The title

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Modul 6b: Empirisch-Experimentelles Praktikum

Dr.

07.08.2023

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline. Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines. One sentence clearly stating the **general problem** being addressed by this particular study. One sentence summarizing the main result (with the words "**here we show**" or their equivalent). Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more **general context**. Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

The title

Methods

Preregistration and version control

The hypotheses, the inclusion/exclusion criteria, used databases, search queries and the basic theoretical foundation of this systematic literature review are preregistered and can be found on Moodle or in the GitHub repository.

As suggested by Lakens (2022) (Chapter 14), the present systemic literature used a GitHub repository to store all data and files. The repository is available at: https://github.com/julianrottenberg/Stereotype_Threat_im_akademischen_Kontext

This approach allows for more transparency and reproducibility as well as accountability.

Artificial Intelligence (AI)

It should be acknowledged that artificial intelligence has been used as an aid in this review - namely, Anthropic's Claude AI 3.5 Sonnet (Anthropic, 2024) and GitHub's Copilot (GitHub & OpenAi, 2024), the latter was directly integrated into RStudio Server (Posit team, 2024). The chats that directly influenced this review are all available on the GitHub repository. For Github Copilot the autocomplete-style suggestions were used.

Claude AI 3.5 Sonnet was used to generate descriptions of the papers used in this review - based on a template. The process here was as follows: First the template was manually filled out by a human, after this process was completed for every paper, a second template was created, the contents of which were filled out by AI and then, later, used in conjunction with the manually created templates. When the different templates differed from one another the primary source (i.e. the paper the template was based on) was checked again. Both, the human-generated as well as the AI-generated templates can be found on the GitHub repository - the AI generated summaries have been marked as such, beginning with "Claude_Ai_" in their file name.

To clarify, AI was not used to generate any of the text in this review, it was used as a

tool to gather a better understanding and overview of the papers involved. The process of having a human and AI created summary of each paper was chosen to gather an extra layer of security regarding the contents of each paper as well as to counteract possible oversights.

Databases, search queries and inclusion/exclusion criteria

The databases used were Web of Science, Google Scholar, PSYNDEX, ResearchRabbit and EBSCOhost Within EBSCOhost, the databases APA PsycArticles, APA PsycInfo, Psychology and Behavioral Sciences Collection, PSYNDEX Literature with PSYNDEX Tests, Education Source Ultimate, and Academic Search Ultimate were searched.

Furthermore, the snowball method was utilised to find additional papers - however, this approach did not deliver any additional papers, the same applies to ResearchRabbit.

The permalinks to each search used can also be found within the GitHub repository.

Within Web of Science the included document types were “Article”, “Other”, or “Clinical Trail”; the excluded document types were “Book”, “Meeting”, “Editorial Material”, or “Review Article”. Furthermore, the database “Preprint Citation Index” was excluded.

In EBSCOhost, “Apply equivalent subjects” was applied as an Expander, while “Peer Reviewed”, “Document Type*”, and “Publication Type*” were used as Limiters.

In Google Scholar, the following was added at the end of the search query: “AND”empirical study” AND “peer-reviewed” -books -meta-analysis)“.

These extra filters were applied in accordance with the inclusion and exclusion criteria outlined in the preregistration. No other changes were made to the search queries. An overview of the search queries can be found in Table 1.

The inclusion and exclusion criteria specified in the preregistration were applied to each paper. The criteria “Stereotype Threat”, which required studies to “explicitly examine, manipulate, or measure stereotype threat as a key study variable or factor” was enforced on a lot of papers and resulted in their exclusion - even when they were otherwise relevant (more on this in the discussion section), same applies to the “Outcomes” criteria, which

required studies to report "at least one of the following: 1. Neural activation patterns/brain imaging data; 2. Cognitive processes (e.g., working memory, cognitive control/executive functions)" also resulted in the exclusion of papers which indirectly measured these outcomes but/or did not specifically focus on "working memory" for example - as an example: a paper might have used a test that is known to measure working memory but did not mention "working memory" within its abstract, methods or results section, so it was excluded.

Screening

The screening process was done using the software Rayyan (Ouzzani et al., 2016). All results were imported onto the platform. The total amount of papers found was 599 ($N = 599$, $n_{\text{EBSCOhost}} = 105$, $n_{\text{Google Scholar}} = 48$, $n_{\text{PSYNDEX}} = 5$, $n_{\text{ResearchRabbit}} = 4$, $n_{\text{Web of Science}} = 437$). Out of these 83 were duplicates (88 were automatically detected by Rayyan, however, 5 were false positives), leaving 516 papers to be screened. During the first screening another 440 papers were excluded. Papers which were excluded did not fit the inclusion criteria, most prominently, they either did not focus on stereotype threat, had the wrong population (e.g., older adults), did not fit the publication type requirements, or did not measure the outcomes of interest - this was assessed using the title, keywords, and abstract. If neither the title, nor the keywords or abstract mentioned enough information to make a decision, the paper was marked as 'maybe. An example for hypothesis 3 would be, a paper measured working memory but just referred to "the participants" in the abstract, without clarifying that they fit the definition of the academic context. After this first screening, 76 papers remained for the second screening. This second screening was done by looking into the full-text of each paper, here another 44 papers were excluded for the following reasons: wrong focus ($n = 30$), wrong population ($n = 7$), wrong study design ($n = 5$), wrong publication type ($n = 2$) - an overview of this can be found in the PRISMA flowchart in Figure 1. In the end, 32 papers were included in this review, $n = 8$ for hypothesis 1, $n = 9$ for hypothesis 2, and $n = 19$ for hypothesis 3 - some were used for multiple hypotheses. Out of the 516 papers, 319 were excluded for 'wrong focus', 57 for

'wrong population', 14 for 'wrong study design', 12 for wrong publication type, 2 for 'foreign language', and 1 for 'wrong study duration' (some papers were excluded for multiple reasons). A full list of all papers found and excluded can be found in the GitHub repository.

A template was created to summarise each paper, with two versions completed: one by the author and one by Claude AI. The template was a mixture of the following checklists: CASP systematic review checklist (Critical Appraisal Skills Programme, 2018), Review guidelines for extracting data and quality assessing primary studies in educational research (EPPI-Centre, 2003), Critical appraisal checklist for a systematic review (University of Glasgow, n.d.), and the Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses (Wells et al., 2014), which are used to describe studies and assess their quality. Redundant and irrelevant items were eliminated, and the remaining questions were consolidated into a single template. This approach provided a comprehensive overview of the final papers. Based on these summaries, the papers were analysed and the results are presented in the following sections.

RStudio and R packages

The following R packages were used to create this review: R (Version 4.4.1; R Core Team, 2024) and the R-packages *citr* (Version 0.3.2; Aust, 2019), *kableExtra* (Version 1.4.0; Zhu, 2024), *papaja* (Version 0.1.2.9000; Aust & Barth, 2023), *RefManageR* (Version 1.4.0; McLean, 2017), *rmarkdown* (Version 2.27; Xie et al., 2018, 2020), and *tinylabels* (Version 0.2.4; Barth, 2023).

Results

Results Hypothesis 1: Stereotype threat induces variations in neutral activation across different brain areas and networks, potentially influencing academic performance.

Beilock et al. (2007)

In their paper, Beilock et al. (2007) focussed on stereotypes effects on working memory, specifically, which parts of working memory are affected by stereotype threat and when these effects linger on and influence performance on unrelated tasks. To investigate this, they focussed on maths stereotype threat, their population consisted of female college students in the United States. Their paper describes five experiments, all of which used a cross-sectional design. Experiment 1, 3 (both A and B), and Experiment 5, consisted of two groups each, with 'stereotype threat' vs. 'no stereotype threat', 'horizontal vs. vertical modular arithmetic (MA) conditions', and 'spatial two-back vs. verbal two-back task', respectively, each with random allocation. Experiment 2 and 4 consisted of one group each. Since neither Experiment 2 nor 3B were relevant to the hypotheses for this review, they will not be discussed further. Across all experiments, participants were female undergraduate students.

Modular Arithmetic (MA) task. The MA task was used to measure maths performance. Participants were asked to judge the validity of equations, like $60 = 19 mod(4)$, which would result in $false$. These equations were either displayed vertically or horizontally and consisted of varying difficulty, thus differed in working memory demand. Using this type of task allowed the researchers to measure the effect stereotype threat had on working memory.

Two-back task. Participants were given a stimuli and had to decide whether or not the given stimuli matched the one presented two trials before. The stimuli in use were a cluster of identical letters (e.g., *cs*, *ns*) and one of six different spatial locations in an ellipse. The two-back task was split into a verbal (letters) or spatial (locations) version, participants

were randomly assigned to one of these versions.

Experiment 1. $N = 31$ women, of equal self-reported maths skill participated in this experiment, $n_{\text{stereotype threat}} = 14$, $n_{\text{no stereotype threat}} = 17$. Firstly, participants were introduced to the MA task, and were then asked to solve 12 practice problems, these differed in demand but only consisted of horizontal problems. Afterwards, 24 problems were performed by each participant over two blocks, with the first one serving as a baseline and the second as the posttest. Stereotype threat manipulation was performed in between these blocks via text on a computer screen. An adaptation of the stereotype threat manipulation used by Aronson et al. (1999) was used by displaying the text on a computer screen. Maths accuracy and reaction time were measured as dependent variables, while Group (stereotype threat vs. control), Problem working memory demand (low vs. high), and Block (baseline vs. posttest) functioned as independent variables.

Within the stereotype threat condition Group $\times$ Block $\times$ Problem Demand , $F(1,29) = 11.18$, $p < .010$, $\eta_p^2 = 0.28$, resulted in a significant interaction effect for accuracy. Further, Group $\times$ Block $\times$ Problem Demand , on reaction time, showed main effect of block, and problem demand; $F(1,29) = 8.33$, $p < .010$, $\eta_p^2 = 0.22$, and $F(1,29) = 754.5$, $p < .010$, $\eta_p^2 = 0.96$, respectively. Thus, individuals were able to increase their speed over time and the more demanding a problem was, the longer it took to solve it. A comparison of the accuracy between the baseline and posttest within the stereotype threat condition showed no difference in terms of accuracy for low-demand problems, while, for high-demand problems, a significant decrease in accuracy between the posttest ($M = 79.3\%$, $SE = 4.6\%$) and baseline ($M = 89.1\%$, $SE = 3.8\%$) was found; CI [81.00% - 97.00%]; $d = 0.61$.

Beilock et al. (2007) conclude that only high- but not low-demand problems affect working memory under stereotype threat.

Experiment 3A. Here, a sample of thirty-three ($N = 33$) women performed, both, vertical and horizontal MA tasks. Similar to Experiment 1 they were introduced to the subject with a practice block, followed by a baseline block and a posttest block. This time,

all participants received the stereotype threat manipulation in between the last two blocks but were randomly assigned to either the vertical or horizontal problem condition. Afterwards, they were given questionnaires to assess their thought during the stereotype threat manipulation, their perceived importance of task performance, and their state anxiety following stereotype threat. The independent variables consisted of Block (baseline vs. stereotype threat), Problem working memory demand (low vs. high), and Problem orientation (horiztonal vs. vertical), while the dependent variables were maths problem accuracy, reaction times, and self-reported thoughts/worries. Neither the perceived importance of performing well (vertical: $M = 4.67$, $SE = 0.35$; horizontal: $M = 5.27$, $SE = 0.37$) nor state anxiety differed between the groups (vertical: $M = 33.22$, $SE = 1.6$; horizontal: $M = 37.00$, $SE = 2.7$), $F(1,31) = 1.53$, $p = .220$. Thoughts/worries were split into four categories, most common were thoughts about the performance monitoring (34.9%), followed by thoughts related about the processes involved in solving the problems (32.4%), unrelated thoughts made up 18.3% and, lastly, 14.5% of the thoughts related to the stereotype threat manipulation. Again, no significant difference between the groups was found (Categories 1, 3, and 4: $F < 1$; Category 2: $F(1,31) = 2.17$, $p = .150$.)

For the MA problems, a three-way interaction between the independent variables was found, $F(1,31) = 4.12$, $p = .050$, $\eta_p^2 = 0.12$.

Similar to Experiment 1, a significant Block $\times$ Problem Demand interaction was found but only for horizontal problems, $F(1,14) = 7.70$, $p < .020$, $\eta_p^2 = 0.36$. Accuracy suffered significantly from the baseline ($M = 91.7\%$, $SE = 3.6\%$) to the stereotype threat ($M = 81.2\%$, SE = 4.6%) block; CI [84.00% - 99.30%]; $d = 0.64$. The three-way ANOVA for, RTs revealed that high-demand problems were slower, compared to low-demand problems; vertical: $F(1,17) = 306.32$, $p < .010$, $\eta_p^2 = 0.95$; horizontal: $F(1,14) = 11.04$, $p < .010$, $\eta_p^2 = 0.44$. This effect was not significant for horizontal problems and revealed a main effect for vertical problems.

It is concluded that low-demand problems do not suffer under stereotype threat,

however, the working memory is impaired for high-demand problems resulting in a decrease in accuracy (only for horizontal problems) and an increase in reaction time.

Experiment 4. All thirty ($N = 30$) women were tasked to solve horizontal and vertical MA under stereotype threat. Procedure was similar to Experiment 3A, albeit, with a bigger practice block, and the repetition of some problems. The independent variables consisted of Block (baseline vs. stereotype threat), Problem Repetition (no repeat vs. multiple repeat), and Problem Working Memory demand (low vs. high), while the dependent variables did not differ from Experiment 1. The significant Block × Problem Repetition × Problem Working Memory demand interaction, $F(1, 29) = 6.13$, $p < .020$, $\eta_p^2 = 0.17$, was further analysed by differentiating between multi- and no=repeat problems.

For the multi-repeat problems no significant interaction was to be found ($F < 1$), meanwhile a significant effect was found for the no-repeat problems, $F(1,29) = 11.11$, $p < 0.01$, $\eta_p^2 = 0.28$. While the accuracy, again, significantly decreased from the baseline ($M = 65.00\%$, $SE = 3.9\%$) to the stereotype threat block ($M = 65.00\%$, $SE = 5.9\%$; CI [52.80% - 77.20%]; $d = 0.70$) in high-demand problems, within the no-repeat condition, no significant effect was found for the low-demand problems (baseline: $M = 95.00\%$, $SE = 1.50\%$, stereotype threat: $M = 94.80\%$, $SE = 2.80\%$). For the RTs, problem demand, $F(1,26) = 144.14$, $p < .010$, $\eta_p^2 = 0.85$, influenced the RTs more than problem repetition,$F(1,26) = 139.94$, $p < .010$, $\eta_p^2 = 0.84$, both showing main effects.

Based on these results, the authors conclude that practised, horizontal problems do not rely on working memory heavily, evidenced by them not being affected by stereotype threat. The opposite is true for no-repeat problems, which did suffer under stereotype threat, given that they were of high verbal working memory demand.

Experiment 5. The last experiment was preceded by a pilot test to establish whether the two-back tasks were of equal difficulty. This pilot test was done with $N = 27$ women, without any stereotype threat manipulation, the procedure is similar to the main

experiment, thus will not be discussed further.

The main experiment consisted of thirty-three ($N = 33$) women. Upon arrival participants completed a two-back practise task, after changing computers, they practised the MA task. Afterwards, the stereotype threat manipulation was performed, followed by twenty high-demand horizontal problems. In the next step, participants went back to the first computer to complete 100 trails of the same version of the two-back task that was practised before.

Condition (stereotype threat vs. control; control being the pilot test) and Two-back task type (verbal vs. spatial) functioned as independent variables, while the dependent variables were accuracy and reaction time, each for, both, the maths problem and the two-back task.

Comparing the MA results with the previous experiments results for the same type of task (horizontal, high-demand), showed that the stereotype threat significantly inhibited performance ($M = 82.20\%$, $SE = 2.00\%$), while the same cannot be said for the no-threat conditions ($M = 91.50\%$, $SE = 1.50\%$).

For the two-back task, RTs between spatial ($M = 895$ ms, $SE = 49$ ms) and verbal ($M = 1087$ ms, $SE = 59$ ms) task differed significantly, $F(1,31) = 6.133$, $p < .020$, $\eta_p^2 = 0.17$ while the difference in accuracy did not reach significance (verbal: $M = 87.30\%$, $SE = 1.70\%$; spatial: $M = 89.00\%$, $SE = 1.50\%$, $F < 1$). Comparing the performance of the stereotype threat condition with the control (pilot test), showed an interaction between Task $\times$ Experiment for RT, $F(1,56) = 4.38$, $p < .050$, $\eta_p^2 = 0.07$. Without stereotype threat, no significant differences in performance between thee verbal and spatial two-back tasks were found, however, under stereotype threat the verbal task was significantly slower than the spatial task.

Contrary to the previous Experiments, Experiment 5 additionally aimed to investigate whether stereotype threat has a spill over effect on unrelated tasks. Since these results are not relevant to the hypotheses of this review, they will not be discussed in detail.

Using multiple regression analyses the authors found that the stereotype threat did indeed spill over to the two-back task, however, only for the verbal task.

According to the authors, the results of the different Experiments in this paper, show stereotype threats effect on working memory, they mention that "especially the phonological aspects" are affected. Stereotype threat's effect onto task-related worries as well as thoughts serve as further indications for this conclusion. Further, it is concluded that stereotype threat likely affects multiple aspects of working memory. A combination of phonological loop and central executive functioning is suggested to be affected by stereotype threat.

H1 is partially confirmed by this paper, central executive functioning is assumed to involve the prefrontal cortex, however, this is not the only area affected. The phonological loop is associated with BA4, BA49, and (approximately) BA44 and BA 45.

Dunst et al. (2013)

In a 2 (sex: male vs. female) $\times$
2 (stereotype exposure: stereotype threat vs. no stereotype threat) cross-sectional between-subjects design, a mixed-sex sample of secondary school students in Austria, was used to investigate the effects of stereotype threat on neural efficiency as well as sex differences in visuo-spatial task performance. The dependent variables consisted of task performance (accuracy and reaction time), brain activation (task-related-power changes), and neural efficiency (correlation between figural intelligence and brain activation); sex, stereotype exposure, and figural intelligence functioned as independent variables. Task-related-power (TRP) changes were measured using an EEG, specifically the upper alpha band (10-12 Hz) were examined.

The final sample consisted of 58 participants ($N = 58$; 26 girls, 32 boys). Participants were randomly assigned to either the stereotype threat or control conditions, additionally, they were IQ-matched between experimental groups.

Firstly, participants were set up with the EEG, 33 electrodes were placed, following the international 10-20 system. Afterwards, the stereotype threat manipulation was

performed using a message claiming boys to be better on the subsequent task, in the no-threat condition the message was neutral, stating that sex differences did not exist. The experimental task followed and consisted of Shepard-Metzler figures, here, figures were presented in a 3D presentation mode, participants had to decide whether the figures were identical or mirroed, to do so, the figures had to be rotated mentally. Previous studies successfully used a similar manipulation in the past.

None of the behavioural analyses were significant, since they do not relate to this reviews hypotheses, they will not be discussed further. The TRP changes were analysed, a main effect for Stereotype Exposure ($F(1,54) = 3.93$, $p = .050$, partial $\eta^2 = 0.07$) was found using a four-way ANOVA, with the between-subjects factors of Stereotype Exposure, and Sex and the within-subjects factors of Hemisphere and Area. A higher cortical activation ($M = 0.07$, $SD = 0.03$) was found in the stereotype threat condition compared to the control condition ($M = -0.03$, $SD = 0.03$). An inverse indication for neural efficiency was found in the correlation of figure intelligence and TRP. While the researchers were able to find a negative IQ-brain activation relationship in both girls and boys under no-threat, the same cannot be said for the threat condition, where no significant correlations were found for either sex. Thus, neural efficiency was only found in the no-threat condition for boys.

H1 is not confirmed by this paper, as the only significant effect under stereotype threat was an increase in cortical activation, which are regions of the cerebral cortex or cerebellar cortex (American Psychological Association, 2018).

Forbes et al. (2015)

In their paper, Forbes et al. (2015) looked into negative subject appraisals under stereotype threat and effect on the default mode network (DMN), specifically individual differences in neural networks that moderate the effect perceived performance of stereotype threat. The researchers hypothesised that for minorities, the greater DMN phase-locking is at rest, the less the stereotype threat will affect their performance perceptions, compared to Whites.

The final sample consisted of 58 ($N = 58$) participants, 25 (11 female) of which were White, the other 33 (22 female) were minorities. The experiment began with preparations for the EEG recording, followed by a resting state EEG, and a stereotype threat manipulation - all participants received the same manipulation. Afterwards, participants tried to solve a probabilistic learning task which was manipulated to evoke similar amounts of correct or wrong feedback. For the stereotype manipulation, participants were told, more intelligent individuals were able to learn the relations in a shorter time frame, in the probabilistic learning task, thus the task was able to predict their intelligence. In between the stereotype threat manipulation and the probabilistic learning task, participants filled out a demographic questionnaire which included a question about their race, to further manipulate stereotype threat. After finishing the task, participants completed a error estimates, as well as self-doubt questionnaires, and a manipulation check.

For the EEG, 32 tin electrodes were placed on the scalp using a stretch-lycra cap. Besides ethnicity (Minority vs. White), the independent variables consisted of the phase-locking between the left lateral parietal cortex (LLPC) and precuneus/posterior cingulate cortex (P/PCC), and the phase-locking between LLPC and the medial prefrontal cortex (MPFC), each at the frequency bands alpha (8-12 Hz) and theta (4-8 Hz), these will also be referred to as DMN phase-locking, if the need to differentiate between them is not given. In short, ethnicity, LLPC-P/PCC phase-locking (alpha and theta), and LLPC-MPFC phase-locking (alpha and theta) formed the independent variables. Error estimates and self-doubt were used as dependent variables.

Minorities and Whites performance on learning rates, error overestimation, self-doubt were similar, this was analysed using a independent samples t-test on learning performance, error estimates, doubt, and stereotype threat manipulation check (all ps $> .050$). However, the stereotype threat manipulation was successful, as indicated by the heightened assumption minorities ($M = 3.08$, $SD = 1.11$) elicited about the researchers expectations about their performance, $t(86) = -2.48$, $p < .020$, compared to Whites ($M = 3.60$, $SD =$

0.77). Despite not being significant (p = .200), minorities overestimated their errors, showing greater self-doubt, compared to Whites. Using regression models, DMN phase-locking during the learning phase was inspected, resulting in no significant effects (ps > .500). The relationship between LLPC-P/PCC phase-locking in the theta band and error estimations showed a tendency to overestimate errors was not related to ethnicity (p = .957), a main effect was found for LLPC-P/PCC theta phase locking (b=-195.29, β=-0.37, SE=81.13, p = .021), which was then moderated by a significant interaction (b=350.13, β=0.37, SE=147.26, p = .021). No significant relationships were found between error estimation and LLPC-P/PCC phase locking, in either alpha or theta bands, for neither ethnic group (ps > .300). For self-doubt, the phase-locking between LLPC-P/PCC in the alpha and theta band, did not result in a significant relationship (ps > .400). For LLPC-MPFC phase-locking and doubt, the researchers were able to effect in the alpha (b=-3.79, β=-0.12, SE=1.28, p = .005 and theta bands (b=-4.41, β=-0.09, SE=1.95, p = .028). LLPC-MPFC phase locking did not interact significantly with ethnicity (p > .200). Among minorities, a correlation between LLPC-MPFC theta phase-locking and self-doubt was found to be significant (r=-0.54, p < .010), while the same cannot be said for Whites (r=-0.04). Moreover, the authors were able to find a significantly greater relationship between these variables for minorities compared to Whites (z=-2.00, p<.050, two-tailed).

Forbes et al. (2015) conclude that phase-locking between DMN regions might help individuals under stereotype threat to mitigate the negative effects of the threat, perhaps by reducing the amount of self-doubt they experience. H1 is supported by this paper.

Forbes et al. (2008)

Forbes et al. (2008) investigated psychological disengagement among minority students under stereotype threat, using cognitive neuroscience methodology. It is hypothesized that error-related negativity (ERN) displays a greater amplitude under stereotype threat and, that greater Error Positivity (Pe) amplitudes to errors would be predicted under stereotype threat.

The study design was cross-sectional with two groups, diagnostic of intelligence (DIQ; stereotype threat) and control (no stereotype threat). These also made up the independent variables, alongside psychological disengagement (devaluing academics/discounting intelligence tests). ERN and Pe, as well as task performance measurements (number of errors, post-error slowing reaction times), and self-reported measurements (perceived task difficulty, self-doubt) functioned as dependent variables.

The sample consisted of 57 ($N = 57$) minority undergraduates, who were randomly allocated to either the DIQ or control group. Beginning with the EEG setup, participants completed a baseline version of the Eriksen-Flankers task, followed by the stereotype threat manipulation, and a second round of the flankers task. Once finished with the second task, participants filled out a final questionnaire. Stereotype threat manipulation was done by describing the flankers task as a predictive measure of intelligence (DIQ), and the goal of the study, as an investigation into the differences of intelligence between different groups. The task was described as a measure of pattern recognition for the control group. Participants in the DIQ group also completed a demographics questionnaire including their race/ethnicity.

In the Eriksen-Flankers task, participants must quickly identify a target stimulus while ignoring distractors, which are either congruent or incongruent with the target. It is a measure of attention and inhibitory control. For the EEG measure, 32 tin electrodes were placed on the scalp using a stretch-lycra cap. Error-specific activity was determined by subtracting the average waveforms of correct responses from error responses. The ERN was measured as the peak negative defelction at Fz (frontal midline electrode) between 50 and 130 ms after the response, while the Pe was measured as the peak positive deflection at site Pz (midline parietal electrode) between 200 and 500 ms after the error, based on these difference waveforms. The final questionnaire asked the participant to assess how doubtful, foolish, inferior, insecure and unsure they felt while completing the task (on a 7-point scale).

Through repeated measures analysis on premanipulation early stage amplitudes, a general ERN pattern was identified, considering site (Fz, Cz [central midline electrode], Pz)

and accuracy (correct, error), main effects for site and accuracy were found, $F_{site}(1,40)=42.34$, $p < .001$; $F_{accuracy}(1,40)=71.43$, $p < .001$. At Fz ($\eta^2=0.53$) and Cz ($\eta^2=0.66$), ERN differences between correct and error trials were most prominent, compared to Pz ($\eta^2=0.47$, $F(1,40)=3.00$, $p = .090$). Analysing premanipulation later stage amplitudes, using repeated measures analysis, Pe was established, again, main effects were found for site and accuracy, $F_{site}(1,40)=55.08$, $p < .001$; $F_{accuracy}(1,40)=77.68$, $p < .001$, with a notable interaction, $F(1,40)=13.29$, $p < .001$. Pe differences between error and correct trails were suggested to be larger at Pz ($\eta^2=0.71$) and Cz ($\eta^2=0.57$), compared to Fz ($\eta^2=0.48$).

Using simple slope analysis, within the DIQ condition ($\beta=0.46$, $p < .010$), smaller ERN amplitudes were found, compared to the control condition ($\beta=-0.21$, $p = .370$), if devaluing was used as a predictor. A interaction at Fz ($\beta=0.33$, $p < .020$, $R^2=0.40$) was observed in the analyses, examining devaluing as a moderator of diagnosticity on ERN amplitudes. No significant effects were found using discounting as a moderator (ps $> .100$). However, on Pe amplitudes a significant moderation effect of discounting on diagnosticity was observed at Pz, $\beta=0.29$, $p < .030$, $R^2=0.52$). Further, in the pre-threat task, discounting was able to predict lower Pe amplitudes, $\beta=-0.41$, $p < .050$, this effect was not found when the stereotype threat was present ($\beta=0.19$, $p= .200$). If participants were low in discounting ($\beta_{Low}=-0.39$, $p < 0.04$), smaller Pe amplitudes were found, compared to control participants, while linking the task to intelligence. In the opposite case, i.e. high discounting ($\beta_{High}=0.20$, $p = .230$), participants showed larger Pe amplitudes. Testing devaluing as a moderator of diagnosticity on Pe amplitudes, only a devaluing main effect, $\beta=-0.27$, $p < .030$, was found, while other effects were not significant (ps $> .100$).

Since the results on error analyses, posterror slowing, and self-reported difficulty as well as self-doubt are not relevant to the hypotheses of this review, they will not be discussed further. One exception is part of the posterror slowing analyses, which indicated that, when paired with effects on ERN and errors, minorities valuing academics, tended to make fewer errors and showed less posterror slowing. H1 is partially being confirmed by this paper, as

neural activation was found due to stereotype threat, however, the results for the affected areas are more vague, being linked to the anterior cingulate of the prefrontal cortex.

Jończyk et al. (2022)

In their study Jończyk et al. (2022) used EEG measurements in combination with behavioural performance tasks to investigate stereotype threats effects on creative thinking. The researchers hypothesise, that a measurable decrease in alpha can be expected under stereotype threat, if it affects creative in a negative way. On the other hand, an increase in alpha power in combination with increased creative thinking, can be expected if stereotype threat does not discourage but rather motivates individuals. Additionally, a positive correlation is expected between elevated creative thinking and heightened alpha power.

The study design was cross-sectional with one group, thus every participant received the stereotype threat manipulation. Measurements were taken before and after threat manipulation, forming the independent variables, while creative thinking and alpha power formed the dependent variables. Alpha power was measured using an elastic cap with 31 active Ag/AgCl (silver/silver chloride) electrodes. Task related power (TRP) was calculated in the lower (8-10 Hz) and upper (10-12 Hz) alpha bands before and after the stereotype threat manipulation.

The final sample consisted of twenty-three ($N = 23$) female undergraduates from an American university. Beginning with a demographics questionnaire, participants were then prepared for the EEG during which they completed further questionnaires. A resting-state EEG was recorded followed by practice trails of the Alternative Uses task (AUT) and Utopian Situations task (UST). While in the AUT participants have to come up with new or unorthodox uses for everyday objects, in the UST were given scenarios and had to come up with creative solutions. To measure originality, participants answers were evaluated, on a 5-point scale, by five independent and trained judges; creative fluency was measured by the number of answers given. Following the practice, the first block of experimental tasks for the AUT and UST were completed. Stereotype threat was manipulated after one block of AUT

and UST, using a text modelled after previous studies. Here, the participants were told that women usually perform worse on the tasks, and thus, the participants should try their best on the following block. After the manipulation, the second block of AUT and UST were completed, followed by another resting-state EEG recording. Finally, participants completed the Stereotype Vulnerability Scale (SVS) as well as the self-efficacy scale and the Big Five Inventory.

Since the results of originality and fluency are not the primary focus of this reviews hypothesis, they will not be discussed in detail, same goes for the self-efficacy, SVS and Big Five Inventory results. Neither idea fluency nor idea originality did differ significantly between pre- and post-threat measures, also, no significant correlations were found between fluency/originality and self-efficacy, SVS, or Big Five Inventory. EEG results were calculated using a 2 (pre- vs. post-threat; i.e., no stereotype threat vs. stereotype threat) × 2 (hemisphere: left vs. right) × 6 (area: anteriofrontal, fronto-central, centrotemporal, centro-parietal, parietal, parieto-occipital) × 2 (block half: first half vs. second half) within-subject repeated measures ANOVA. A main effect of threat was found in the lower alpha range (8-10 Hz), $F(1,21)=19.41$, $p<.001$, $\hat{\eta}^2_G=0.05$, 90% CI [0.00, 0.26], with greater alpha Event-Related Synchronisation (ERS) after the administration of stereotype threat ($M_{\text{post-threat}}=10.00$, 95% CI [-4.38, 24.39]). For hemisphere a main effect was found, with greater ERS in the right hemisphere, $F(1,21)=9.20$, $p<.006$, $\hat{\eta}^2_G=0.02$, 90% CI [0.00, 0.20]. Higher ERS in the right hemisphere was found for frontocentral ($M_{\text{right}}=7.88$, 95% CI [-7.15, 22.92]; $M_{\text{left}}=-4.51$, 95% CI [-19.55, 10.52]), centrotemporal ($M_{\text{right}}=6.76$, 95% CI [-8.28, 21.79]; $M_{\text{left}}=-12.07$, 95% CI [-27.11, 2.96]), centroparietal ($M_{\text{right}}=9.26$, 95% CI [-5.78, 24.30]; $M_{\text{left}}=-5.12$, 95% CI [-20.16, 9.92]), and parietal regions ($M_{\text{right}}=8.08$, 95% CI [-6.96, 23.11]; $M_{\text{left}}=-6.08$, 95% CI [-21.12, 8.95]), in an interaction between area and hemisphere, $F(2.74,57.61)=3.15$, $p=.036$, $\hat{\eta}^2_G=0.00$, 90% CI [0.00, 0.00]. Similarly, a main effect was observed in the upper alpha band, $F(1,21)=15.42$, $p<.001$, $\hat{\eta}^2_G=0.05$, 90% CI [0.00, 0.26], along with a hemispheric difference, $F(1,21)=11.43$, $p<.003$, $\hat{\eta}^2_G=0.02$, 90% CI [0.00, 0.20].

The area-by-hemisphere interaction also indicated greater ERS in the right hemisphere across various regions, $F(2.66,55.86)=4.06$, $p=.014$, $\hat{\eta}^2_G=0.00$, 90% CI [0.00, 0.00], namely centrotemporal (M_{right}=0.25, 95% CI [-14.51, 15.00]; M_{left}=-19.35, 95% CI [-34.10, -4.60]), centroparietal (M_{right}=6.20, 95% CI [-8.56, 20.95]; M_{left}=-9.21, 95% CI [-23.97, 5.54]), and parietal regions (M_{right}=9.05, 95% CI [-5.71, 23.80]; M_{left}=-8.90, 95% CI [-23.65, 5.85]).

Discussion

References

American Psychological Association. (2018). Cortical activation [Dictionary]. In *APA dictionary of psychology*. https://dictionary.apa.org/.

Anthropic. (2024). *Claude Ai 3.5 Sonnet*. https://claude.ai/.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*(1), 29–46. https://doi.org/10.1006/jesp.1998.1371

Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. https://github.com/crsh/citr

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels*. https://cran.r-project.org/package=tinylabels

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

Critical Appraisal Skills Programme. (2018). CASP Systematic Review Checklist [Organization]. In *CASP - Critical Appraisal Skills Programme*. https://casp-uk.net/casp-tools-checklists/.

Dunst, B., Benedek, M., Bergner, S., Athenstaedt, U., & Neubauer, A. C. (2013). Sex differences in neural efficiency: Are they due to the stereotype threat effect? *Personality and Individual Differences*, *55*(7), 744–749. https://doi.org/10.1016/j.paid.2013.06.007

EPPI-Centre. (2003). *Review guidelines for extracting data and quality assessing primary studies in educational research* (Guidelines Version 0.9.7). Social Science Research Unit.

Forbes, C. E., Leitner, J. B., Duran-Jordan, K., Magerman, A. B., Schmader, T., & Allen, J. J. B. (2015). Spontaneous default mode network phase-locking moderates performance perceptions under stereotype threat. *Social Cognitive and Affective Neuroscience*, *10*(7), 994–1002. https://doi.org/10.1093/scan/nsu145

Forbes, C. E., Schmader, T., & Allen, J. J. B. (2008). The role of devaluing and discounting in performance monitoring: A neurophysiological study of minorities under threat. *Social Cognitive and Affective Neuroscience*, *3*(3), 253–261. https://doi.org/10.1093/scan/nsn012

GitHub, & OpenAi. (2024). *GitHub Copilot*. copilot.github.com.

Jończyk, R., Dickson, D. S., Bel-Bahar, T. S., Kremer, G. E., Siddique, Z., & Van Hell, J. G. (2022). How stereotype threat affects the brain dynamics of creative thinking in female students. *Neuropsychologia*, *173*, 108306. https://doi.org/10.1016/j.neuropsychologia.2022.108306

Lakens, D. (2022). *Improving Your Statistical Inferences*. Zenodo. https://doi.org/10.5281/ZENODO.6409077

McLean, M. W. (2017). RefManageR: Import and manage BibTeX and BibLaTeX references in r. *The Journal of Open Source Software*. https://doi.org/10.21105/joss.00338

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Posit team. (2024). *RStudio: Integrated development environment for R* [Manual]. Posit Software, PBC.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

University of Glasgow. (n.d.). *Critical appraisal checklist for a systematic review* [Checklist]. Department of General Practice, University of Glasgow.

Wells, G., Shea, B., O'Connell, D., Robertson, J., Welch, V., Losos, M., & Tugwell, P.

(2014). The newcastle-ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. *Ottawa Health Research Institute Web Site*, *7*.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook

Zhu, H. (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax*. https://CRAN.R-project.org/package=kableExtra

Table 1

Search queries used for the systematic literature review.

Hypothesis	Search Query
H1	("stereotype threat") AND (neural OR neuroimaging OR "functional magnetic resonance imaging" OR fMRI OR electroencephalo* OR EEG OR ERP OR "brain activation" OR amygdala OR "prefrontal cortex" OR "default mode network" OR "salience network") AND (academ* OR education* OR stud* OR learn* OR perform* OR school OR university OR college)
H2	("stereotype threat") AND ("cognitive control" OR "executive function" OR "executive function network" OR "cognitive control network" OR "brain activation" OR "brain activation patterns" OR "cognitive tasks" OR "executive tasks" OR "cognitive assessment" OR "executive assessment") AND (academ* OR education* OR stud* OR learn* OR perform* OR school OR university OR college)
H3	("stereotype threat") AND ("working memory*" OR "processing speed" OR accuracy) AND (academ* OR education* OR stud* OR learn* OR perform* OR school OR university OR college)

Note. The search queries were used in the databases Web of Science, Google Scholar, PSYNDEX, ResearchRabbit, and EBSCOhost. The permalinks to each search used can be found within the GitHub repository.

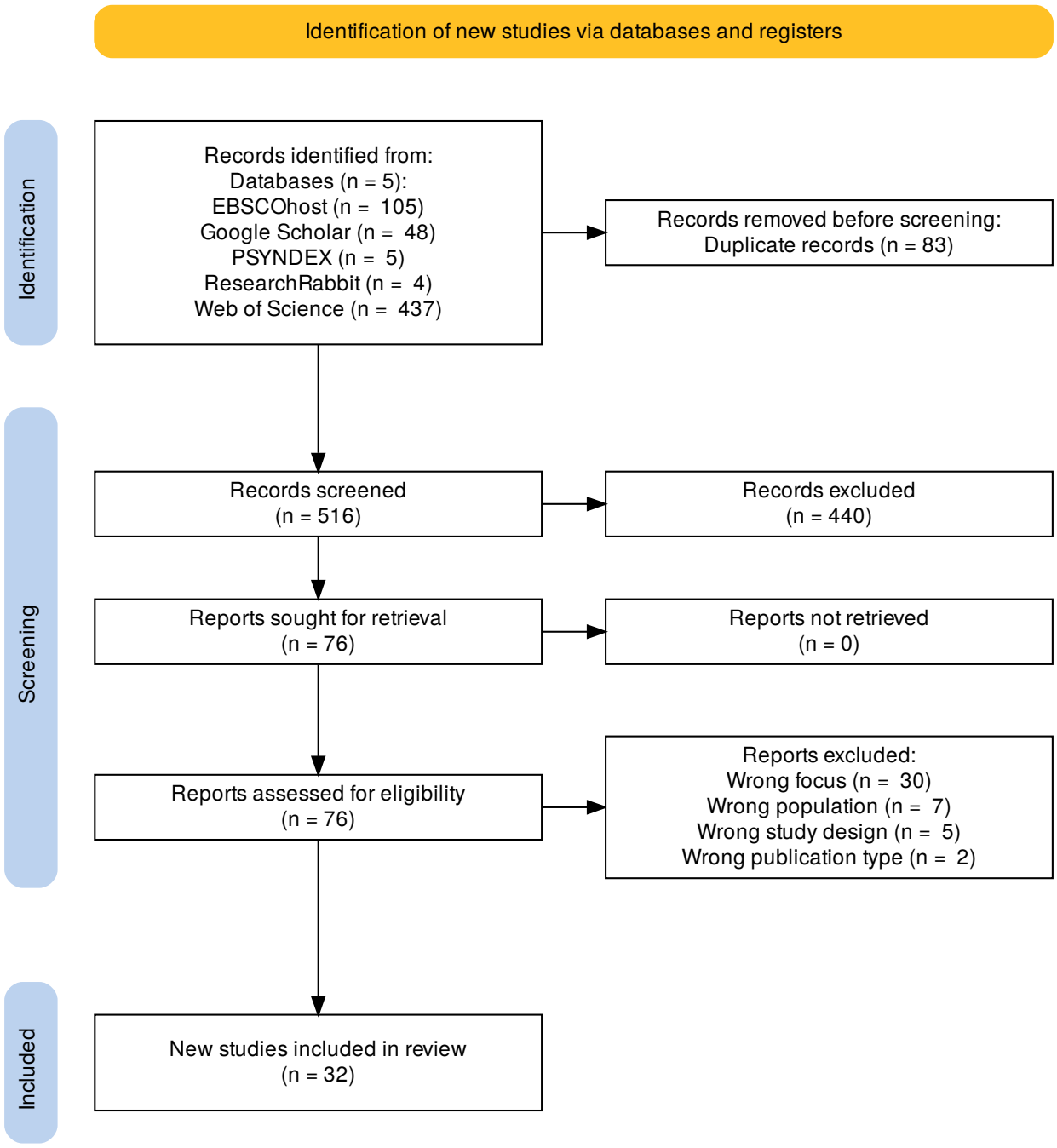

Figure 1

PRISMA flowchart of the screening process.