Journal of Cognitive Psychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/pecp21

The effect of mindfulness and stereotype threat in mental rotation: a pupillometry study

Robert Bauer, Leonardo Jost & Petra Jansen

To cite this article: Robert Bauer, Leonardo Jost & Petra Jansen (2021) The effect of mindfulness and stereotype threat in mental rotation: a pupillometry study, Journal of Cognitive Psychology, 33:8, 861-876, DOI: 10.1080/20445911.2021.1967366

To link to this article: https://doi.org/10.1080/20445911.2021.1967366

Published online: 19 Aug 2021.

JOURNAL OF COGNITIVE PSYCHOLOGY
2021, VOL. 33, NO. 8, 861–876
https://doi.org/10.1080/20445911.2021.1967366

Routledge
Taylor & Francis Group

The effect of mindfulness and stereotype threat in mental rotation: a pupillometry study

Robert Bauer, Leonardo Jost and Petra Jansen

Faculty of Human Sciences, University of Regensburg, Regensburg, Germany

ABSTRACT
The effects of a brief mindfulness intervention and a stereotype threat activation on behavioural performance and cognitive load in a chronometric mental rotation test with cube figures were investigated. 107 participants (55 women and 52 men) were divided into four test groups (mindful or control intervention combined with or without stereotype threat activation). They completed two sets of 150 items each with cube figures of similar complexity and filled out a test of state mindfulness. Cognitive load was measured by changes in pupil dilation using eye tracking. Cognitive load, reaction time, and accuracy were analysed with linear mixed models with the factors time, stereotype threat, mindfulness, angular disparity, and sex. The analyses revealed no sex differences in either of the measurements. Neither stereotype threat nor mindfulness influenced task performance significantly. However, we found that high levels of state mindfulness inhibit stereotype threat to negatively influence task performance and cognitive load.

ARTICLE HISTORY
Received 12 October 2020
Accepted 7 August 2021

KEYWORDS
Eye tracking; pupil diameter; cognitive effort; object-based transformation; sex differences

In everyday activities, spatial abilities play an important role, for instance, in navigation, natural sciences, and engineering. They are related to various abilities like mathematical ability (Xie et al., 2019) or problem solving (Geary et al., 2000), among others. One of these spatial abilities is mental rotation, which describes the ability to mentally rotate 2D or 3D objects fast and accurately (Linn & Petersen, 1985; Shepard & Metzler, 1971) and can be classified according to Uttal et al. (2013) as an intrinsic dynamic ability. In mental rotation research, there is an ongoing discussion about whether and how sex influences the performance. Whereas in psychometric mental rotation tests, sex differences favouring men are more stable (Voyer et al., 1995), this seems not so evident in chronometric mental rotation tests, where reaction time and accuracy are measured in a computerised test (Jansen-Osmann & Heil, 2007). However, if sex differences appear, several explanation possibilities arise, for instance from biological ones (e.g. sex hormone levels), task performance mechanisms (e.g. time limit in tests), or sociocultural factors like spatial toys choice, socioeconomic factors, or also stereotype threat (Heil et al., 2012; Moè & Pazzaglia, 2006). Stereotype threat can be defined as the risk of confirming an existing negative stereotype of the *own* group. This can lead to a disruptive state that undermines the performance in this specific domain (Spencer et al., 2016). In their meta-analysis, Doyle and Voyer (2016) found a significant effect of stereotype threat on women's performance for math but not for spatial tasks. They pointed out that stereotype manipulation might not be as effective when a stereotype is not widely accepted already. Aronson and McGlone (2009) also described that the experience of a stereotype threat can only start with the awareness that one is a member of a negatively stereotyped group. Explicitly in mental rotation, some studies reported significant effects in mental rotation experiments (e.g. Heil et al., 2012; Moè & Pazzaglia, 2006). Stereotype threat can be measured implicitly by its effect on the dependent variables (e.g. Weger et al., 2012) or explicitly by using manipulation questionnaires with Likert-type scales (e.g. Neuburger et al., 2015; Pennington et al., 2019).

CONTACT Robert Bauer robertbauer13@gmail.com Faculty of Human Sciences, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

An important research question is how potential stereotype threats can be reduced. In one study with a mathematical task, Weger et al. (2012) used a brief mindfulness intervention of five minutes to possibly alleviate the effects of stereotype threat. 71 female psychology students participated, who had to complete two mathematical tests with or without a mindfulness intervention, and with or without stereotype activation. The results showed that performance decrements that typically occur under stereotype threat could be reversed after a brief mindfulness induction, which is a short single session of mindfulness, for instance to eat a raisin mindfully (Schofield et al., 2015). In general, longer mindfulness meditation practices over several weeks lead to an improvement in attentional control (Lutz et al., 2008), but also a short meditation practice can impact cognitive control tasks in a different manner depending on the kind of the short meditation form (Colzato et al., 2016). Although mental rotation is a widely investigated cognitive task, to our best knowledge only one study exists where it was shown that participants of both sexes responded faster after mindful than mindless learning (Geng et al., 2011). With an overall sample of 32 students, they found significant performance improvements due to mindful learning with no significant differences between men and women. In contrast, according to the study of Weger et al. (2012), mindfulness might not improve specific cognitive performance per se. Even more, as stated above, they reported an interesting interaction between stereotype threat and mindfulness intervention, that is, that the math performance was only ameliorated in a stereotype threat condition with a mindfulness induction before. The authors explained it with disassociation of the cues linked to social comparison from their threatening value, which would lead to a reinterpretation of the threat because of the brief mindfulness intervention. At least three explanations for this result could be considered: One of them is that mindfulness could improve emotion regulation (Heppner et al., 2008), which is debilitated by stereotype threat (Johns et al., 2008; see also, Weger et al., 2012). Another explanation is that a short mindful induction influences working memory, which might be worsened in stereotypy threat situations (Beilock et al., 2007). Furthermore, one can assume that stereotype threat takes up cognitive resources and increases cognitive effort. Last, mindfulness, emotion regulation, working memory, and stereotype threat could be linked together, as emotion regulation can influence working memory and vice versa (see, Barkus, 2020; Groves et al., 2020). For these reasons, we wanted to investigate the influence mindfulness could have on cognitive effort in a stereotype threat situation.

To our knowledge, Campbell et al. (2018) were the first to implement the physiological correlate of cognitive effort in a chronometric mental rotation task design using pupillometry. Pupillometry describes the measurement of the rapidly changing pupil diameter during cognitive processes. When set in relation to baseline values, the task-evoked changes of the pupil diameter do normally not exceed 0.5 mm (Beatty & Lucero-Wagoner, 2000), and can be used as a “psychophysiological index or correlate of cognitive activity” (Campbell et al., 2018, p. 20) that changes because of task difficulty (Kahneman & Beatty, 1966). Kahneman (1973) used the terms “capacity”, “effort”, and “attention” interchangeably to describe the limited working memory resources available to participants while solving cognitive tasks. Therefore, pupil dilation was used to measure arousal resulting from “cognitive load” and it could differentiate varying difficulty between tasks. Indeed, the findings of Campbell et al. (2018) confirmed that pupil dilation was modulated by angular disparity, with higher angular disparity, that is, more difficult tasks, increasing the pupil diameter. Apart from task difficulty, monitoring the pupil diameter might illustrate other effects. One might be the effect of stereotype threat to deplete working memory capacity (Spencer et al., 2016). Working memory capacity is connected to cognitive tasks such as reasoning (Suess et al., 2002) and inhibitory control (Redick et al., 2011). In summary, in this study, cognitive load was analysed via pupillometric measures to assess task difficulty as well as the limited working memory resources.

Main goal of this study

With mathematical abilities and mental rotation performance being related (Xie et al., 2019) and mental rotation also being partially stereotyped (e.g. Heil et al., 2012; Moè & Pazzaglia, 2006), the reported effects of Weger et al. (2012) in mathematics might also apply to mental rotation tasks. Our goal was to conduct a study designed in close relation to Weger et al. (2012). In their study, they tested the effects of stereotype threat on women

(N = 71) in a 2 × 2 mixed design. Similarly, our participants performed mental rotation tasks and were also assigned to four condition groups with or without mindfulness intervention and with or without stereotype threat. In addition to this, we also tested men to compare the results for both sexes. Even more, we used pupillometry to investigate if the sex differences in mental rotation are due to cognitive load. The following hypotheses were investigated:

In line with the findings of Weger et al. (2012), we predicted an interaction between stereotype threat and mindfulness, with mindfulness only improving the mental rotation performance in the stereotype threat condition for women (Hypothesis 1). We expected these changes to be larger for tasks of higher difficulty, that is, a higher angular disparity between both objects being compared. Mindfulness together with stereotype threat was expected to reduce cognitive load by showing lower pupil dilation in the posttest (Hypothesis 2). Stereotype threat by itself was expected to increase cognitive load (Hypothesis 3). In form of additional hypotheses, we expected an influence of reaction time on pupil dilation due to their close relation indicating cognitive effort. Furthermore, we expected a time effect of angular disparity, that is, greater performance improvements in the posttest for the higher angular disparity. The effect of sex differences will be analysed, too, exploratively.

Methods

Participants

In the study, 125 students (61 women) participated and received study credits. Eye tracking recording errors led the data of 18 participants to be excluded from further analysis. Consequently, 107 students (55 women, mean age (SD) = 20.8 (1.8) years; 52 men, mean age (SD) = 21.4 (2.2) years) form the sample for statistical analysis. In the study of Weger et al. (2012), the interaction effect of mindfulness on stereotype threat for women was detected with a $d = .92$. The required sample size for women was estimated using the software G*power (Faul et al., 2007). To ensure a sufficient sample size, the effect size of $d = .92$ ($f = .46$), a power of 0.90, and an alpha value of .05, a sample size of 52 women was estimated for this study (see Brysbaert & Stevens, 2018). For participation, unrestricted eyesight at close range or corrected eyesight with contact lenses was required, as was no practical experience with mindfulness. Thirty participants reported minor theoretical knowledge with mindfulness, and the others reported no knowledge of mindfulness at all. Participants were sports and movement sciences students, of which 68 reported no experience with mental rotation tasks, 19 reported having participated in one mental rotation task experiment, and 20 reported having participated in two or three mental rotation task experiments prior to this experiment. In total, the test sample had an average of 0.61 mental rotation task experiment experience prior to this study. To mitigate possible experience effects, participants were required not to have participated in any mental rotation task experiment within the six months preceding this experiment. Participants were free from eye injuries and reported no relevant physical or mental limitations. Informed consent was obtained from all individual participants. The experiment was conducted according to the ethical declaration of Helsinki. We communicated all considerations necessary to assess the question of ethical legitimacy of the study.

Conditions

Participants were assigned to one of four test groups in a counterbalanced order. The intervention after completing the first block of mental rotation tasks depended on this assigned group. This intervention was either a five-minute mindfulness intervention, during which the participants listened to recorded instructions by a professional *Mindfulness-Based Stress Reduction* instructor to eat two raisins in a mindful manner, or a control intervention, during which the participants were instructed to eat two raisins in a five-minute break sitting silently in their chair. Following this, participants either did or did not receive a stereotype threat condition from the investigator. The stereotype threat consisted of informing the participant that he or she was taking part in an experiment about “why women/men are always better in solving mental rotation tasks” (always using the respective counterpart to induce the stereotype threat). This was similar to the approach of Weger et al. (2012) to present a standardised brief stereotype threat in one sentence (see also, Colzato et al., 2016). Thirteen men were assigned to each of the four conditions (control/noST, control/ST, mindfulness/noST, or mindfulness/ST). Thirteen women were assigned to the control condition without stereotype threat, and 14 to each of the other three.

Setup

Stimulus presentation and response handling were controlled with Presentation® software (Version 20.1 Build 12.04.17, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com) on a Dell Latitude E5540 Laptop, 15″, 1366 × 768 px, 60 Hz. Below the bottom screen border, a RED250mobile (SensoMotoric Instruments GmbH, 2017) eye tracker running at 250 Hz was applied. iViewRED software (Version 4.4.26.0, SensoMotoric Instruments GmbH, 2017) was used to integrate all screen properties and the position of the tracker relative to the screen. With the iViewX_SDK (Version 4.4.10.0, SensoMotoric Instruments GmbH, 2017), the Presentation® script conducted a 13-point calibration at the start of each run. Calibration accuracy in form of vertical and horizontal dispersion was reported to be lower than 0.3° for all participants. The iViewRED software depicts the distance of the eyes to the screen. All participants were seated close to a table while placing the head on a sturdy chinrest. Then, the laptop was positioned with the screen being at 60 cm from participants' eyes. Two tables at different heights were used. On the lower table, participants had their arms and chinrest. The laptop with the eye tracker was placed on the higher table to prevent participants' movement in form of possible small vibrations to corrupt data collection. The participant used a wired mouse for task response to control for input lag. The laboratory was silent and dimly lit to control for pupil dilation due to light conditions. The only light sources were a ceiling lamp outside the peripheral visual field of the participant and the laptop screen (luminance 1.04 cd/m^2 for all model figures; measured with a spot metre, Chroma Meter CS-100, Minolta Co., Ltd., Japan), resulting in a constant illuminance of 22 lx (measured with a lux metre, testo 540, Testo AG, Germany).

Stimuli

Stimuli consisted of a selection from the stimulus library of Peters and Battista (2008). Ten cube figure models with rotations around x- and z-axis in 45° steps and mirrored/non-mirrored orientation (i.e. figures one to ten with *a* and *b* orientation in the stimulus library) were used with a checkered pattern on a black background (see Figure 1).

In the main experiment, which was divided into two parts, each part consisted of 150 stimuli, resulting in a total of 300 different stimuli. The first 150 stimuli were figures one to five, and the second 150 stimuli were figures six to ten of the stimulus library (Peters & Battista, 2008). The order of the stimuli in each block was first randomised once and then counterbalanced in their order over all participants. On the left side of the screen, every model was presented in orientation a, rotated by 30° in x direction and 15° in z direction, so that the base model for x or z rotation was identical. On the right side of the screen, a rotated and mirrored/non-mirrored stimulus was presented. Stimulus pictures were sized 400px times 400px and presented vertically centred and horizontally positioned 300px to the left or right of the centre of the screen until a response was given. In the practice block, three different models were presented in overall 24 practice stimuli (figures 11 to 13 of the stimulus library), that is, four rotations in x-axis and z-axis for each model, combined with using each possible angular disparity once and half of the trials being mirrored. In the practice session, between the stimuli pairs, participants received feedback for 1000 ms (✓- right, ×- wrong) shown

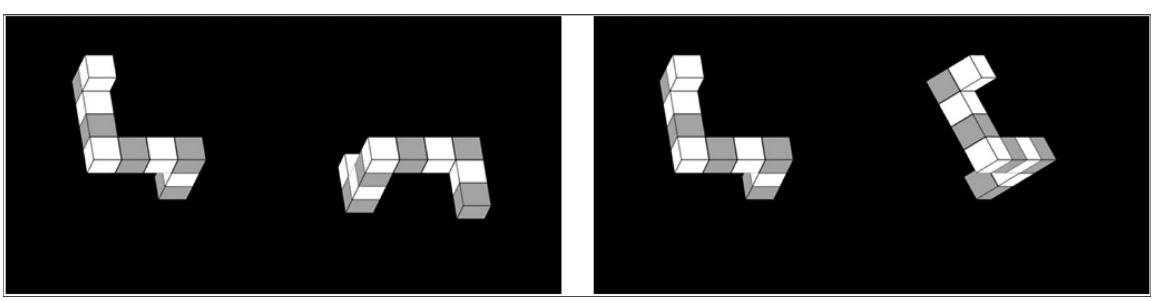

Figure 1. Examples of mental rotation stimuli used in the experiment (selected from the stimulus library of Peters and Battista (2008)). Left: 90° rotation in x-axis. Right: 45° rotation in z-axis.

at the centre of the screen. In experimental sessions, a fixation cross ("+") was shown there for 500 ms.

Pupil diameter

Before the practice block, baseline measurements were conducted to later calculate comparable pupil diameter data. All ten models were presented in randomised order and only for the 0° non-mirrored condition. Participants were instructed that they were about to see some pictures, which they only had to look at. The illumination of the screen for all individual figures is presented once and can further be used as a controlled baseline for light condition influences (Holmqvist et al., 2011). To address these two aspects of the baseline data, participants looked at each figure for six seconds. For the time interval from three to five seconds after stimulus onset, the average pupil diameter of each eye was calculated for each figure. This time range was chosen to reduce the effects of the onset of new visual stimuli and to control for general physiological variations. The response in pupil dilation typically occurs within the first few hundred milliseconds after stimulus onset, peaks around one to two seconds after stimulus onset and continues asymptotically until it returns to baseline values (Beatty & Lucero-Wagoner, 2000; Nieuwenhuis et al., 2011). For this reason, longer baseline and stimulus durations are recommended, where participants must maintain fixation (e.g. Hayes & Petrov, 2016). In the main experiment, the maximum pupil diameter for each eye was recorded. Then, the difference between this value and its respective baseline value was calculated for each item (Beatty & Lucero-Wagoner, 2000). Based on the fixation duration of each eye, their weighted mean of dilation difference was calculated. Unscrambled pictures were chosen over scrambled ones due to most of the screen (background) being black with the latter, resulting in nearly completely black pictures.

Toronto mindfulness scale

As a manipulation check, whether the mindfulness manipulation had the desired effect, we used the Toronto Mindfulness Scale (TMS; Lau et al., 2006) as a measure of state mindfulness. The TMS was translated to German and included 13 items with five answer alternatives (0 = not at all; 4 = very much). For each participant, we computed summed scores for all items (see Lau et al., 2006). To prevent priming the participants to focus on mindfulness behaviour prior to the treatment (either control or mindfulness condition), the TMS was only done once after the second mental rotation task block. Thus, these values served as a check, whether the treatments showed an overall different effect between the groups. In this study, the Cronbach's alpha of the TMS score was 0.78.

Procedure

The experiment was a single session and lasted between 40 and 50 minutes, depending on participants' speed to complete all items. Upon arrival, the participant was informed about the study protocol (questionnaire, calibration, a mental rotation task block, a short middle part, another mental rotation task block, questionnaire, end) and the *general intention* of the study (i.e. investigation of mental rotation with eye tracking; no information was given about the conditions or pupillometry), then read and signed the informed consent, and filled out a demographic questionnaire. After positioning the participant with the chinrest at the table and a brief explanation of the test protocol, the calibration and baseline were run. The practice session with feedback followed, which was introduced by a digitally presented instruction. Participants used the right hand for mouse handling and received written instructions to press the left mouse button, if the stimuli could be rotated into congruence (non-mirrored, "same"), and the right mouse button, if the two stimuli were mirrored ("different"), and to answer as quickly and precisely as possible. After completing all practice trials, the calibration and the first part of the main experiment were run. Depending on the assigned test group, an intervention (control, control/ST, mindfulness, or mindfulness/ST) followed. After another calibration, the second part of the experiment was conducted. Following this part, the participant filled out the TMS (Lau et al., 2006). In the end, the participant was debriefed. All verbal instructions were standardised.

Study design

To analyse cognitive performance, the dependent variables are reaction time (RT) and accuracy (ACC). The difference between the maximum pupil diameter and the respective baseline pupil diameter for each trial (PD) serves as a measure of cognitive load. Independent variables are pre-post experiment setup (PP), mindfulness intervention (MF), stereotype threat (ST), sex, angular disparity (DEG), and their respective interactions. DEG describes the angular disparity between the two figures shown on the screen. Since the left image was always presented with 0° rotation, the angular disparity depicts the rotation in degrees of the right image. The PD analysis includes reaction time as an additional fixed effect. Regarding MF and sex, the overall TMS score is analysed as a measure for state mindfulness.

Data processing

For the behavioural data, outliers were determined by a deviance of more than three standard deviations from the mean reaction time of all stimulus pairs with the same rotation angle and were excluded from all analyses. Because angular disparity is not defined for mirrored responses in cube figures (Jolicœur et al., 1985; Shepard & Metzler, 1971), only non-mirrored stimulus pairs were analysed, and reaction time was additionally only analysed for correct responses. Using the SMI software BeGaze 3.7, build 58 (SensoMotoric Instruments GmbH, 2017), the implemented Raw Data Export with a velocity dependent algorithm (peak velocity threshold = 40°/s, min. fixation duration = 50 ms, peak velocity between 20 and 80% of saccade length) was used to determine fixation, blink, and saccade events.

The data was further processed in R (version 3.5.1; R Core Team, 2018) to obtain valid pupil diameter data. If a trial started with a fixation, its starting time was set to 0 (keeping the pupil diameter). If the ending time of the eye tracking data was longer than the reaction time received from Presentation® for an item, events with the starting time being after the reaction time ending got erased (including the respective pupil diameter). Overlapping events at the trial end got cut to the reaction time (keeping the pupil diameter). If the eye tracking time ended more than 200 ms before the reaction time, the trial was treated as an outlier. The baseline average pupil diameter is the weighted mean (by time) of all event fixations during a two-second interval for each eye. The maximum pupil diameter takes the maximum value of all event fixations that have a pupil diameter and is then calculated as a weighted mean (by fixation duration) of the two eyes. In the baseline, the values did neither differ between men (mean (SD) = 4.35 (0.76)) and women (mean (SD) = 4.49 (0.66), $\chi^2(1) = 1.14$, $p = .29$), nor between eyes (left, mean (SD) = 4.43 (0.70); right, mean (SD) = 4.42 (0.72), $\chi^2(1) = 0.93$, $p = .34$), but did differ between models, $\chi^2(9) = 153.69$, $p < .001$. Based on the same reasoning for behavioural data analysis in mental rotation tasks, our pupil dilation analysis was conducted with a dataset filtered similarly to the reaction time dataset, that is, with the maximum pupil diameter excluding mirrored items and wrongly answered ones. This is also in line with the database selection used by Campbell et al. (2018) for pupillometric measurements with mental rotation tasks.

Statistical analysis

Statistical analysis was performed according to Bauer et al. (2021), and Jost et al. (2019) using *Linear Mixed-Effects Models using "Eigen" and S4* package (*lme4,* version 1.1-21; Bates et al., 2015b) in R (R Core Team, 2018). Reaction time and pupil diameter were analysed using linear mixed models, and accuracy was analysed using generalised linear mixed models with a binomial distribution. Model parameters were estimated by maximum likelihood estimation using the *bobyqa* algorithm wrapped by *optimx* package (version 2018-7.10; Nash & Varadhan, 2011) as optimizer. Model fit was calculated using likelihood ratio tests to compare models with and without the fixed effect of interest. The resulting *p*-values were compared to a significance level of .05. For multiple comparisons of the same variables, the significance level was Bonferroni corrected. Visual inspection of residual plots did not reveal deviations from homoscedasticity or normality in any model. For the significant effects and main effects, we report the unstandardised effect sizes, and the confidence intervals that were calculated using parametric bootstrapping with 1000 simulations, in line with recommendations of Baguley (2009), and Pek and Flora (2018).

Model building was based on the research of Barr et al. (2013), and Bates et al. (2015a), starting with a model with random intercepts and slopes for every appropriate fixed effect and reducing the model complexity by dropping non-significant variance components. Non-significant fixed effects were further removed stepwise from the model, that is, effects that least decreased the model fit were removed first, and a model only containing significant fixed effects remained. Then, non-significant effects were tested for improving the model fit by including them in the resulting model. Also, significant effects were tested for worsening the model fit by exclusion of the effect. The main effects for significant interactions were tested separately by splitting the interaction (also see Jost & Jansen, 2020). The resulting models for each parameter are described in the results section. In the tables, all results (i.e. partial interactions, test statistics, and confidence intervals) are depicted for effects with $p < .05$. For $.1 > p > .05$, the test statistics are also depicted. Furthermore, for $p > .1$, only the test statistics for hypothesis-relevant effects (interactions of MF, ST, PP, SEX, and DEG, and selected main effects) are depicted. That means, all other effects in the models resulted in $p > .1$. The TMS scores were analysed with Mann-Whitney tests in R (R Core Team, 2018). In all graphs, the depicted data are combined for both sexes for better readability. All data were visualised using *ggplot2* package (Wickham, 2016) in R (R Core Team, 2018).

Results

Toronto mindfulness scale scores

As a measure of state mindfulness, the total TMS scores were analysed. Visual inspection and test for normality (Shapiro-Wilk) revealed the data to be significantly different from a normal distribution. A Mann-Whitney test indicated that participants' scores with MF (Mdn = 30.5) did not differ significantly to those without MF (Mdn = 27), $U = 1693.5$, $p = .102$. A similar test also revealed no significant difference between men (Mdn = 28) and women (Mdn = 29), $U = 1473$, $p = .791$.

Reaction time

As shown in Figure 2, the graphs of the reaction times are closely related. All graphs show a positive slope with increasing angular disparity.

For reaction time, model construction resulted in a model with random intercepts and slopes for PP and DEG by participant. PP*MF*ST*SEX*DEG and all respective interactions and main effects were analysed as fixed effects. Significant differences were found for DEG*SEX*ST and DEG*PP (see Table 1). The interaction DEG*PP shows a significant increase in reaction time by DEG, with higher increases in pretest than posttest. Partial interactions show a significant decrease in reaction time for SEX*ST and DEG*ST. Reaction time increased significantly by DEG.

Accuracy

In Figure 3, the graphs of the proportion of correct answers are closely related. The visually interconnected points of the means show, that all graphs are similar at zero degrees (with large standard deviations) and show a negative slope with increasing angular disparity.

Model construction for ACC resulted in a model with random intercepts and slopes for PP and DEG by participant. PP*MF*ST*SEX*DEG and all respective interactions and main effects were analysed as fixed effects. Significant differences were only found for DEG (see Table 2). Accuracy decreased significantly by DEG.

Pupil diameter

The task-evoked pupil responses are depicted in Figure 4 and show an inclination of all graphs with increasing angular disparity. Here, the control group shows constantly lower values than the other conditions. The graphs of the other conditions lie closer together.

The model building for the pupil diameter differences resulted in a model with random intercepts and slopes for PP and RT by participant. PP*MF*ST*SEX*DEG, RT, and all respective interactions and main effects were analysed as fixed effects. The pupil diameter increased significantly by PP*SEX, DEG*ST, and RT (see Table 3). Pairwise comparisons showed significant differences between the groups for PP (woman/man), but not for SEX (pretest/posttest). The effect of RT in the model indicates an increase of the pupil size change with increasing reaction times.

Discussion

Toronto mindfulness scale scores

All participants showed overly high scores for state mindfulness in the experiment (see Lau et al., 2006). The scores are similar to the ones Weger et al. (2012) reported for their mindfulness groups. Interestingly, even the participants without mindfulness intervention in this study showed high levels of state mindfulness in the scores, in contrast to significantly lower scores of the respective group in Weger

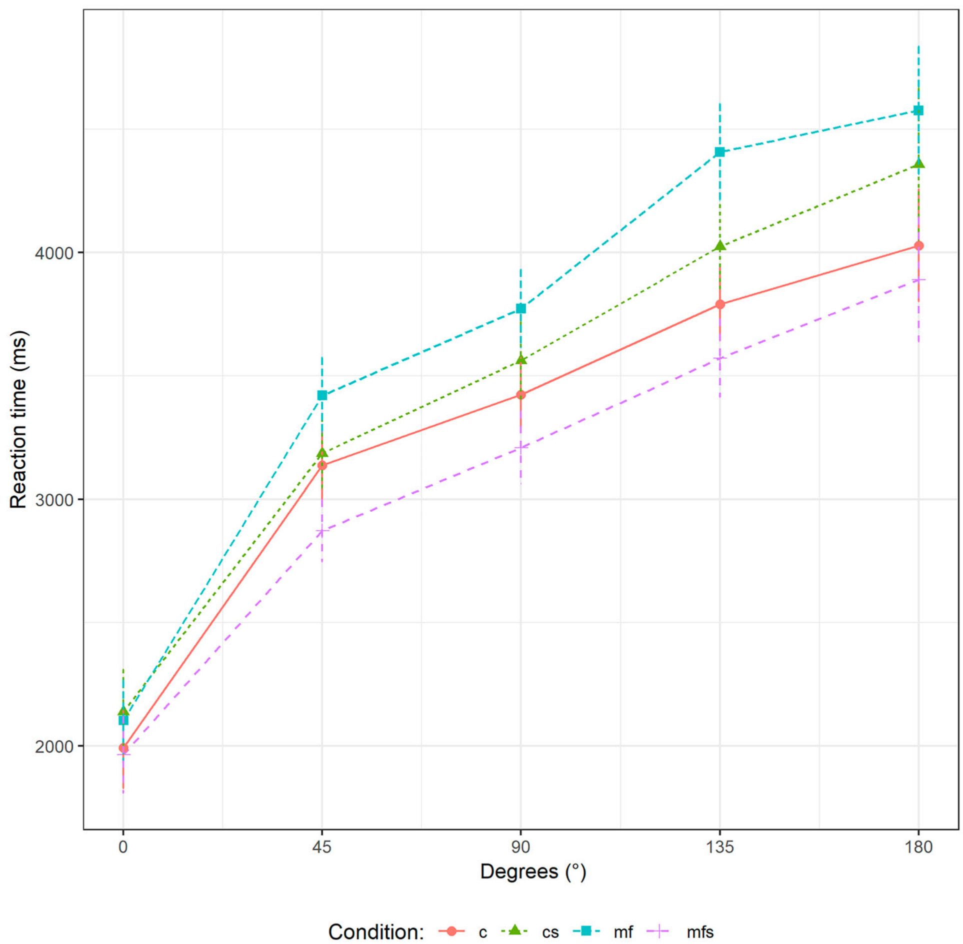

Figure 2. Reaction time plotted against angular disparity and condition (control, C; stereotype threat, CS; mindfulness, MF; mindfulness and stereotype threat, MFS) for the combined data of both sexes.

Table 1. Statistical analysis of reaction time (in seconds).

Variable	Estimate	SE	Test Statistic	*p*	95% CI
Intercept	1.89	0.15			1.59, 2.21
Hypothesis-relevant Effects					
MF*ST*PP*SEX*DEG			$\chi^2(1) = 0.01$	.927	
MF*ST*PP*SEX			$\chi^2(1) = 2.20$	.138	
MF*ST*PP*DEG			$\chi^2(1) = 1.14$	.286	
MF*ST*PP			$\chi^2(1) = 0.09$	.771	
MF*PP			$\chi^2(1) = 0.01$	.914	
ST*PP			$\chi^2(1) = 2.39$	.122	
DEG*SEX			$\chi^2(1) = 0.58$	.455	
DEG*PP(pre-post)	0.15	0.05	$\chi^2(1) = 8.11$	.004	0.04, 0.24
DEG(0°)*PP(pre-post)	0.22	0.20			−0.17, 0.60
DEG*PP(post)	1.12	0.12			0.90, 1.36
DEG(100°)	1.17	0.06	$\chi^2(1) = 164.10$	<.001	1.05, 1.28
Additional Effects					
DEG*SEX*ST			$\chi^2(1) = 5.25$	.022	
DEG*(ST-noST)	−0.24	0.11	$\chi^2(1) = 4.12$	.041	−0.49, −0.03
DEG*(noST)	1.21	0.09			1.04, 1.38
DEG(0°)*(ST-noST)	0.39	0.20			−0.03, 0.81
SEX(w-m)*(ST)	−0.55	0.27	$\chi^2(1) = 3.94$	.047	−1.13, 0.00
SEX(w-m)*(noST)	0.21	0.19			−0.16, 0.59
SEX(m)*(ST)	0.29	0.19			−0.11, 0.67
MF*ST*DEG			$\chi^2(1) = 3.05$	.081	
MF*PP*DEG			$\chi^2(1) = 3.21$	.073	

Intercept in this model represents the estimate for SEX = men, no stereotype threat, angular disparity (DEG) = 0°, pre-posttest (PP) = post, and no mindfulness (no MF). Effect of SEX represents the difference between women (w) and men (m). Effect of ST represents the difference between stereotype threat (ST) and no stereotype threat (no ST). Effects of angular disparity (DEG) represent changes of 100°.

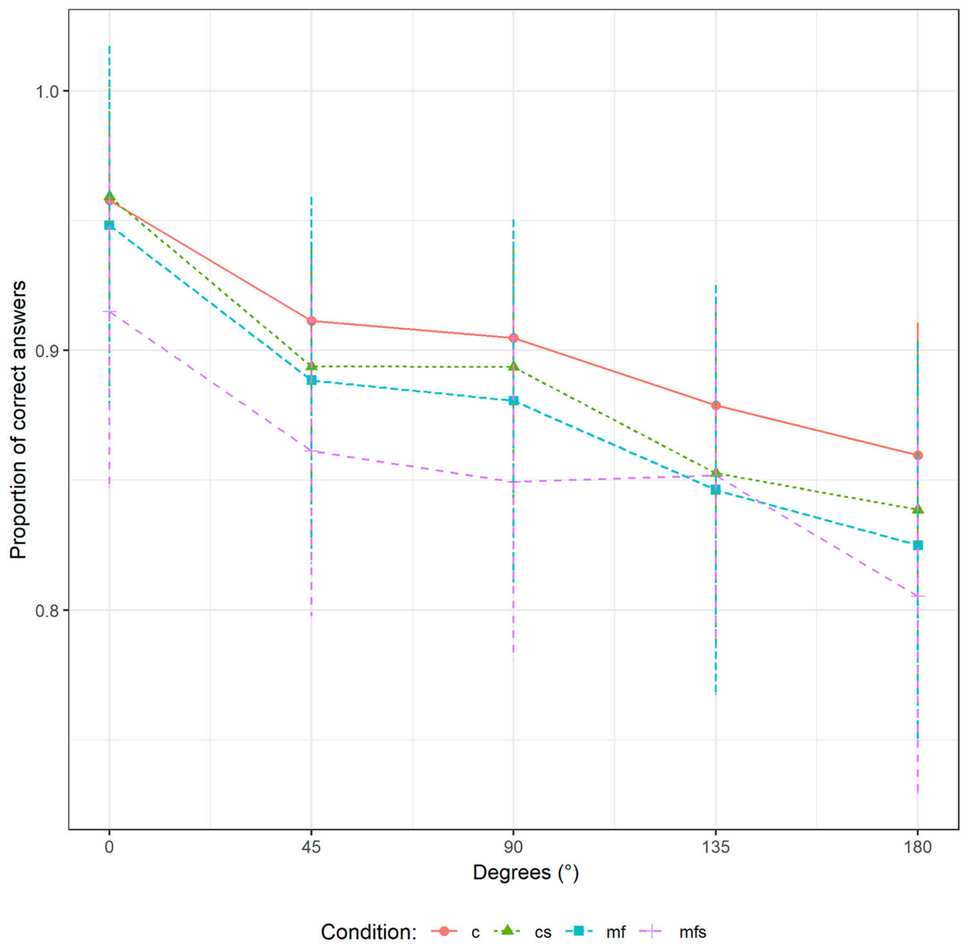

Figure 3. Accuracy plotted against angular disparity and condition (control, C; stereotype threat, CS; mindfulness, MF; mindfulness and stereotype threat, MFS) for the combined data of both sexes.

et al. (2012). It is important to note that the Toronto Mindfulness Scale scores indicate that the brief mindfulness intervention did not have a significant main effect on state mindfulness, which was not the expected effect to form the basis for the rest of the analysis. Thus, this must be considered regarding the following interpretations of the results.

Mental rotation performance

Based on our hypothesis, the focus was on the interaction between mindfulness, stereotype threat, and pre-post testing, which would illustrate the changes resulting from both treatments. We expected these changes to be larger for tasks of higher difficulty, which is why angular disparity was added to the modeling. Additionally, we examined exploratively whether sex had an influence.

No significant interactions between pre-post testing, mindfulness, and stereotype threat emerged in the analysis for reaction time and accuracy, which indicates that no effects regarding the treatments of stereotype threat and mindfulness can be observed. For accuracy, the resulting model is even more reduced than the one for reaction time, with angular disparity being the only significant variable. A possible explanation is that in mental rotation tasks, a speed-accuracy-tradeoff always happens. Here, accuracy undergoes smaller changes than reaction time and is, therefore, less influenced by different variables (see e.g. Hertzog et al., 1993; Wickelgren, 1977).

In general, all groups improved in their reaction times from pretest to posttest. However, no significant differences in the performance changes manifested between the four condition groups. This statistically renders both treatments, that is, mindfulness and stereotype threat, and their respective combinations, as overall ineffective concerning all dependent variables in this study. Additionally, the results show no significant sex differences in

Table 2. Statistical analysis of (logarithmic odds of) accuracy.

Variable	Estimate	SE	Test Statistic	*p*	95% CI
Intercept	3.62	0.19			3.23, 4.06
Hypothesis-relevant Effects					
MF*ST*PP*SEX*DEG			$\chi^2(1) = 0.01$	.918	
MF*ST*PP*SEX			$\chi^2(1) = 1.00$	.317	
MF*ST*PP*DEG			$\chi^2(1) = 2.31$	.129	
MF*ST*PP			$\chi^2(1) = 0.03$	.871	
MF*PP			$\chi^2(1) = 0.01$	.926	
ST*PP			$\chi^2(1) = 1.21$	.272	
DEG*SEX			$\chi^2(1) = 1.55$	.214	
DEG(100°)	−1.02	0.09	$\chi^2(1) = 90.95$	<.001	−1.21, 0.83
Additional Effect					
MF*PP*DEG*SEX			$\chi^2(1) = 3.63$	.057	

Intercept in this model represents the estimate for the logarithmic odds at angular disparity (DEG) = 0° for SEX = men, no stereotype threat, pre-posttest (PP) = post, and no mindfulness (no MF). Effect of SEX represents the difference between women (w) and men (m). Effect of ST represents the difference between stereotype threat (ST) and no stereotype threat (no ST). Effects of angular disparity (DEG) represent changes of 100°.

overall behavioural performance. The main effects and interactions for both accuracy and reaction time do not indicate any influence of sex in the models. This is in line with other chronometric mental rotation studies. For instance, Voyer et al. (2006), and Jansen-Osmann and Heil (2007) also reported no sex differences in mental rotation performance of 3D cube figures.

Altogether, our results did not confirm our hypothesis that task performance would be better

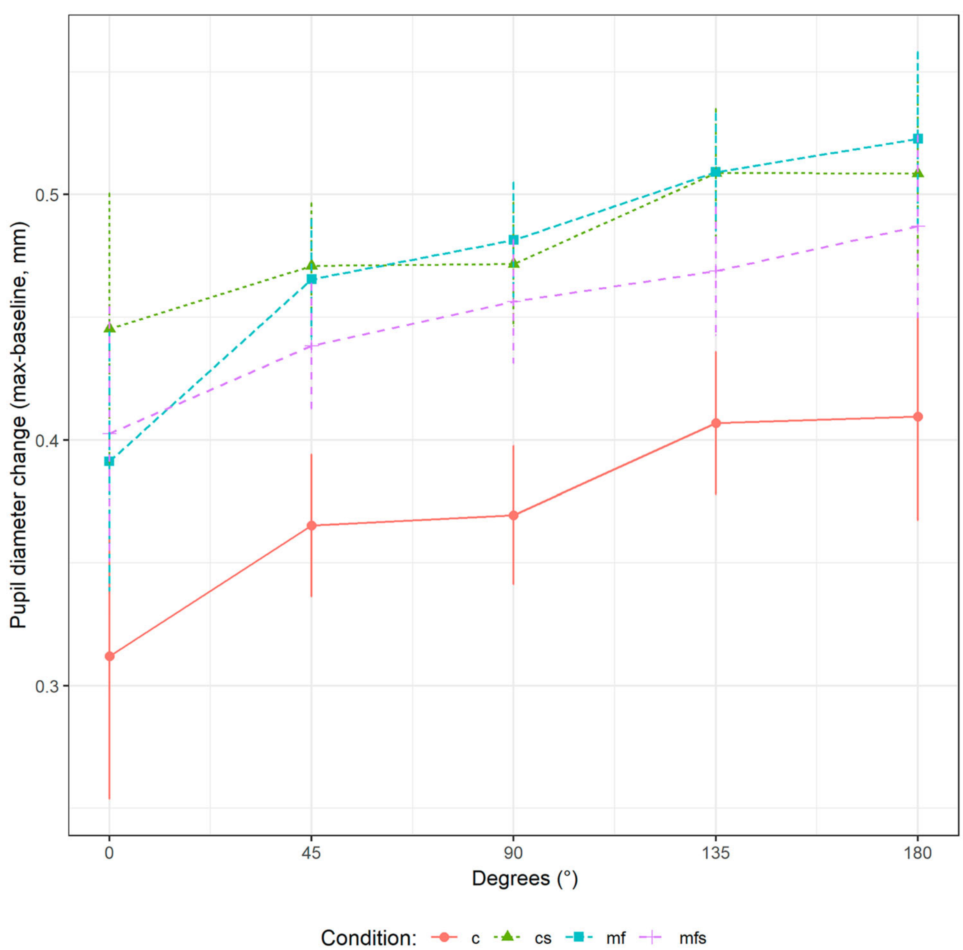

Figure 4. Changes of pupil size (max-baseline) plotted against angular disparity and condition (control, C; stereotype threat, CS; mindfulness, MF; mindfulness and stereotype threat, MFS) for the combined data of both sexes.

Table 3. Statistical analysis of pupil diameter (in 10^{-1} mm).

Variable	Estimate	SE	Test Statistic	p	95% CI
Intercept	3.15	0.53			1.05, 5.00
Hypothesis-relevant Effects					
MF*ST*PP*SEX*DEG			$\chi^2(1) = 0.06$	.812	
MF*ST*PP*SEX			$\chi^2(1) = 0.30$	.587	
MF*ST*PP*DEG			$\chi^2(1) = 0.30$	.585	
MF*ST*PP			$\chi^2(1) = 0.31$	.580	
MF*PP			$\chi^2(1) = 1.98$	.159	
ST*PP			$\chi^2(1) = 2.15$	.143	
DEG*SEX			$\chi^2(1) = 0.87$	.366	
RT	0.33	0.02	$\chi^2(1) = 166.46$	<.001	0.30, 0.37
DEG(100°)	0.23	0.52	$\chi^2(1) = 22.97$	<.001	0.14, 0.37
PP(pre-post)*SEX(w-m)	−0.88	0.43	$\chi^2(1) = 3.94$	.047	−1.72, −0.04
PP(pre-post)*SEX(m)	3.11	0.38			2.34, 3.89
PP(post)*SEX(w-m)	−0.10	0.61			−1.22, 1.13
Split Interaction Main Effects					
PP	2.65	0.31	$\chi^2(1) = 25.93$	<.001	2.05, 3.26
PP(w)	2.23	0.38	$\chi^2(1) = 20.63$	<.001	1.50, 2.99
PP(m)	3.11	0.37	$\chi^2(1) = 27.40$	<.001	2.34, 3.84
SEX	−0.74	0.52	$\chi^2(1) = 1.89$	.169	−1.72, 0.26
SEX(pre)	−0.83	0.54	$\chi^2(1) = 2.33$	.127	−1.93, 0.23
SEX(post)	−0.13	0.60	$\chi^2(1) = 0.05$	.827	−1.26, 1.07
Additional Effects					
DEG*(ST-noST)	−0.28	0.09	$\chi^2(1) = 9.29$	.002	−0.46, −0.09
DEG*(noST)	0.38	0.07			0.23, 0.51
DEG(0°)*(ST-noST)	0.15	0.52			−0.92, 1.25
DEG*ST*PP			$\chi^2(1) = 2.97$	.085	

Intercept in this model represents the estimate for SEX = men, no stereotype threat, angular disparity (DEG) = 0°, pre-posttest (PP) = post, and no mindfulness (no MF). Effect of SEX represents the difference between women (w) and men (m). Effect of ST represents the difference between stereotype threat (ST) and no stereotype threat (no ST). Effects of angular disparity (DEG) represent changes of 100°.

after the mindfulness induction in the stereotype threat condition. Additionally, all results apply to both sexes in the same way in this study. This stands in contrast to the findings of Weger et al. (2012), on which our hypothesis was based. With an applied stereotype threat, the authors reported better performance due to a brief mindfulness intervention. With our intervention being recorded by a professional instructor and lasting the same amount of time, that is, five minutes, the intervention aspects of both studies can be expected to be very similar and not likely the cause for the differences.

First of all, the missing stereotype threat effect is astonishing, because in general, laboratory effects of stereotype threats are robust, with moderate to small effect sizes (for a review, see Spencer et al., 2016). Nguyen and Ryan (2008) reported subtle cues to trigger larger stereotype threat effects for women in math than did blatant or moderate ones. However, in order to remove the threat, blatant strategies reduced stereotype threat effects sizes more than subtle strategies did. In this study, the stereotype threat cue could have been too unconcealed to elicit a sufficiently strong effect. Additionally, the mindfulness intervention was not targeted on removing the stereotype threat per se, and cannot be described as a blatant strategy to reduce it. However, its effects were expected to reduce stereotype threat effects as a consequence. The findings of Nguyen and Ryan (2008) also indicated that stereotype threat effects for women in math were largest among women who moderately identified with math. Spencer et al. (2016) described this as an unfortunate aspect of stereotype threat: that the most dedicated and caring people are affected most by negative stereotypes. Participants' performance diminishes more under stereotype threat, when they identify with the stereotype domain, with their performance in the domain being self-relevant (Spencer et al., 2016). Here, mental rotation tasks might be different from math tasks. At school, people are confronted with math for several years and establish an attitude toward the domain and their respective abilities in it. Specifically, women's performance can suffer through stereotype threat, because they are "aware of the widely held stereotype threat in our Western culture that women are not as good at math as men" (Doyle & Voyer, 2016, p. 10; see also, Matlin, 2011; Nosek et al., 2002; Tartre & Fennema, 1995). The participation in a mental rotation experiment might not cause similarly strong effects, because people would not be able to identify and range their own performance level in this domain right away. Even more, after the practice trials, no more indication for correct or wrong answers was given in this experiment. The participants were therefore oblivious to how well they were currently performing, in contrast to mathematical tasks, where the own competence might be evaluated to a greater extent. Hence, low levels of stereotype threat effects regarding mental rotation task performance might be due to the unfamiliarity with the test and the lack of knowledge about one's own competence compared to others. Here, the difference between psychometric and chronometric mental rotation experiments might also have an effect. Sex differences have been reported more in psychometric experiments than in chronometric studies (Voyer et al., 1995). It is possible, that the chronometric tests do not only show smaller sex differences in performance but are also perceived as less stereotypical. However, it is generally the mental rotation ability and not the test design that is assumed to be stereotypical. For example, Hausmann et al. (2009) activated and found gender stereotypes for the item "imaging abstract objects and rotate them mentally", which should apply to both psychometric and chronometric tests. However, the activated stereotypes only resulted in performances difference in a psychometric test and not in other spatial abilities (with no direct comparison with a chronometric test). Thus, while it is possible that stereotypes exist for spatial abilities in general, these might transfer to performance differences only in specific tests. Furthermore, it has been shown that not all women suffer from stereotype threat activation in mental rotation. In one study, this effect was only visible in women with a feminine gender role orientation (Tempel & Neumann, 2016). For future research, it might be worth analysing the gender role orientation. It is also possible that task difficulty was not sufficiently high enough. Other studies have shown that only very difficult tasks were affected by stereotype threat (Spencer et al., 1999; Blascovich et al., 2001; Keller, 2007).

Another possible explanation for our interventions not to show any effects might be our test sample. Facing stereotype threat, strong coping abilities are associated with resilience. Inzlicht et al. (2006a, 2006b) reported high self-monitors not to underperform under stereotype threat. The same applies to people with a high coping sense of humour (Ford et al., 2004). The participants in this study were students of sports and movement sciences. It is possible that this encompasses better coping mechanisms, that is, showing stronger minds in terms of failure (Seeley & Gardner, 2003) and therefore not underperforming under stereotype threat (Inzlicht et al., 2006a, 2006b). The results of the analysis of the Toronto Mindfulness Scale scores are in line with these interpretations (see above). The high TMS scores might indicate one reason, why the mindfulness treatment was statistically insignificant. Having all test groups on high scores already, the extent, to which an intervention might elevate state mindfulness, might get rendered insignificant.

In line with the paradigm for chronometric mental rotation tasks, changes in angular disparity significantly influenced all dependent variables. Higher angular disparity between the two pictures resulted in higher reaction times and lower accuracy. Differences between the reaction times of pre- and post-test can be observed, with higher values in the pretest. The interaction between pre-post and angular disparity shows a higher reaction time increase for each degree in the pretest. In addition, the performance increase from pre- to posttest is higher at greater angles. These outcomes can be explained as typical learning effects in a pre-post experimental design, and have more impact for more difficult tasks, that is, at higher angular disparities.

The role of cognitive effort in the mental rotation task

Before addressing the hypotheses, we would like to elaborate on possible problems in the analysis of pupil size means. In our last study (Bauer et al., 2021), we exploratively conducted both a baseline correction similar to this study and a trial-dependent baseline correction (see, Mathôt, 2013; Mathôt et al., 2018; Mathôt et al., 2015), which lead to the same results. Both types of baseline correction have advantages (see, Holmqvist et al., 2011; Mathôt et al., 2018). However, by using short breaks between the trials (i.e. fixation point sections), the risk of possible carry-over effects should be of concern for mental rotation experiments. Considering our fixation point duration of 500ms in this study, we used the described baseline correction, as it should be slightly better regarding possible carry-over effects. In our last paper, we pointed out that future experimenters should (1) keep showing the task and measuring pupil dilation even after the response for at least 500 ms, and (2) increase the break between trials to at least two seconds (see, Bauer et al., 2021).

Regarding to the analysis of pupil diameter differences as a measure for cognitive load, we expected reduced load for the mindfulness and stereotype threat condition. The significant interaction of sex and pre-post testing indicates that a cognitive improvement in form of lower cognitive load was higher for men than for women, which might be due to men having started at a higher level and having more room for improvement than women had (i.e. a ceiling effect affecting the women more). These differences were not significant in the main effect for sex, however. The main effect for pre-post testing shows an improvement of both sexes over the two experiment parts. Furthermore, the interaction solely shows a sex difference for the temporal effects of pre-post testing, for instance through greater learning improvements in men, but does not seem to interact with any of the treatments of the experiment. Neither mindfulness nor stereotype threat had a statistically significant impact on the cognitive load for both sexes. This finding is congruent with our results for behavioural performance.

Applied stereotype threat leads to the motivation in affected people to disconfirm negative stereotypes that target their social identity or at least to avoid confirming it (for a review, see Spencer et al., 2016). This inflicts a pressure to succeed that can negatively influence the performance through three mechanisms: mere effort, working memory depletion, and conscious attention to automatic processes (Spencer et al., 2016). Based on this, stereotype threat should indeed increase cognitive load in laboratory experiments, which we also predicted in our hypothesis. As discussed above, our stereotype threat application as well as our test sample with already high state mindfulness might not have been ideal to establish an impactful stereotype threat on the participants. However, for this study, our third hypothesis was rejected.

In the model, changes in pupil dilation did not get fully explained by variations in angular disparity. Reaction time itself still predicted a significant portion of the pupil diameter in the statistic model, which is in line with the findings of our last study (Bauer et al., 2021). That is, both reaction time and angular disparity had an impact on cognitive load, but none of them alone seemed to be sufficient to describe the connection between each other, and to account for task difficulty. The relationship between task performance and cognitive load is also modulated by task difficulty, which is hard to control (Coyne et al., 2017). Therefore, the relationship between difficulty and cognitive load is probably not linear (Bauer et al., 2021).

In Campbell et al. (2018), women showed higher cognitive load – or used more attentional resources – than men in mental rotation tasks of abstract figures. Our findings did not confirm this, as no differences emerged. As a conclusion, both sexes were able to solve mental rotations tasks with cube figures of similar difficulty with the same cognitive effort. Our last study showed similar results regarding the sex differences in cognitive effort (Bauer et al., 2021). Hence, our findings in this study provide further evidence that men and women need the same cognitive effort for the solution of cube figure tasks. The possible effects might get analysed further in future studies using different kinds of stimuli and stereotype threat application to be further illustrated.

Limitations and conclusion

In pupillometry studies, the pupil foreshortening error should be of concern to have an influence on the acquired data, especially when looking at areas, which are further away from the screen centre (Hayes & Petrov, 2016). In our mental rotation tasks, all analysed image pairs are shown at the same positions on the screen. Possible pupil foreshortening errors were not expected to systematically differ between conditions, and were neglectable for that reason (see, Mathôt et al., 2018).

We conducted this experiment with close similarity to the study of Weger et al. (2012). They presented the stereotype threat in one sentence in the instructions and did not include a manipulation check after that. We replicated their approach, taking also into account the already long participation time for the experiment. Retrospectively, we would recommend using a manipulation check questionnaire, as for instance Neuburger et al. (2015), and Pennington et al. (2019) did.

As mentioned in the discussion, the experimental design led to possible carry-over effects of cognitive effort, which must be taken into account for interpretation of the results.

No sex differences appeared in either of the measurements, which contributed to the different results in chronometric mental rotation test concerning sex differences (see Jansen-Osmann & Heil, 2007). This study replicated the experimental design of Weger et al. (2012) to scrutinise the impact of stereotype threat and mindfulness and showed that the here induced stereotype threat had no impact on mental rotation task performance, which was successfully applied in former mental rotation studies, but with psychometric tests (Heil et al., 2012). However, we found that high levels of state mindfulness inhibit stereotype threat to negatively influence task performance and cognitive load. This must be investigated in more detail.

Acknowledgements

We thank Prof. Dr. Kenneth Holmqvist (University of Lund) for his professional advice and Prof. Dr. Mark W. Greenlee (University of Regensburg) for providing us with a spot meter. Data collection was supported by Alexander Kalus, Stefan Gruber, and Philipp Hofmann (all from University of Regensburg). We thank Dr. Michael Pilling (Oxford Brookes University) for his fast responses and work, and the anonymous reviewers for their insightful suggestions and careful reading of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability

The data that support the findings of this study are openly available in "zenodo" at https://doi.org/10.5281/zenodo.4081753. The Presentation® (Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com) and R (R Core Team, 2018) code used for this study is available from the corresponding author upon reasonable request.

ORCID

Robert Bauer http://orcid.org/0000-0003-1034-2007
Leonardo Jost http://orcid.org/0000-0003-3508-4315
Petra Jansen http://orcid.org/0000-0002-0973-2158

References

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Aronson, J., & McGlone, M. S. (2009). Stereotype and social identity threat. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 153–178). Taylor & Francis Group.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Barkus, E. (2020). Effects of working memory training on emotion regulation: Transdiagnostic review. *PsyCh Journal*, *9*(2), 258–279. https://doi.org/10.1002/pchj.353

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015a). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, R., Jost, L., Günther, B., & Jansen, P.. (2021). Pupillometry as a measure of cognitive load in mental rotation tasks with abstract and embodied figures. *Psychological Research*, 1–15. https://doi.org/10.1007/s00426-021-01568-5

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (Vol. 2, pp. 142–162). Cambridge University Press.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, *136*(2), 256–276. https://doi.org/10.1037/0096-3445.136.2.256

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science*, *12*(3), 225–229. https://doi.org/10.1111/1467-9280.00340

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Campbell, M. J., Toth, A. J., & Brady, N. (2018). Illuminating sex differences in mental rotation using pupillometry. *Biological Psychology*, *138*, 19–26. https://doi.org/10.1016/j.biopsycho.2018.08.003

Colzato, L. S., van der Wel, P., Sellaro, R., & Hommel, B. (2016). A single bout of meditation biases cognitive control but not attentional focusing: Evidence from the global–local task. *Consciousness and Cognition*, *39*, 1–7. https://doi.org/10.1016/concog2015.11.003

Coyne, J. T., Foroughi, C., & Sibley, C. (2017). Pupil diameter and performance in a supervisory control task: A measure of effort or individual differences? In *Proceedings of the Human Factors and Ergonomics Society* (Vol. 2017-October, pp. 865–869). Human Factors an Ergonomics Society Inc. https://doi.org/10.1177/1541931213601689.

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences*, *47*, 103–116. https://doi.org/10.1016/j.lindif.2015.12.018

Faul, F., Erdfelder, E., Lang, A–G, & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ford, T. E., Ferguson, M. A., Brooks, J. L., & Hagadone, K. M. (2004). Coping sense of humor reduces effects of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *30*(5), 643–653. https://doi.org/10.1177/0146167203262851

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*(4), 337–353. https://doi.org/10.1006/jecp.2000.2594

Geng, L., Zhang, L., & Zhang, D. (2011). Improving spatial abilities through mindfulness: Effects on the mental rotation task. *Consciousness and Cognition*, *20*(3), 801–806. https://doi.org/10.1016/j.concog.2011.02.004

Groves, N. B., Kofler, M. J., Wells, E. L., Day, T. N., & Chan, E. S. (2020). An examination of relations among working memory, ADHD symptoms, and emotion regulation. *Journal of Abnormal Child Psychology*, *48*(4), 525–537. https://doi.org/10.1007/s10802-019-00612-8

Hausmann, M., Schoofs, D., Rosenthal, H. E. S., & Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences-A psychobiosocial approach. *Psychoneuroendocrinology*, *34*(3), 389–401. https://doi.org/10.1016/j.psyneuen.2008.09.019

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527. https://doi.org/10.3758/s13428-015-0588-x

Heil, M., Jansen, P., Quaiser-Pohl, C., & Neuburger, S. (2012). Gender-specific effects of artificially induced gender beliefs in mental rotation. *Learning and Individual Differences*, *22*(3), 350–353. https://doi.org/10.1016/j.lindif.2012.01.004

Heppner, W. L., Kernis, M. H., Lakey, C. E., Campbell, W. K., Goldman, B. M., Davis, P. J., & Cascio, E. V. (2008). Mindfulness as a means of reducing aggressive behaviour: Dispositional and situational evidence. *Aggressive Behavior*, *34*(5), 486–496. https://doi.org/10.1002/ab.20258

Hertzog, C., Vernon, M. C., & Rypma, B. (1993). Age differences in mental rotation task performance: The influence of speed/accuracy tradeoffs. *Journal of Gerontology*, *48*(3), P150–P156. https://doi.org/10.1093/geronj/48.3.P150

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Inzlicht, M., Aronson, J., Good, C., & McKay, L. (2006a). A particular resiliency to threatening environments. *Journal of Experimental Social Psychology*, *42*(3), 323–336. https://doi.org/10.1016/j.jesp.2005.05.005

Inzlicht, M., McKay, L., & Aronson, J. (2006b). Stigma as ego depletion: How being the target of prejudice affects self-control. *Psychological Science*, *17*(3), 262–269. https://doi.org/10.1111/j.1467-9280.2006.01695.x

Jansen-Osmann, P., & Heil, M. (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain and Cognition*, *64*(3), 217–227. https://doi.org/10.1016/j.bandc.2007.03.002

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General*, *137*(4), 691–705. https://doi.org/10.1037/a0013834

Jolicœur, P., Regehr, S., Smith, L. B., & Smith, G. N. (1985). Mental rotation of representations of two-dimensional and three-dimensional objects. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(1), 100–129. https://doi.org/10.1037/h0080118

Jost, L., & Jansen, P. (2020). *A novel approach to analyzing all trials in chronometric mental rotation and description of a flexible extended library of stimuli.* Manuscript submitted for publication.

Jost, L., Weishäupl, A., & Jansen, P. (2019). *Interactions between simultaneous aerobic exercise and mental rotation*. Manuscript submitted for publication.

Kahneman, D. (1973). *Attention and effort (Vol. 1063)*. Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology*, *77*(2), 323–338. https://doi.org/10.1348/000709906X113662

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., & Devins, G. (2006). The Toronto mindfulness scale: Development and validation. *Journal of Clinical Psychology*, *62*(12), 1445–1467. https://doi.org/10.1002/jclp.20326

Linn, M., & Petersen, A. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*(6), 1479–1498. https://doi.org/10.2307/1130467

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences*, *12*(4), 163–169. https://doi.org/10.1016/j.tics.2008.01.005

Mathôt, S. (2013). A simple way to reconstruct pupil size during eye blinks. https://doi.org/10.6084/m9.figshare.688001.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

Mathôt, S., van der Linden, L., Grainger, J., & Vitu, F. (2015). The pupillary light response reflects eye-movement preparation. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(1), 28–35. https://doi.org/10.1037/a0038653

Matlin, M. W. (2011). *The psychology of women*. Cengage Learning.
Moè, A., & Pazzaglia, F. (2006). Following the instructions!: Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, *16*(4), 369–377. https://doi.org/10.1016/j.lindif.2007.01.002
Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for R. *Journal of Statistical Software*, *43*(9), 1–14. https://doi.org/10.18637/jss.v043.i09
Neuburger, S., Ruthsatz, V., Jansen, P., & Quaiser-Pohl, C. (2015). Can girls think spatially? Influence of implicit gender stereotype activation and rotational axis on fourth graders' mental-rotation performance. *Learning and Individual Differences*, *37*, 169–175. https://doi.org/10.1016/j.lindif.2014.09.003
Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702
Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology*, *48*(2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x
Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math=male, me=female, therefore math≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44–59. https://doi.org/10.1037/0022-3514.83.1.44
Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126
Pennington, C. R., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European Journal of Social Psychology*, *49*(4), 717–734. https://doi.org/10.1002/ejsp.2540
Peters, M., & Battista, C. (2008). Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. *Brain and Cognition*, *66*(3), 260–264. https://doi.org/10.1016/j.bandc.2007.09.003
Redick, T. S., Calvo, A., Gay, C. E., & Engle, R. W. (2011). Working memory capacity and go/no-go task performance: Selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 308–324. https://doi.org/10.1037/a0022216
Schofield, T. P., Creswell, J. D., & Denson, T. F. (2015). Brief mindfulness induction reduces inattentional blindness. *Consciousness and Cognition*, *37*, 63–70. https://doi.org/10.1016/j.concog.2015.08.007
Seeley, E. A., & Gardner, W. L. (2003). The "selfless" and self-regulation: The role of chronic other-orientation in averting self-regulatory depletion. *Self and Identity*, *2*(2), 103–117. https://doi.org/10.1080/15298860309034
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701
Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*(1), 415–437. https://doi.org/10.1146/annurev-psych-073115-103235
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373
Suess, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working memory capacity explains reasoning ability – And a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3
Tartre, L. A., & Fennema, E. (1995). Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables [grades 6–12]. *Educational Studies in Mathematics*, *28*(3), 199–217. https://doi.org/10.1007/BF01274173
Tempel, T., & Neumann, R. (2016). Gender role orientation moderates effects of stereotype activation on test performances. *Social Psychology*, *47*(2), 63–73. https://doi.org/10.1027/1864-9335/a000259
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446
Voyer, D., Butler, T., Cordero, J., Brake, B., Silbersweig, D., Stern, E., & Imperato-McGinley, J. (2006). The relation between computerized and paper-and-pencil mental rotation tasks: A validation study. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 928–939. https://doi.org/10.1080/13803390591004310
Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250
Weger, U. W., Hooper, N., Meier, B. P., & Hopthrow, T. (2012). Mindful maths: Reducing the impact of stereotype threat through a mindfulness exercise. *Consciousness and Cognition*, *21*(1), 471–475. https://doi.org/10.1016/j.concog.2011.10.011
Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9
Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.
Xie, F., Zhang, L., Chen, X., & Xin, Z. (2019). Is spatial ability related to mathematical ability: A meta-analysis. *Educational Psychology Review*, *32*, 113–155. https://doi.org/10.1007/s10648-019-09496-y