") exemplars. However, if they capture a more automatic truth evaluation, then typical exemplars should produce larger PEP effects than atypical exemplars.

Method

Participants. Our final sample consisted of 48 participants (37 female and 11 male) with a mean age of 21.85 years (SD = 4.83). Similarly to Experiment 1, 48 participants provides 90% power to detect a medium Cohen's d effect size (d = 0.48) in our paired-samples t-tests.

Materials. The MT-PEP was programmed and administered using PsychoPy 3.0 (Peirce et al., 2019).

MT-PEP.

We use only PEPs with 30% catch trials for the remainder of our experiments. Stimuli in the MT-PEP consisted of animal-category statements in the form "A(n) [animal] is a(n) [category]". Sentence stimuli were taken from Dale et al. (2007; see OSM). Importantly, these sentences consisted of typical (e.g., "a rabbit is a mammal"), or atypical-exemplars (e.g., "a whale is a mammal").

Results

In all subsequent experiments we analyse only AUC scores (AUC and MD were near-perfectly correlated in Experiment 1, r = .98). For all analyses, results did not differ by using AUCs vs. MDs. In subsequent experiments, we utilise linear mixed-effects models when conducting analyses relating to group-level PEP effects in order to maximise statistical power.

Hypothesis Testing. Scores were larger on inconsistent than consistent trials, for AUCs ($t(47) = 9.41$, $p < .001$, Cohen's d = 1.36, 95% CI [0.97, 1.77]; see Figure 2) and RTs ($t(47) = 7.11$, $p < .001$, Cohen's d = 1.03, 95% CI [0.68, 1.39]).

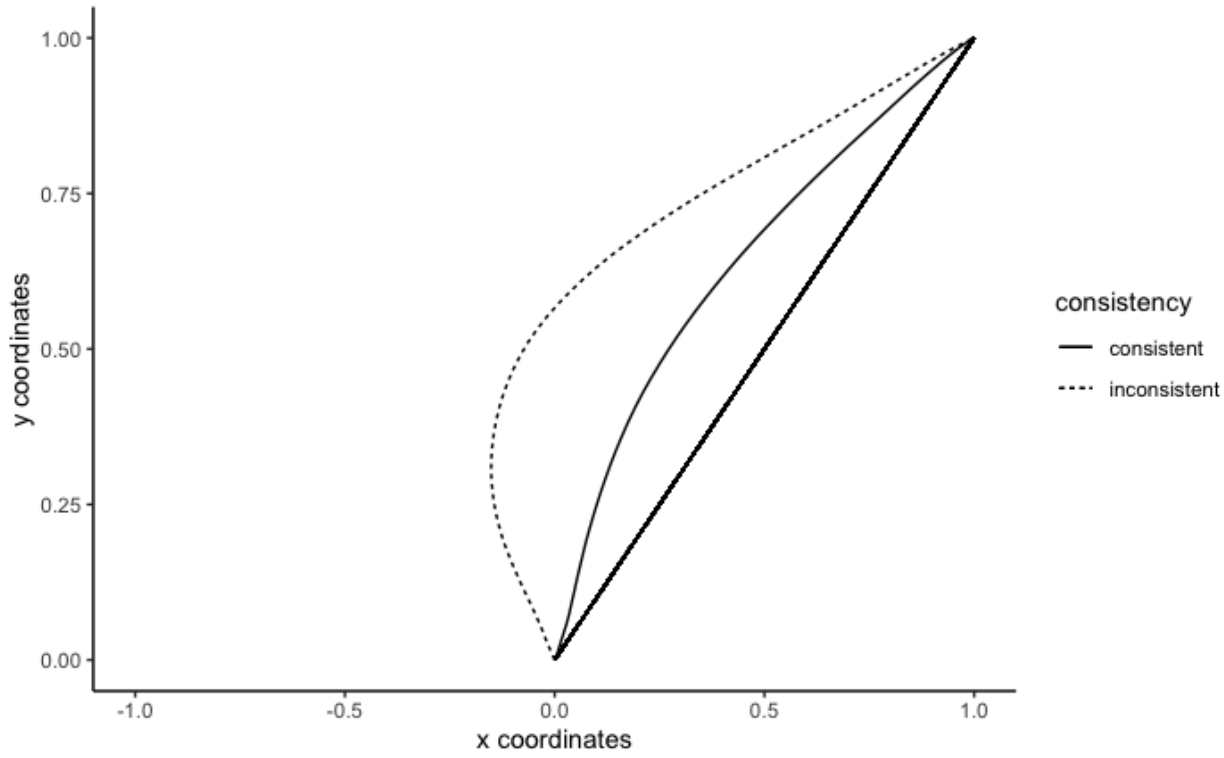

Figure 2. Normalized mouse trajectories for typical-consistent and typical-inconsistent trial types in Experiment 2.

In our mixed-effects model we found the expected consistency * typicality interaction for AUCs, β = 0.07, 95% CI [0.04, 0.10], $p < .001$ (see Figure 3). We did not find this significant interaction in an equivalent fixed effects analysis, $F(1, 184) = 0.66$, $p = .419$. For RTs, we did not find a significant interaction in either the mixed-effects model (β = 0.03, 95% CI [-0.00, 0.06], $p = 0.057$) or the fixed-effects model ($F(1, 184) = 0.003$, $p = .958$). In follow-up analyses[1] recommended by a reviewer, we investigated which point in the time-course of responding the consistency * typicality interaction emerged: in other words, at what point the difference in x coordinates between consistent and inconsistent trials became greater for typical, compared to atypical, trials. To do this, we conducted linear mixed-effects analyses at each of the 101 normalised response timepoints, modelling consistency and typicality as fixed effects, and x coordinates as the DV. The typicality * consistency

[1] Presented in the analysis file for Experiment 2.

interaction was significant (with Bonferroni correction) between the 3rd and 55th normalised timepoints, indicating this effect emerged very early in the response time-course.

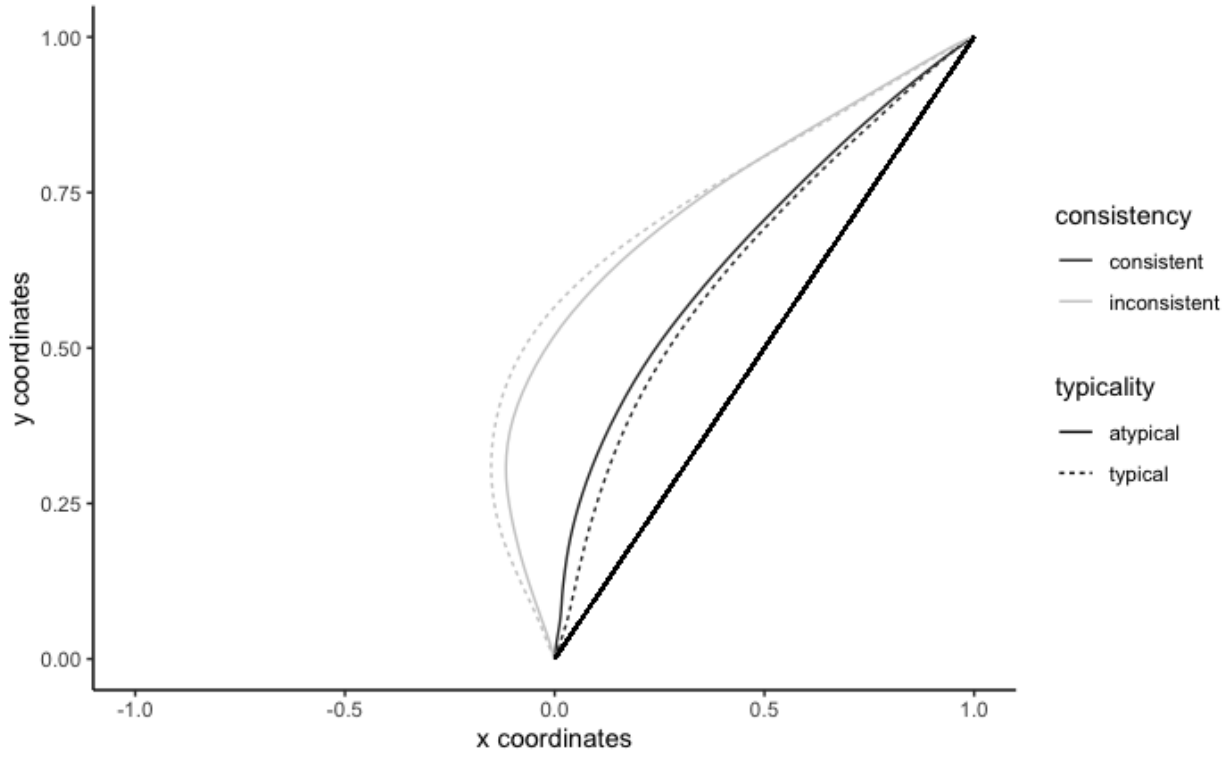

Figure 3. Normalized mouse trajectories for all 4 consistency-typicality trial types.

Discussion

We again demonstrated the basic MT-PEP effect (for typical-exemplars). Analyses of AUCs showed that this effect was moderated by sentence typicality: MT-PEP effects were larger for typical- compared to atypical-exemplar trials, albeit only in the more statistically-powerful mixed-effects analysis. This suggests that MT-PEP effects at least in part reflect the immediate, automatic (in the sense of fast) truth evaluation of sentences.

Experiment 3

The MT-PEP appears sensitive to factually true and false sentences. However, it is unclear whether MT-PEP effects arise for subjective beliefs (see Heiphetz et al., 2014). As Müller and Rothermund (2019) previously did using the RT-PEP, in the current experiment we tested whether the MT-PEP (and RT-PEP) were sensitive to beliefs relating to

immigrants, and whether PEP effects correlated with deliberated beliefs. In addition, we examined whether effects in either PEP could be faked, both at the group-level or at the individual differences level. Because participants are less aware of mouse movements than RTs, and knowledge of the measured variable increases the fakeability of tasks (Hughes et al., 2016; Röhner et al., 2013; Steffens, 2004; though see Fiedler & Bluemke, 2005), we expected that the MT-PEP would be less fakeable than the RT-PEP.

Method

Participants. Our final sample consisted of 205 participants (79 men, 123 women, 1 agender, 1 non-binary, 1 no gender given) with a mean age of 31.66 years (SD = 10.84). 107 completed the RT-PEP; 98 completed the MT-PEP. Across our analyses, this provides us with 90% power to detect a minimal Cohen's d effect size of d = 0.50 (i.e., in our linear regression with three interactive predictors).

Materials. All materials were programmed and administered using lab.js (Henninger et al., 2019).

MT-PEP. We modified the MT-PEP very slightly to facilitate online data collection (see OSM).

RT-PEP. The RT-PEP was identical to the MT-PEP, except that participants responded using key presses ('A' and 'L' keys).

Self-reported beliefs. Self-reported anti-immigrant beliefs were assessed using the Modern and Classic Racism Scales (McConahay, 1986; Pettigrew & Meertens, 1995). Scores from the two scales were averaged into one compound score, similarly to Müller and Rothermund (2019). Reliabilities (Cronbach's alpha) for both the Classic (a = 0.90, 95% CI [0.88, 0.92]) and Modern (a = 0.86, 95% CI [0.83, 0.89]) Scales were acceptable.

Procedure. Participants were collected in two groups: first the control group completed the study after receiving standard instructions. From this we established that our

participants generally exhibited pro-immigrant beliefs, and thus that faking instructions should require participants to fake anti-immigrant beliefs. We then ran a second group of participants who were instructed to fake strong anti-immigrant beliefs. Participants provided demographic information, then completed the MT- or RT-PEP. After the PEP, faking participants were told to cease faking. Participants then completed the self-report measures.

Results

Hypothesis Testing. When assessing anti-immigrant beliefs, we refer to two trial types: pro-immigrant trials and anti-immigrant trials. Pro-immigrant trials consisted of either a pro-immigrant sentence followed by the probe word 'true', or anti-immigrant sentences followed by the probe word 'false' (i.e., these responses indicate pro-immigrant beliefs). Anti-immigrant trials consisted of converse configurations to pro-immigrant trials. MT-PEP (RT-PEP) scores were calculated based on the difference between AUCs (RTs) in pro- vs. anti-immigrant trials, such that a more positive score indicated stronger anti-immigrant beliefs.

At the group-level, RTs in the RT-PEP were significantly shorter on pro-immigrant trials compared to anti-immigrant trials, $\beta = -0.03$, 95% CI [-0.05, -0.01], $p = .004$. Likewise, AUCs in the MT-PEP were smaller for pro-immigrant trials, $\beta = -0.23$, 95% CI [-0.26, -0.21], $p < .001$ (see Figure 4). We also found this pattern in t-tests for both the RT-PEP, $t(55) = 3.040$, $p = .003$, Cohen's d = 0.41, 95% CI [0.13, 0.68] and MT-PEP, $t(50) = 8.946$, $p < .001$, Cohen's d = 1.25, 95% CI [0.89, 1.63]. At the individual-level, larger anti-immigrant effects in both the RT-PEP ($r = .28$, 95% CI [.02, .51], $p = .034$) and the MT-PEP ($r = .41$, 95% CI [.15, .61], $p = .003$) correlated with greater self-reported beliefs.

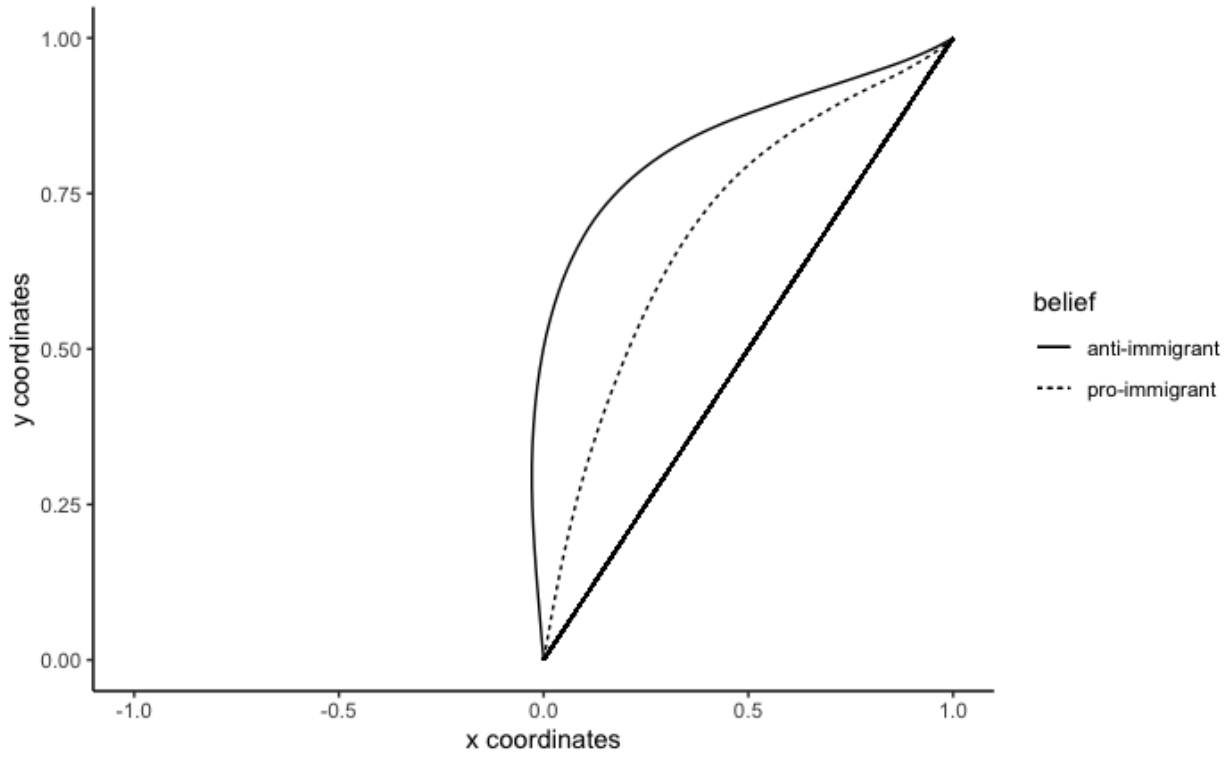

Figure 4. Time-normalised mouse trajectories for pro- and anti-immigrant trials in the MT-PEP control condition (Experiment 3).

We next investigated group-level faking. We firstly compared PEP effects in the control and faking conditions for each PEP. Effects in the faking condition were significantly different, in the opposing direction, to effects in the control condition, for both the RT-PEP, β = 0.06, 95% CI [0.03, 0.08], $p < .001$, and the MT-PEP, β = 0.35, 95% CI [0.32, 0.38], $p < .001$. When comparing the PEPs, we found a significant interaction effect between PEP variant and condition in predicting (standardised) PEP effects, β = -0.64, 95% CI [-0.96, -0.31], $p < .001$. Faking was stronger in the MT-PEP compared to the RT-PEP (see Table 1).

Table 1. Mean (standardised) group-level effects in the RT- and MT-PEP in Experiment 3.

Experimental condition	PEP variant	
	RT-PEP	MT-PEP
Control	-0.07 (0.13)	-0.47 (0.35)
Faking	0.05 (0.11)	0.34 (0.46)

We next examined whether faking impacted the ability of the PEPs to predict individual differences in self-reported beliefs. Counter to our expectations, condition (control vs. faking) did not influence the prediction of self-report scores by PEP effects differently for the MT-PEP than for the RT-PEP, β = 0.35, 95% CI [-0.33, 1.03], *p* = .310. The ability of PEP scores to predict self-reported beliefs was not significantly moderated by experimental condition for the MT-PEP (β = -0.44, 95% CI [-0.99, 0.11, *p* = .119) or the RT-PEP (β = -0.03, 95% CI [-0.44, 0.38], *p* = .875; see Table 2).

Table 2. Coefficients (95% CIs in square brackets) for the prediction of self-report scores for each condition and each PEP variant in Experiment 3. Coefficients marked with an asterisk denote significance from zero (p < .05).

MT-PEP	Control	.41* [.15, .61]
	Faking	.11 [-.19, .41]
RT-PEP	Control	.28* [.02, .51]
	Faking	.18 [-.10, .46]

Exploratory Testing. Standardized effect sizes in the MT-PEP were larger than in the RT-PEP, *t*(105) = 2.05, p = .042, Cohen's d = 0.40, 95% CI [0.01, 0.78]. Split-half reliabilities for the MT-PEP were high, but the reliabilities for the RT-PEP were poor (see Table 3). We examined AUC scores on the catch trials of the MT-PEP between trials where participants responded in a pro-immigrant manner and trials where participants responded in an anti-immigrant manner. We found a significant interaction between trial type and faking condition on catch trial AUCs, β = -0.63, 95% CI [-1.12, -0.14], *p* = .011. AUCs were larger for anti-immigrant trials than for pro-immigrant trials in the faking condition (β = -.61, 95% CI [-.93, -.29], *p* < .001, but did not differ in the control condition (β = -.00, 95% CI [-.37, .36], *p* = .993; see Figure 5). Notably, we did not find this significant interaction effect when examining RTs in catch trials of the RT-PEP, β = -0.22, 95% CI [-0.59, 0.16], *p* = .259.

Table 3. The split-half reliabilities of the MT-PEP and RT-PEP in the control and faking conditions in experiments 3 and 4.

Experiment 3	MT-PEP	Control	*Rsb* = .83, [.71, .90]
		Faking	*Rsb* = .85, [.73, .92]
	RT-PEP	Control	*Rsb* = -.27, [~-1, .25]
		Faking	*Rsb* = -.06, [-.64, .46]
Experiment 4	MT-PEP	Control	*Rsb* = .45, [.21, .62]
		Faking	*Rsb* = .50, [.28, .66]
	RT-PEP	Control	*Rsb* = -.39, [~-1, .04]
		Faking	*Rsb* = .13, [-.24, .40]

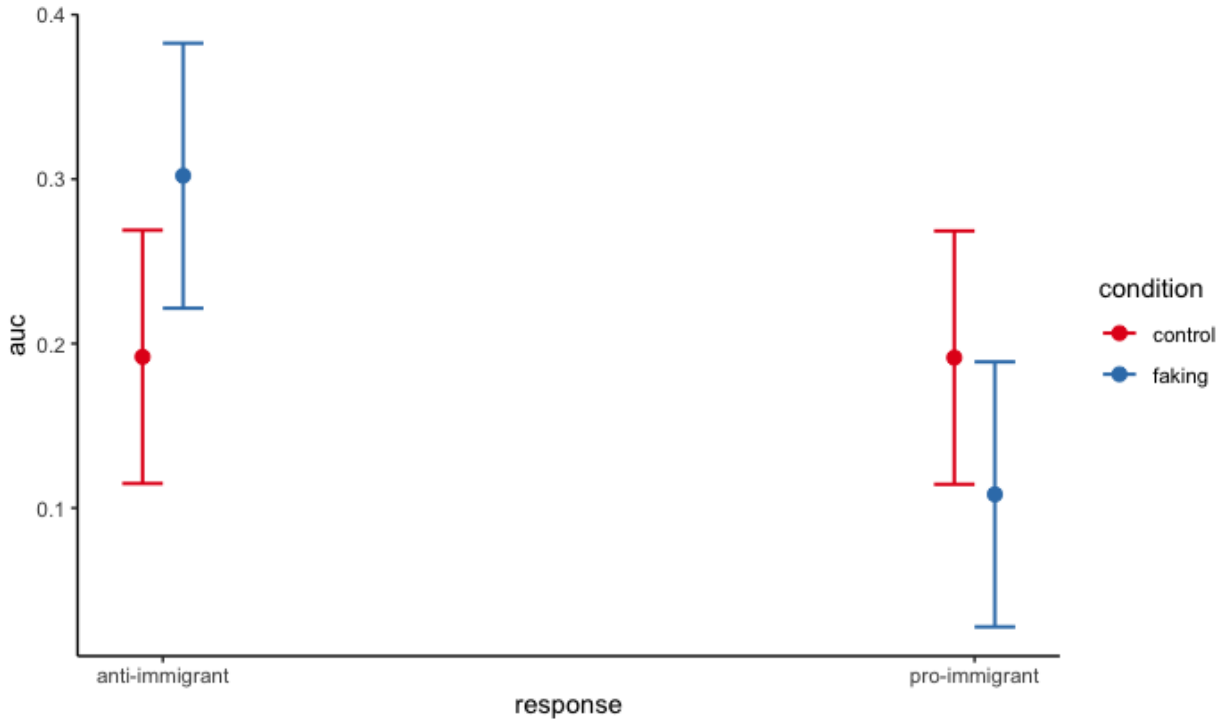

Figure 5. The AUC scores for catch trials in the MT-PEP on which participants responded in a pro- or anti-immigrant manner in the control and faking conditions.

Discussion.

Our results suggest that the MT-PEP and RT-PEP were sensitive to group-level immigrant-related beliefs. MT- and RT-PEP effects also correlated with individual differences in anti-immigrant beliefs. Both PEPs were fakeable at the group-level, but faking impacted the MT-PEP more than the RT-PEP. However, when faking in the MT-PEP, participants still exhibited pro-immigrant responses on MT-PEP catch trials. Notably, faking did not reduce the relationship between either PEP and self-report scores.

Why might the PEPs be susceptible to group-level faking? One potential explanation could be that our faking instructions resulted in the absence of an important component of PEP effects: a truth evaluation mindset. That is, a truth evaluation mindset is arguably induced via the catch trials (Wiswede et al., 2013). However, catch trials served a different purpose for participants in the faking condition: they offered a context where participants must respond *as if* they have a specific belief (i.e., not evaluating based on their own opinions). As a result, it is likely that no truth evaluation mindset was induced.

Experiment 4

PEP effects might be less fakeable if an evaluation mindset could be maintained while faking. A variant of the PEP already exists which can help us test this: in their second and third experiments, Müller and Rothermund (2019) used catch trials where participants needed to identify whether there was a spelling mistake present in sentences, rather than truth-evaluate them. Responses to these catch trials should not vary based on whether a participant is faking or not. Thus, an evaluation mindset should be induced for all participants, which should reduce the impact of faking. Müller and Rothermund state this this manipulation will lead specifically to a *truth* evaluation mindset. However, it might well be that the mindset which is induced is more a "spelling evaluation" mindset. Regardless of those specifics, this

manipulation produced meaningful effects in PEP probe trials for Müller and Rothermund, and is thus worth utilising here. Our fourth experiment was identical to our third experiment, but now with spelling evaluation as the catch trial task.

Method

Participants. Our final sample consisted of 434 participants (144 men, 288 women, 1 non-binary, 1 no gender given) with mean age of 30.85 years (SD = 10.82). 225 participants completed the RT-PEP; 209 completed the MT-PEP. Across our analyses, this provides us with 90% power to detect a minimal Cohen's d effect size of d = 0.41 (i.e., in our linear regression with three interactive predictors).

Materials. Sentences in the PEPs were presented in two forms: with and without a spelling mistake. Reliabilities (Cronbach's alpha) for both the Classic (a = 0.88, 95% CI [0.87, 0.90]) and Modern (a = 0.86, 95% CI [0.84, 0.88]) Scales were again acceptable.

Results

Hypothesis Testing. In mixed-effects analyses, we found significant pro-immigrant effects in both the RT-PEP, β = -0.03, 95% CI [-.0, 0.02], p < .001, and the MT-PEP, β = -.05, 95% CI [-.07, -.04], p < .001. We found a similar pattern of results in fixed-effects analyses: significant pro-immigrant PEP effects for the RT-PEP, t(109) = 4.99, p < .001, Cohen's d = 0.48, 95% CI = [0.28, 0.68], and for the MT-PEP, t(107) = 7.83, p < .001, Cohen's d = 0.75, 95% CI = [0.54, 0.97] (see Figure 6). However, we did not find a significant correlation between PEP scores and self-reported beliefs for the RT-PEP (r = .12, 95% CI [-.07, .30], p = .205) or the MT-PEP (r = .04, 95% CI [-.15, .22], p = .710).

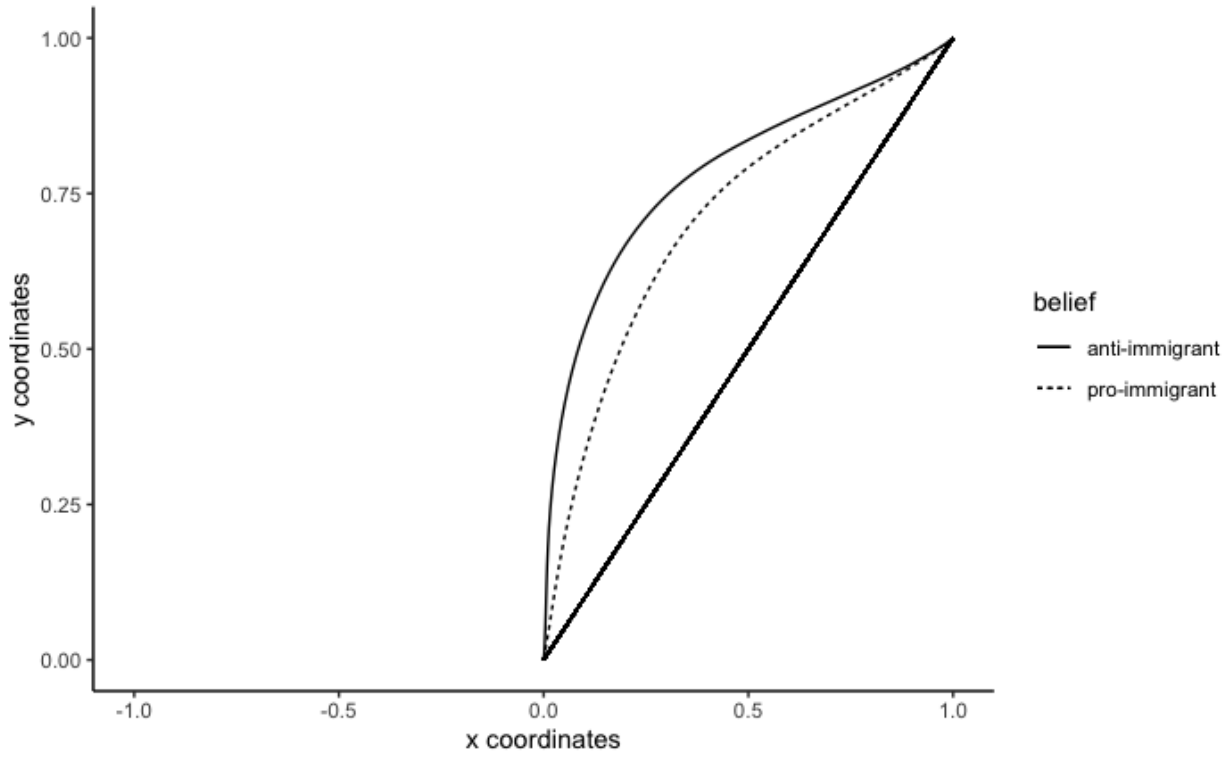

Figure 6. Time-normalised mouse trajectories for pro- and anti-immigrant trials in the MT-PEP control condition (Experiment 4).

Group-level faking effects were larger in the MT-PEP than the RT-PEP, β = -0.30, 95% CI [-0.57, -0.04], p = .025. As in our third experiment, this was in the opposite direction to our expectations (see Table 4). PEP effects did not significantly differ between the faking and control conditions in RT-PEP, β = 0.01, 95% CI [-0.01, 0.03], p = .231, but did in the MT-PEP, β = 0.05, 95% CI [0.03, 0.07], p < .001, with faking participants exhibiting smaller MT-PEP effects than control participants.

Table 4. Mean (standardised) group-level effects in the RT- and MT-PEP in Experiment 4.

Experimental condition	PEP variant	
	RT-PEP	MT-PEP
Control	-0.10 (0.17)	-0.19 (0.24)
Faking	-0.05 (0.17)	-0.06 (0.25)

We did not find the expected impact of condition and PEP variant on the relationship between PEP scores and self-reported beliefs (β = 0.11, 95% CI [-.29, .51], p = .584). Put briefly, only the RT-PEP in the faking condition significantly predicted self-report scores. However, this effect did not significantly differ between the control and faking conditions of the RT-PEP, β = 0.05, 95% CI [-.23, .33], p = .726 (see Table 5).

Table 5. Beta coefficients (95% CIs in square brackets) for the prediction of self-report scores for each condition and each PEP variant in Experiment 4.

MT-PEP	Control	.04 [-.15, .23]
	Faking	-.02 [-.22, .18]
RT-PEP	Control	.12 [-.07, .31]
	Faking	.20* [.02, .38]

Exploratory Testing. Standardised effect sizes did not significantly differ between the MT-PEP and RT-PEP, t(216) = 21.29, p = .195, Cohen's d = 0.18, 95% CI [-0.09, 0.44]. MT-PEP split-half reliabilities were higher than RT-PEP reliabilities, but all reliabilities were poor (Table 4). RTs in the MT-PEP were shorter on pro- compared to anti-immigrant trials, β = -0.03, 95% CI [-.04, -.01], p < .001. We also found a faking effect for RTs in the MT-PEP: β = .02, 95% CI [.00, .04], p = .014.

Across all four experiments, we assessed participants' beliefs about what was being measured within the PEP. Table 6 outlines the trends of these responses (coded independently by the first and second author). In brief: substantially fewer participants expected mouse movements, compared to RTs, were being measured.

Table 6. The number (and percentage) of participants who answered that either (i) mouse movements or (ii) response times were the dependent variable of interest, for each of our four experiments.

Experiment	Procedure	Outcome variable	Suspected	Inter-rater agreement

Experiment 1	MT-PEP	Mouse movements	2 (4%)	94%
		Response times	20 (40%)	
Experiment 2	MT-PEP	Mouse movements	0 (0%)	76%
		Response times	16 (35%)	
Experiment 3	MT-PEP	Mouse movements	5 (5%)	95%
		Response times	42 (43%)	
	RT-PEP	Mouse movements	0 (0%)	
		Response times	15 (14%)	
Experiment 4	MT-PEP	Mouse movements	4 (2%)	94%
		Response times	67 (32%)	
	RT-PEP	Mouse movements	0 (0%)	
		Response times	50 (23%)	

General Discussion

Is the MT-PEP valid and reliable?

Our results represent support for the MT-PEP as a valid and reliable relational implicit measure. For group-level beliefs, the MT-PEP demonstrated effects when using factual stimuli (Cohen's d = 1.02 in Experiment 1, Cohen's d = 1.36 in Experiment 2) and for race-related stimuli (Cohen's d = 1.25 in Experiment 3, Cohen's d = 0.75 in Experiment 4).

In similar contexts, the IAT produces an effect size around Cohen's d = 0.6 – 0.75 (e.g., Agosta & Sartori, 2013; Frantz et al., 2004; Bar-Anan & Nosek, 2014). The MT-PEP demonstrated mixed validity as a measure of *individual differences* in self-reported beliefs. In Experiment 3, MT-PEP scores correlated strongly with self-reported beliefs (r = 0.41). For comparison, the correlation between IAT scores and self-report beliefs tends to be between .20 and .35 (Hofmann et al., 2005; Bar-Anan & Nosek, 2014). However, when catch trials did not involve truth evaluation, the MT-PEP no longer correlated with self-reported beliefs (Experiment 4). This difference also affected the MT-PEP's split-half reliability (r = .83 in Experiment 3, r = .39 in Experiment 4).

Are MT-PEP effects implicit?

MT-PEP effects can be considered implicit in the sense of "fast" (Moors & De Houwer, 2007). Responses in the MT-PEP generally occurred quickly (Experiment 1 M = 995ms; Experiment 2 M = 805ms; Experiment 3 Control M = 603ms; Experiment 3 Faking M = 565ms; Experiment 4 Control M = 575ms, Experiment 4 Faking M = 588ms). These RTs are consistent with RTs in other "fast" implicit measures such as the IAT, AMP, and EPT (Bar-Anan & Nosek, 2014). One might argue that responding in the MT-PEP is less fast than other measures because participants can process parts of the sentence stimuli before the final word of the sentence appeared. However, the meaning of sentences was most often apparent only after the final word had been presented. Hence, the impact of the truth value of the sentence on responding likely occurred quickly. Effects in the MT-PEP also appear to occur without awareness, as few participants (0-5%) correctly stated that mouse movements were being measured.

Some have suggested the PEP reflects unintentional truth evaluation (e.g., Müller & Rothermund, 2019). However, the results of Experiments 3 and 4 suggest that participants can intentionally fake beliefs (indicating *intentional* truth evaluation). In hindsight, this is

unsurprising. Until participants see the prompt word, they cannot know whether the truth value of the sentence will be relevant (catch trials) or not (probe trials). Hence, they might *always* intentionally process whether sentences are true or false. PEP effects thus likely reflect an intentionally-initiated truth evaluation that occurs quickly, is measured *indirectly* (i.e., by assessing responding to the probes), and without awareness. The intentional nature of PEP effects is supported by the fact that MT-PEP effects increase with greater proportions of catch trials, and that PEP effects predict self-reports only when catch trials involve truth evaluation. Group-level PEP effects still arise when catch trials do not involve truth evaluation, but to the detriment of reliability and predictive utility.

PEP effects reflecting quickly-emitted intentional truth evaluation is also in line with the observation in Experiment 2 that MT-PEP effects were influenced by the automatic truth value of sentences. When participants have little time to intentionally truth-evaluate a sentence (e.g., "a whale is a fish"), they might be inclined to believe that it is false, even when it is true. Hence, (MT-)PEP effects might provide a useful tool to tap into the early stages of intentional truth evaluation, that is, before participants evaluate the truth in a deliberate, well-considered manner.[2] The idea that PEP effects reflect quickly-emitted intentional truth evaluation also aligns with the unexpected finding that the MT-PEP was less fakeable than the RT-PEP at the group-level. Both the MT-PEP and RT-PEP reflected intentional truth evaluation in line with the goals of the task (i.e., to express anti-immigrant beliefs), and the mouse trajectories within the MT-PEP simply reflected this to a greater extent than the RTs of the RT-PEP.

The MT-PEP vs. the Race-MT.

[2] Exploratory analyses of Experiment 2 data showed that participants erred more on catch trials for atypical- compared to typical-exemplars. This suggests that automatic truth value influences responding even when participants deliberatively truth-evaluate. It also supports the idea that the PEP effects on probe trials reflect what happens at an early stage during intentional truth evaluation: both PEP effects on probe trials and errors on catch trials reflect spontaneous (i.e., fast) beliefs.

After conducting our experiments, we discovered that our procedure is similar to another mousetracking procedure used to assess automatic racial biases: the Race-MT (Melnikoff et al., 2020; based on Wojnowicz et al., 2009). This procedure involved presenting Black/White names and positive/negative concepts as stimuli, with the response options "Like" or "Dislike". Participants were required (via feedback) to respond "Like" to names of Black and White people, and exhibited larger AUCs in responding "Like" to Black names compared to White names. Although similar, the MT-PEP has the distinct advantage of incorporating relational information into its stimuli. Additionally, Melnikoff et al.'s procedure involved tying response options to a single feature of stimuli (e.g., liking). However, racial groups (and groups in general) can be attached to multiple *features* (e.g., intelligence, perceived dangerousness, etc.; Hughes et al., 2020). By focusing on the truth evaluation of presented sentences, the MT-PEP can investigate multiple features simultaneously, as well as the different ways social groups *relate* to these features.

The Race-MT is similar to catch trials on the MT-PEP (with the exception that participants are given feedback in the Race-MT). The Race-MT and MT-PEP in principle could be interfaced to take advantage of both procedures. Indeed, Melnikoff et al. identified two limitations of the Race MT method: its small effect sizes, and the fact that (in some contexts) a more indirect measure may outperform it. The MT-PEP offers large group-level effect sizes, and is more indirect. As such, interfacing the Race-MT method into the catch trials of the MT-PEP may help to overcome the limitations of both procedures, and provide researchers with a robust and powerful tool to gain insight into beliefs.

Conclusion.

We found that the MT-PEP consistently produced strong group-level effects for both factual and subjective beliefs. These effects can be described as automatic: they were produced quickly by participants and occurred outside of awareness. The PEP also predicted

self-reported beliefs, but only when catch trials involved intentional truth evaluation. Our findings provide some initial support for the MT-PEP as a measure of automatic beliefs, but also highlight boundary conditions for its usefulness. However, at a time when confidence in "classic" implicit measures is waning (Cummins et al., 2019; Schimmack, 2019), the MT-PEP offers a number of novel advantages, and overcome issues associated with other procedures.

References

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00519

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Chester, D., & Lasko, E. (2019). Construct Validation of Experimental Manipulations in Social Psychology: Current Practices and Recommendations for the Future [Preprint]. https://doi.org/10.31234/osf.io/t7ev9

Cummins, J., & De Houwer, J. (2019). An inkblot for beliefs: The Truth Misattribution Procedure. *PloS One*, *14*(6), e0218661. https://doi.org/10.1371/journal.pone.0218661

Cummins, J., Hussey, I., & Hughes, S. (2019). The AMPeror's New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes. [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/d5zn8

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28. https://doi.org/10.3758/BF03195938

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00319

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 61, pp. 127–183). Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.004

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and Unaided Response Control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*(4), 307–316. https://doi.org/10.1207/s15324834basp2704_3

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A Threat in the Computer: The Race Implicit Association Test as a Stereotype Threat Experience. *Personality and Social Psychology Bulletin*, *30*(12), 1611–1624. https://doi.org/10.1177/0146167204266650

Freeman, J. B. (2018). Doing Psychological Science by Hand. *Current Directions in Psychological Science*, *27*(5), 315–323. https://doi.org/10.1177/0963721417746793

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). The Guilford Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401. https://doi.org/10.1177/1368430214538325

Heider, N., Spruyt, A., & De Houwer, J. (2015). Implicit beliefs about ideal body image predict body image dissatisfaction. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01402

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2014). What do Different Beliefs Tell us? An Examination of Factual, Opinion-Based, and Religious Beliefs. *Cognitive Development*, *30*(April-June 2014), 15–29. https://doi.org/10.1016/j.cogdev.2013.12.002

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). *lab.js: A free, open, online study builder* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fqr49

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 & Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hughes, S., De Houwer, J., Mattavelli, S., & Hussey, I. (2020). The Shared Features Principle: If two objects share a feature, people assume those objects also share other features [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/hp5kr

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology*, *46*(5), 632–648. https://doi.org/10.1002/ejsp.2207

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

Melnikoff, D. E., Mann, T. C., Stillman, P. E., Shen, X., & Ferguson, M. J. (2020). Tracking Prejudice: A Mouse-Tracking Measure of Evaluative Conflict Predicts Discriminatory Behavior: *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619900574

Moors, A., & Houwer, J. D. (2007). What is Automaticity? An Analysis of Its Component Features and Their Interrelations. In J. A. Bargh, *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 11–50). Psychology Press.

Müller, F., & Rothermund, K. (2019). The Propositional Evaluation Paradigm: Indirect Assessment of Personal Beliefs and Attitudes. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02385

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in western Europe. *European Journal of Social Psychology*, *25*(1), 57–75. https://doi.org/10.1002/ejsp.2420250106

Remue, J., Hughes, S., De Houwer, J., & De Raedt, R. (2014). To Be or Want to Be: Disentangling the Role of Actual versus Ideal Self in Implicit Self-Esteem. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0108837

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality*, *47*(4), 330–338. https://doi.org/10.1016/j.jrp.2013.02.009

Schimmack, U. (2019). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, 1745691619863798. https://doi.org/10.1177/1745691619863798

Smeding, A., Quinton, J.-C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. *Journal of Personality and Social Psychology*, *111*(6), 817–834. https://doi.org/10.1037/pspa0000063

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*(3), 165–179. https://doi.org/10.1027/1618-3169.51.3.165

Van Dessel, P., Cummins, J., Hughes, S., Kasran, S., Cathelyn, F., & Moran, T. (2020). Reflecting on Twenty-Five Years of Research Using Implicit Measures: Recommendations for their Future Use. *Social Cognition.*

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: Behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience*, *8*(6), 647–653. https://doi.org/10.1093/scan/nss042

Yu, Z., Wang, F., Wang, D., & Bastin, M. (2012). Beyond Reaction Times: Incorporating Mouse-Tracking Measures into the Implicit Association Test to Examine its Underlying Process. *Social Cognition*, *30*(3), 289–306. https://doi.org/10.1521/soco.2012.30.3.289