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.

< !– https://tinyurl.com/ybremelq – >

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, further, follow-up questions were asked to clear up uncertainties. 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 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 and 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 create a 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 utilized 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

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.

enforced on plenty 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 and paper details

The screening process was done using the software Rayyan (Ouzzani et al., 2016). All results were imported onto the platform. The total number of papers found was 600 ($N = 600$, $n_{\text{EBSCOhost}} = 105$, $n_{\text{Google Scholar}} = 48$, $n_{\text{PSYNDEX}} = 5$, $n_{\text{ResearchRabbit}} = 5$, $n_{\text{Web of Science}} = 437$). Out of these, 83 were duplicates (88 were automatically detected by Rayyan; however, 5 were false positives), leaving 517 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, 77 papers remained for the second screening. This second screening was done by looking into the full-text of each paper, here another 49 papers were excluded for the following reasons: wrong focus ($n = 30$), wrong study design ($n = 10$), wrong population ($n = 7$), wrong publication type ($n = 2$) — an overview of this can be found in the PRISMA flowchart (Haddaway et al., 2022) in Figure 1. In the end, 28 papers were included in this review, $n = 8$ for Hypothesis 1, $n = 9$ for

Hypothesis 2, and $n = 14$ for Hypothesis 3 — some were used for multiple hypotheses. Out of the 517 papers, 382 were excluded for 'wrong focus', 73 for 'wrong population', 43 for 'wrong study design', 13 for wrong publication type, 4 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 summarize 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.

Reporting of *p*-values

It should be noted that some *p*-values are reported for example, $p < .010$, this is not in accordance with APA guidelines ("6.36 Decimal Fractions," 2020); however, this format was chosen, since the papers it was taken from used this format, and it was deemed important to keep the original format, especially since it is unknown what the actual *p*-value was.

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).

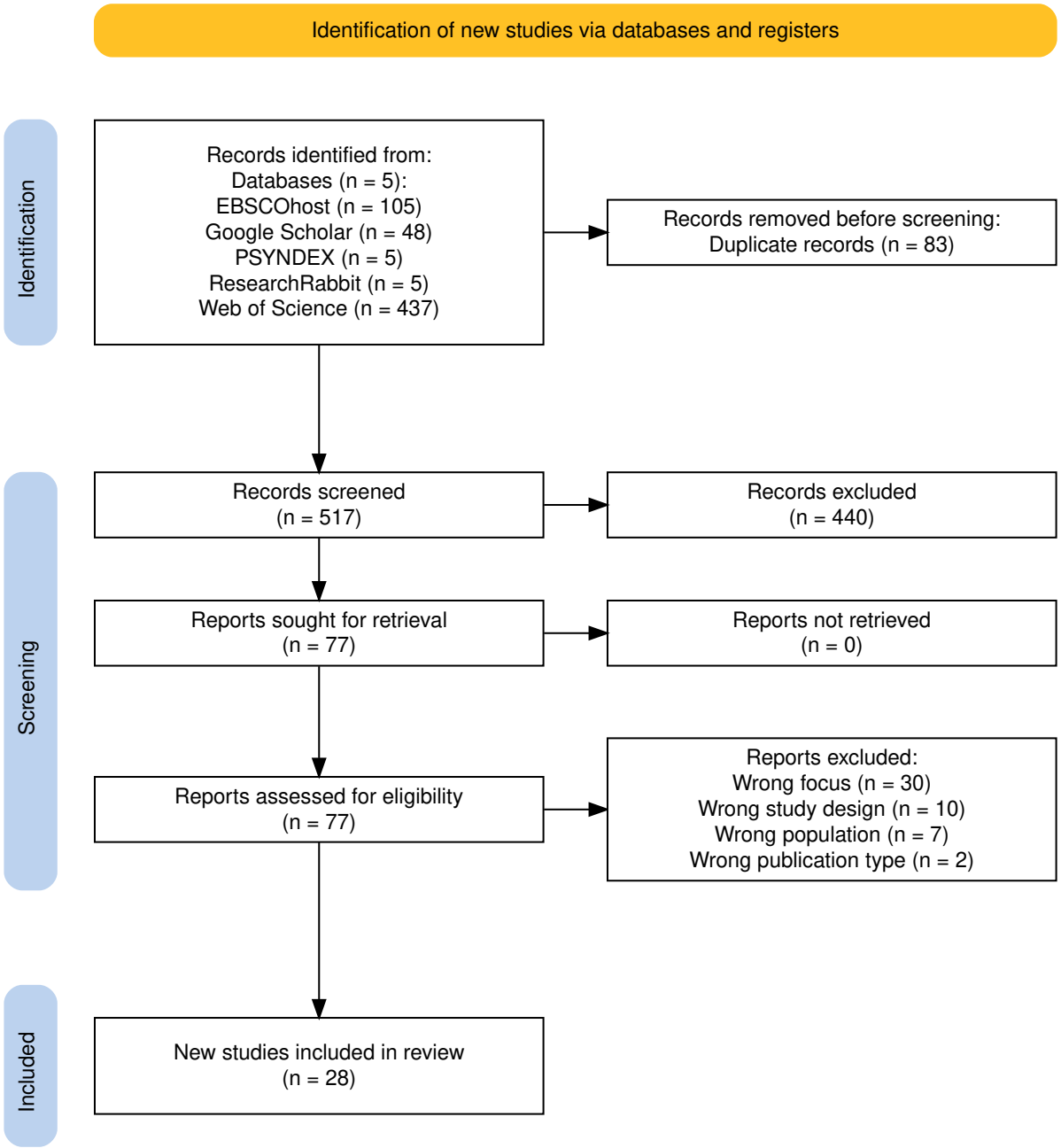

Figure 1

PRISMA flowchart of the screening process.

Results

Hypothesis 1: Stereotype threat, brain areas/networks, performance.

Beilock et al. (2007) hypothesized that high-pressure situations would impair maths performance, especially under high working memory load. The dependent variables were maths accuracy, self-reported thoughts, and reaction time (RT), while the independent variables were group (stereotype threat vs. control), problem working demand (low vs. high), block (baseline vs. post-test), and *2*-back task (verbal, spatial). 31 women participated in Experiment 1, 33 in Experiment 3A, and 30 in Experiment 4. Experiment 1: Under threat, the Group × Block × Problem Demand interaction was found to be significant for accuracy, $F(1,29) = 11.18$, $p < .010$, $\eta_p^2 = .280$. High-demand problems, showed a significant decrease in accuracy at the post-test, CI [81.00% - 97.00%]; $d = 0.61$.

Experiment 3A: For the MA problems, a three-way interaction between the independent variables was found, $F(1,31) = 4.12$, $p = .050$, $\eta_p^2 = .120$. Accuracy suffered significantly under stereotype threat; CI [84.00% - 99.30%]; $d = 0.64$.

Experiment 4: A significant Block × Problem Repetition × Problem Working Memory demand interaction was found, $F(1, 29) = 6.13$, $p < .020$. Accuracy significantly decreased under stereotype threat; CI [52.80% - 77.20%]; $d = 0.70$) in high-demand problems.

Experiment 5: Under stereotype threat, an interaction between Task × Experiment for RT, $F(1,56) = 4.38$, $p < .050$, $\eta_p^2 = .070$ was found, with verbal tasks being significantly slower than spatial ones. 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 BA45.

Dunst et al. (2013) investigated the effects of stereotype threat on neural efficiency and sex differences in visuospatial task performance, using task performance, brain activation, and neural efficiency as dependent variables, and sex, stereotype

exposure, and figural intelligence as independent variables The final sample consisted of 58 participants ($N = 58$; 26 girls). The TRP changes revealed a main effect for Stereotype Exposure ($F(1, 54) = 3.93$, $p = .050$, partial $\eta^2 = 0.07$), with heightened cortical activation ($M = 0.07$, $SD = 0.03$) in the stereotype threat condition. 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).

In a sample of 58 participants (33 minorities), Forbes et al. (2015) investigated how DMN phase-locking at rest affects stereotype threat's impact on performance perceptions in minorities versus Whites. 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. Error estimates and self-doubt were used as dependent variables.

The relationship between LLPC-P/PCC theta phase-locking showed a main effect on error estimation (b = -195.29, β = -0.37, $SE = 81.13$, $p = .021$), which was then moderated by a significant interaction ($b = 350.13$, $\beta = 0.37$, $SE = 147.26$, $p = .021$). Among minorities, a correlation between LLPC-MPFC theta phase-locking and self-doubt was found to be significant (r = -0.54, $p < .010$), differing significantly from Whites (z = -2.00, $p < .050$. H1 is supported by this paper.

Forbes et al. (2008) 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). Diagnostic of intelligence (DIQ; stereotype threat) and control, alongside psychological disengagement (devaluing

academics/discounting intelligence tests) were the independent variables and ERN, Pe, task performance, and self-reported measurements were the dependent variables. The sample consisted of 57 ($N = 57$) minority undergraduates. The ERN was measured as the peak negative deflection 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. Under stereotype threat ($\beta = 0.46$, $p < .010$), smaller ERN amplitudes were found, with devaluing was a predictor. On Pe amplitudes a significant moderation effect of discounting on diagnosticity was observed at Pz, $\beta = 0.29$, $p < .030$, $R^2 = 0.52$). If participants were low in discounting (β_{Low} = -.390, $p < 0.04$), smaller Pe amplitudes were found under stereotype threat. In the opposite case, i.e. high discounting ($\beta_{High} = 0.20$, $p = .230$), participants showed larger Pe amplitudes. 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) examined the impact of stereotype threat on alpha power and creative thinking in 23 female undergraduates. Measurements were taken before and after threat manipulation, forming the independent variables, while creative thinking (measured using Alternative Uses task; AUT, and Utopian Situations task; UST) and alpha power formed the dependent variables. 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.

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 Synchronization (ERS) after the administration of stereotype threat ($M_{\text{post-threat}} = 10.00$, 95% CI [-4.38, 24.39]). A main effect of threat was also found in the upper alpha range, $F(1, 21) = 15.42$, $p < .001$, $\hat{\eta}^2_G = 0.05$, 90% CI [0.00, 0.26], with greater upper alpha ERS after the administration of stereotype threat ($M_{\text{post-threat}} = 3.75$, 95% CI [-10.09, 17.59]).

Comparing the upper alpha power directly before and after stereotype threat with one another, reached significance, $F(1, 21) = 15.28$, $p < .001$, $\hat{\eta}^2_G = 0.08$, 90% CI [0.00, 0.31], with an increase after the manipulation. For the lower alpha power, the same comparison was significant, $F(1, 21) = 4.46$, $p = .047$, $\hat{\eta}^2_G = 0.03$, 90% CI [0.00, 0.23], with an increase after the manipulation. H1 is partially supported by this paper, while stereotype threat did result in increased neural activity, the paper did not explicitly investigate stereotype effects on any of the mentioned areas. However, parts that are associated with the DMN were affected. Furthermore, performance was not found to be inhibited to a significant degree under stereotype threat.

Krendl et al. (2008) used functional magnetic resonance imaging (fMRI) to investigate underlying neural processes of stereotype threat, specifically women under maths stereotype threat. Twenty-eight ($N = 28$) female undergraduates were randomly assigned to either a stereotype threat or a control condition. Neural activation patterns and maths performance (measured by accuracy, i.e., number of correct maths items; reaction time on maths problems) were used as dependent variables, stereotype threat condition (threat vs. control) and time of measurements (Time 1: pre-manipulation vs. Time 2: post-manipulation) functioned as independent variables.

For performance, a significant condition × time interaction was found, $F(1,26) = 11.41$, $p < .005$, $\eta^2_p = .310$. Individuals under stereotype threat performed slightly worse over time, $t(13) = 1.98$, $p = .070$). No main effect of condition was found on performance. The threatened group showed heightened activity in the ventral anterior cingulate cortex (vACC; BA 32/10) during the second test. Significant interactions for BA47, $F(1,26) = 7.35$, $p < .020$, and a trend for BA40, $F(1,26) = 2.93$, $p < .100$, were found. Threatened participants did show increased vACC activation over time, $t(13) = 5.64$, $p < .001$, compared to controls, resulting in a significant interaction, $F(1,26) = 5.97$, $p = .020$. A significant three-way interaction was found for BA47, $F(1,26) = 13.94$, $p < .005$, left BA 40, $F(1,26) = 10.99$, $p < .005$, left BA 39 $F(1,26) = 11.31$, $p < .005$, and right BA39, $F(1,26) = 13.39$, $p < .005$. Heightened activation, for threatened individuals, was found in the affective region (vACC). Regarding H1, neural activation across different brain

areas and networks, was found in this study, furthermore, heightened activation of vACC, which is part of the DMN, further support it. However, BA47 is part of the prefrontal cortex and BA40, as well as BA39 are part of the DMN, all three areas only showed increased activation in the control group, not the threatened group; thus more evidence against H1 is found in this paper. These findings do point out a significant flaw in H1, more on that in the discussion section.

Mangels et al. (2012) investigated the effects of stereotype threat on maths performance and neutral responses to feedback using three event-related potentials (ERPs): anterior P3 (P3a), medial frontal feedback-related negativity (FRN) and posteriorly-maximal late positive potential (LPP). The study design was prospective, with sixty-eight participants ($N = 68$) in total. Stereotype threat vs. no stereotype threat formed the independent variable, maths performance (first-test vs. retest accuracy), ERP responses to feedback, use of the tutor (number of uses and clicks when in use), and learning success (error correction on retest) were used as dependent variables. While maths performance was impaired under stereotype threat, $F(1,64) = 4.30$, $p < .050$, it did not affect retest performance or error correction. While LPP and learning success were closely linked, particularly under threat. Participants showing greater initial detection of negative feedback (indicated by FRN differences) were more likely to disengage from tutor exploration. Those struggling to regulate attention and arousal in response to negative feedback (indicated by LPP differences) benefited less from the tutor. Relating these results to H1, only neural activation is supported by this paper, as the results of the affected areas are more vague, partially due to the ERPs not being significantly influenced by stereotype threat, and partially because it is not possible to link LPP, FRN, and P3a to the suggested brain areas/networks, if so, only to a small degree.

Wu and Zhao (2021) examined the effects of maths stereotype threat using resting-state fMRI and degree centrality (DC) analysis on 48 female Chinese undergraduates. Significant main effects were found for the hippocampus, middle cingulate gyrus (MCG), right cerebellum, and left precentral gyrus (PCG), using a 2

(test: pre- vs post-test) × 2 (condition: stereotype threat vs. control) mixed-effect analysis for the binary graph. However, out of these only MCG had increased RSDC z-values under stereotype threat, $F(1,45) = 4.88$, $p = .032$. Further, an interaction was found to be significant between test and condition in the left cerebellum anterior lobe, left hippocampus, left precuneus, and left middle occipital gyrus (MOC). Here, the mean RSDC z-values were only higher in the stereotype threat condition in the right superior parietal gyrus, left precuneus, left MOG, and right angular gyrus. The weighted graph analysis showed similar patterns, with increased RSDC z-values under threat in the middle cingulate gyrus and the aforementioned regions, except for the left cerebellum, which showed lower values. For interactions, lower RSDC z-values under stereotype threat were only found for the left cerebellum ($F(1, 45) = 4.23$, $p = .046$). H1 is partially supported by this paper, increases in the DMN (associated areas) were found as well as neural activation. However, it is unclear whether performance did suffer as a result of stereotype threat and further, while the ACC and prefrontal cortex are mentioned in the researcher's hypothesis, they are not mentioned in the results.

An overview of the included papers for Hypothesis 1 can be found in Table 2.

Table 2

Overview of the Included Papers for Hypothesis 1

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Beilock et al. (2007)	Experimental	Female college students in US	Behavioral tasks	Stereotype threat, working memory efficiency	ANOVA	Reduced performance on high-demand problems under threat	Yes
Dunst et al. (2013)	Experimental	58 secondary school students in Austria	EEG	Stereotype threat, neural efficiency, task performance	ANOVA	Higher cortical activation under threat	Partially
Forbes et al. (2015)	Experimental	58 participants (25 White, 33 minorities)	EEG	DMN phase-locking, error estimates, self-doubt	Regression models	DMN phase-locking may mitigate stereotype threat effects	Yes

Table 2 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Forbes et al. (2008)	Experimental	57 minority undergraduates	EEG	ERN, Pe, task performance	Repeated measures analysis	Smaller ERN amplitudes under threat	Partially
Jończyk et al. (2022)	Experimental	23 female undergraduates in US	EEG	Creativity, alpha power	Repeated measures ANOVA	Increased alpha power after threat	Partially
Krendl et al. (2008)	Experimental	28 female undergraduates	fMRI	Neural activation, maths performance	Mixed-model ANOVA	Increased vACC activation, decreased cognitive region activation under threat	Partially
Mangels et al. (2012)	Prospective	68 participants	EEG	Maths performance, ERP responses, learning success	ANOVA	LPP and learning success link more pronounced under threat	Partially

Table 2 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Wu and Zhao (2021)	Experimental	48 female undergraduates in China	RS-fMRI	RSDC of brain regions	Mixed-effect analysis	Increased RSDC in DMN areas, decreased in cerebellum and hippocampus	Partially

Note. This table summarizes studies investigating neural activation patterns under stereotype threat. The 'Variables' column focuses on brain areas and networks of interest, such as the amygdala, prefrontal cortex, default mode network, and salience network. 'Methods of Data Analysis' includes neuroimaging techniques like fMRI and EEG. 'Results' highlight changes in neural activation patterns related to stereotype threat.

Hypothesis 2: Stereotype threat, cognitive control.

Guardabassi and Tomasetto (2020) hypothesized that BMI and working memory were negatively, whether this effect was increased by stereotype threat, and whether this effect can be moderated by endorsement. Body Mass Index (BMI), stereotype threat condition (threat vs. control), personal endorsement of obesity-related stereotypes (stereotype endorsement), and weight-based teasing, as independent variables, working memory performance (N-back task performance; 0-back up to two-back) was used as the dependent variable. The final sample consisted of 176 primary school children, 106 of which were boys ($M_{\text{age}} = 116.07$ months, $SD = 10.43$). A significant main effect was found for N-back difficulty, F(2, 198.70) = 43.43, $p < .001$, and the condition × zBMI interaction, F(1, 153.07) = 5.07, $p = .026$. Under stereotype threat, zBMI scores negatively correlated with working memory. H2 is weakly supported by this paper, while stereotype threat did decrease N-back task performance, it cannot be divided which parts of working memory were affected, since the N-back task does involve multiple parts of working memory.

Hirnstein et al. (2014) investigated the effects of stereotype threat on sex differences in cognitive performance and its relation to testosterone levels. A final sample of 136 participants ($n_{\text{female}} = 70$) were tested in a 2 (stereotype threat vs. control) × 2 (mixed vs. same sex) × 2 (male vs. female) factorial design, which formed the independent variables. Cognitive performance (tested using mental rotation, verbal fluency, and perceptual speed tests) alongside testosterone levels, were used as dependent variables. The MRT-3D showed a significant main effect of Sex, with men performing better than women, $F(1,128) = 10.97$, $p = .001$, $\eta^2 = 0.08$, $d = 0.57$, while Condition showed a significant main effect in the MP-2D, $F(1,128) = 4.70$, $p = .032$, $\eta^2 = 0.04$. For verbal fluency, interactions were found between Condition and Group Sex Composition, in both, the WF, $F(1,128) = 4.49$, $p = .036$, $\eta^2 = 0.03$,and the 4W, $F(1,128) = 6.30$, $p = .013$, $\eta^2 = 0.05$). Between Sex × Condition, another interaction was found, in the 4W, $F(1,128) = 6.77$, $p = .011$, $\eta^2 = 0.05$. Furthermore, in the 4W, Condition showed a significant main effect, $F(1,128) = 4.67$, $p = .033$, $\eta^2 = 0.04$. PS also showed the

significant Condition and Group Sex Composition interaction, $F(1,128) = 6.89$, $p = .009$, $\eta^2 = 0.05$. Condition, again showed a significant main effect, this time in the PS, $F(1,128) = 12.65$, $p = .001$, $\eta^2 = 0.09$. H2 is weakly but partially being supported, since individuals under stereotype threat performed worse on the 4W (but not WF) and perceptual speed tests, which measure cognitive control among other things.

In this study, the Jordano and Touron (2017) investigated the effects of stereotype threat on both, cognitive performance and task-related mind-wandering. In both experiments sixty female undergraduates were assigned to either a stereotype threat or control condition; mind-wandering — measured using Task-Unrelated Thoughts (TUTs) and Task-Related Inference (TRI) probes —, task performance (on the Operation Span task; OSPAN), and self-reported measures (i.e., emotional states, cognitive load/perceived difficulty, experience of mind-wandering), were used as dependent variables. In the first experiment, it was found that participants under stereotype threat reported significantly more TRI compared to the control group, $F(1,58) = 5.67$, $p = .021$, $d = 0.64$.

Neither maths verification accuracy nor letter recall accuracy were worse under stereotype threat. In the second experiment, the same pattern was found, with participants under stereotype threat reporting significantly more TRI, $F(1,58) = 5.53$, $p = .022$, $d = 0.42$. Maths verification accuracy was worse under stereotype threat, $F(1,58) = 12.11$, $p = .001$, $d = 0.16$. H2 is partially confirmed. While cognitive control did suffer under stereotype threat, as evidenced by increased mind-wandering (TRI), regardless of task difficulty, task performance only worsened in the more difficult maths component of the OSPAN task.

Krendl et al. (2008) has already been reported in the H1 section. H2 is supported, due to decreased accuracy under stereotype threat, as well as neural activation patterns (vACC, DLPFC, IFG, BA47, BA40).

Lin et al. (2023) investigated the mediating role of executive function in stereotype threat effects on spatial perspective-taking tasks. Stereotype threat (threat vs. control) served as the independent variable, while spatial perspective-taking performance and

executive function performance (inhibition, updating, shifting) were used as dependent variables. Seventy-six undergraduates participated A significant decrease in performance was found for the threat group, $F(1,74) = 10.06$, $p = .002$, $\eta^2 = 0.12$).

In Experiment 2, seventy-seven undergraduates participated.

The effect of stereotype threat was significant for inhibition and updating, $F(1,75) = 11.40$, $p = .001$, $\eta^2 = 0.13$), and $F(1,75) = 5.54$, $p = .021$, $\eta^2 = 0.07$), respectively. Females under threat performed worse on the spatial perspective-taking task. Only inhibition showed a significant (indirect) mediating effect. Further, between spatial perspective-taking and stereotype threat, a significant direct effect was found, direct effect = 0.38, $SE = 0.18$, $t = 2.06$ 95% CI [0.01, 0.75], also, significant negative effects of stereotype threat on inhibition (b = -0.73, $SE = 0.22$, t = -3.38 CI [-1.51, -0.30]) and updating (b = -0.52, $SE = 0.22$, t = -2.35, 95% CI [-0.96, -0.08]) were found. H2 is partially supported by this paper, as cognitive control (updating and inhibition) and performance did suffer under stereotype threat; however, shifting did not reach significance.

Rydell et al. (2014) examined the effects of stereotype threat on executive functions and maths performance across three experiments. The independent variable was condition (stereotype threat vs. control), while maths performance, executive function (inhibition, updating, shifting), and risk-taking behaviour functioned as dependent variables. In the first experiment, 168 ($n_{\text{female}} = 75$) undergraduates participated. Stereotype threat significantly impaired women's inhibition, $F(1,164) = 7.95$, $p = .005$, $\eta^2_{\text{p}} = .046$) and updating, $F(1,164) = 20.89$, $p < .001$, $\eta^2_{\text{p}} = .113$. A performance decrease under threat was found, for accuracy, $F(1,164) = 20.22$, $p < .001$, $\eta^2_{\text{p}} = .110$.

Experiment 2 used a sample of ninety female undergraduates. Regarding maths performance, under threat, accuracy, $t(88)$ = -3.15, $p = .002$, d = -0.66 and the correct item count, $t(88)$ = -5.15, $p < .001$, d = -1.09 were impaired, with updating mediating this effect.

In Experiment 3, a sample of eighty-two female undergraduates participated. Inhibition, $t(79)$ = -2.34, $p = .020$, d = -0.50, and updating, $t(79)$ = -2.29, $p = .030$, d =

-0.50, were significantly affected by stereotype threat The results for maths performance, showed accuracy, $t(79)$ = -3.28, p = .010, d = -0.70, and the correct item count, $t(79)$ = -2.14, p = .035, d = -0.48 being impaired. Under threat, women were more likely to take risks; however, no correlation between it and maths performance was found. The paper more so supports H2 than it does not. Shifting repeatedly did not reach significance, while the other executive functions did.

Ståhl et al. (2012) investigated the role of regulatory focus in stereotype threat effects on cognitive control and maths performance across three experiments. The study used varying combinations of independent variables: condition (stereotype threat vs. no threat), regulatory focus (prevention focus vs. promotion focus vs. no focus), and task order (maths task first vs. maths task last). Dependent variables were cognitive control capacity (measured by a Stroop task) and maths performance (accuracy and response time on an MA task). Sixty-three social science students (n_{female} = 50) participated in Experiment 1, one hundred eight social science students (n_{female} = 82, n_{male} = 26) in Experiment 2, and one hundred sixty-four female students in Experiment 3. A significant effect of stereotype threat on Stroop inference was found, $F(1,61)$ = 8.69, p = .004, η^2_{p} = .130. Experiment 2 found a main effect of threat, $F(1,102)$ = 8.91, p = .004, η^2_{p} = .080. Further, Experiment 1's results for the Stroop task, were replicated, $F(1, 101)$ = 2.80, p < 0.10, η^2_{p} = .030, and an interaction on Stroop inferences were found, $F(2,101)$ = 3.07, p = .050, η^2_{p} = .060). The Stroop inference was only significant under threat, in the prevention condition, $F(1,101)$ = 3.60, p = .240, η^2_{p} = .010.

In Experiment 3, an interaction was found to be significant on maths performance (% of correct responses), $F(1,150)$ = 13.30, p < .001, η^2_{p} = .080. Furthermore, only in the prevention focus condition, when the maths task came first, individuals under threat were found to perform better. For maths performance, measured as RT, the same three-way interaction was significant, $F(1,152)$ = 4.69, p = .030, η^2_{p} = .030. For the maths first condition, prevention focus results in better performance under threat. In the maths last condition, the two-way interaction was marginally significant. H2 is mostly supported by this paper, albeit, only under prevention focus - and here also only after the initial

temporary increase in cognitive control (Experiment 3).

Wister et al. (2013) hypothesized that menstruation stereotype threat would impair performance on cognitive tasks, this effect would be amplified if women believed menstruation to be a hindrance to their cognitive abilities. However, if menstruation was primed as a positive influence, cognitive performance would not be impaired. Ninety-two female undergraduates were assigned into four groups, 2 stereotype threat (menstruation threat vs. no threat) × 2 menstruation prime (positive vs. no positive prime), forming the independent variables, while cognitive performance (measured by a Stroop test and SAT-like maths test) and menstrual attitudes (measured by Menstrual Attitude Questionnaire; MAQ) were used as dependent variables. The experiment ended for everyone with the completion of the Menstrual Attitudes Questionnaire.

While a main effect of threat on Stroop performance was found, Lambda = 0.87, $F(1,68) = 4.91$, $p < .010$. Participants under threat were able to complete less items correctly, $F(1,69) = 9.48$, $p < 0.01$, and a correlation between closeness to menstruation and both, Stroop performance, $r = -0.56$, $p = .011$, and positive prime effectiveness, $r = -0.46$, $p = .610$ were found, with individuals in the Menstruation Threat/Positive Prime condition performing worst. H2 is partially supported by this paper, cognitive performance under threat did only suffer if measured on the Stroop task, further, the prime also influenced performance.

Wulandari and Hendrawan (2020) hypothesized that how gender-stereotype threat is being activated, affects performance differently, depending on gender and task difficulty. Participants were assigned to one of four groups, differing in the gender-stereotype threat activation (blatant vs. moderately explicit vs. subtle vs. control), forming the independent variables, alongside gender (male vs. female) and task difficulty (easy vs. medium vs. hard), while letter fluency performance (number of correct words/errors) and cognitive processes (clusterings/switching) were used as dependent variables. The sample consisted of 168 undergraduates ($n_{\text{female}} = 91$).

No significant effect was found for switching, neither for gender, nor for threat; for

cluster size, only a significant effect of gender was found, $F(1,159) = 4.12$, $p < .050$, $\eta_p^2 = .025$, with males scoring higher than females. Also, positive correlations between total errors and gender-stereotype score, $r = 0.22$, $p < .010$, were found, with gender-stereotype score being the result of the gender-stereotype questionnaire. H2 is not supported by this paper, more so, it provides evidence against it.

An overview of the included papers for Hypothesis 2 can be found in Table 3.

Table 3

Overview of the Included Papers for Hypothesis 2

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Guardabassi and Tomasetto (2020)	Cross-sectional	176 primary school children	N-back task	BMI, stereotype threat, working memory	Mixed-effects models	zBMI negatively correlated with working memory under threat	Partially
Hirnstein et al. (2014)	Factorial	136 participants (66 male, 70 female)	Cognitive tests	Stereotype threat, sex, group composition, cognitive performance	ANOVA	Performance decreased on 4W and perceptual speed under threat	Weakly

Table 3 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Jordano and Touron (2017)	Experimental	120 female undergraduates	OSPAN task, mind-wandering probes	Stereotype threat, mind-wandering, task performance	ANOVA	Increased mind-wandering, decreased maths performance under threat	Partially
Krendl et al. (2008)	Experimental	28 female undergraduates	fMRI	Neural activation, maths performance	Mixed-model ANOVA	Increased vACC activation, decreased cognitive region activation under threat	Yes
Lin et al. (2023)	Cross-sectional	153 female undergraduates	Spatial perspective-taking, executive function tests	Stereotype threat, executive function, spatial performance	ANCOVA, mediation analysis	Decreased performance, impaired inhibition and updating under threat	Partially

Table 3 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Rydell et al. (2014)	Experimental	340 undergraduates across 3 experiments	Executive function tasks, maths tests	Stereotype threat, executive function, maths performance	ANOVA, mediation analysis	Impaired inhibition and updating, decreased maths performance under threat	Mostly
Ståhl et al. (2012)	Experimental	335 students across 3 experiments	Stroop task, maths task	Stereotype threat, regulatory focus, cognitive control	ANOVA	Initial increase then decrease in cognitive control under threat (prevention focus)	Mostly
Wister et al. (2013)	Experimental	92 female undergraduates	Stroop test, SAT-like maths test	Menstruation threat, cognitive performance	MANOVA	Impaired Stroop performance under menstruation threat	Partially

Table 3 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Wulandari and Hendrawan (2020)	Experimental	168 undergraduates (91 female)	Letter fluency test	Stereotype threat activation, gender, task difficulty	ANOVA	No significant effects of threat on performance	No

Note. This table presents studies examining cognitive control processes under stereotype threat. The 'Variables' column includes both cognitive processes (e.g., inhibition, updating, shifting) and related performance measures. 'Methods of Data Analysis' specifies cognitive tasks used, such as the Stroop task, n-back task, or task-switching paradigms. 'Results' emphasize changes in cognitive control performance under stereotype threat conditions.

Hypothesis 3: Stereotype threat, working memory.

Bedyńska et al. (2020) investigated the relationships between chronic stereotype threat, language achievement, and domain identification in a sample of 319 male secondary school students. The study utilized chronic stereotype threat as the independent variable, while language achievement and domain identification served as dependent variables. Additionally, working memory and intellectual helplessness were examined as mediators, and gender identification was explored as a moderator. The researchers found an overall moderate level of stereotype threat in the sample, with only a slight correlation between stereotype threat and both intellectual helplessness and working memory ($r = 0.32$). Stereotype threat did negatively impact working memory capacity, with the latter mediating the relationship between stereotype threat and language achievement, $b = 2.81$, $\beta = 0.45$, $SE = 0.06$, $p < .001$, 95% CI [0.34, 0.55]. Language achievement was indirectly affected by stereotype threat, through impaired working memory and intellectual helplessness. H3 is supported by this paper; however it should be noted that it was chronic stereotype threat that was investigated and that the level of threat was only moderate.

Bedyńska et al. (2018) examined the effects of chronic stereotype threat on mathematical achievement in a sample of 624 female secondary school students. The study employed chronic stereotype threat as the independent variable and mathematical achievement as the dependent variable, with working memory and intellectual helplessness serving as mediators and gender identification as a moderator. The researchers found a slight correlation between stereotype threat and intellectual helplessness ($r = 0.20$),while maths achievement showed significant correlations with working memory ($p < .010$), and both stereotype threat and intellectual helplessness ($p < .010$), with the former being positive and the latter negative. Working memory was negatively impacted by chronic stereotype threat, $b = -0.01$, $\beta = -0.11$, $SE = 0.13$, $p = .378$, albeit not significantly. Gender identity did moderate the effect of stereotype threat on working memory, amplifying the negative effect, $b = -0.01$, $\beta = -0.29$, $SE = 0.14$, $p = .039$. H3 is supported by this paper; however, it needs to be noted that it was chronic

stereotype threat that was investigated.

Beilock et al. (2007) was already discussed in the H1 section. H3 is supported by this paper, clear impairments on maths performance are found.

Brown and Harkins (2016) investigated the effects of stereotype threat on mind-wandering in 73 female undergraduates, using the Sustained Attention to Response Task (SART). The study employed a 2 × 2 design with condition (stereotype threat vs. control) and SART framing (related to maths ability vs. unrelated to maths ability) as independent variables, and mind-wandering (measured by SART performance) as the dependent variable. The researchers hypothesized that participants under threat would show more mind-wandering, unless the SART was framed as related to maths ability, in which case they would show less mind-wandering due to increased motivation and effort. Both manipulations were successful. In the Stereotype Threat/SART Related condition, commission errors were fewer ($M = 4.89$, $SD = 3.03$), compared to Stereotype Threat/SART Unrelated ($M = 10.44$, $SD = 4.55$), $F(1, 69) = 28.78$, $p < .001$, η_p^2 0.29, anticipations were lower in the Stereotype Threat/SART Related condition ($M = 0.56$, $SD = 1.10$), compared to Stereotype Threat/SART Unrelated ($M = 4.44$, $SD = 7.09$), $F(1, 69) = 11.42$, $p < .010$, $\eta_p^2 = .142$, and RT coefficient of variation showed less variability in the Stereotype Threat/SART Related condition ($M = 258.00$, $SD = 76.00$), compared to Stereotype Threat/SART Unrelated ($M = 341.00$, $SD = 97.00$), $F(1, 69) = 11.75$, $p < .001$, $\eta_p^2 = .146$. H3 is not supported by this paper, as the mere effort account was found to be the more likely explanation for the results.

Guardabassi and Tomasetto (2020) was already discussed in the H2 section. H3 is supported by this paper, as N-back task performance was found to be impaired under stereotype threat.

Hutchison et al. (2013) examined the effects of stereotype threat on mind-wandering and Stroop task performance in a sample of 187 men (88.5% Caucasian). The study used working memory capacity (measured by OSPAN), list congruency (mostly

congruent vs. mostly incongruent), and stereotype threat condition (threat vs. control) as independent variables, while Stroop task performance (error rates and reaction times) served as the dependent variable. The researchers hypothesized that mind-wandering would increase under threat, impairing Stroop task performance, particularly in incongruent trials. While the two-way interaction between stereotype threat and working memory capacity was moderate, β = -0.11, t = -1.93, p = .054, a main effect was found for stereotype threat, β = .120, t = 2.11, $p < .050$, showing the Stroop effect to be larger under threat. Further, the interaction with list congruency was found to qualify this main effect, with mostly congruent lists, β = .240, t = 2.55, $p < .050$, showing a greater Stroop effect under threat, than mostly incongruent lists, β = -0.02, t = -0.17, p = .860, Building on this, the interaction between all independent variables was found to be significant, β = -0.12, t = -1.99, $p < .050$. Reaction times on the Stroop task were also decreased under threat to a significant degree, R^2 = .530, $F(7, 174) = 28.95$, $p < .001$.

H3 is partially supported, a significant effect of threat on Stroop performance was only found for low working memory capacity individuals. Further, working memory capacity might moderate the effect.

Jamieson and Harkins (2007) investigated the effects of stereotype threat on tasks requiring inhibitory control across four experiments. The study used condition (stereotype threat vs. control), task type (anti-saccade vs. pro-saccade), and cognitive load (2-back vs. 0-back task) as independent variables, while accuracy, reaction time (RT), and eye movements served as dependent variables. Experiment 1 displayed the target in the saccade task for 150 ms, Experiment 2, 3, and 4 for 250 ms. In Experiment 1 the sample consisted of eighty undergraduates, Experiments 2 and 3 had thirty-six female undergraduates, and Experiment 4 seventy-two. Performance on the anti-saccade task was impaired for participants under threat, compared to controls, $F(1, 72) = 17.28$, $p < .001$, d = 0.98. The interaction between Condition × Task indicated that under threat, response times were lower, to a significant degree, $F(1, 72) = 4.85$, p = .050, further, under threat, accuracy was also higher for the anti-saccade trails, $F(1, 32) = 9.06$, p = .010, d = 1.06, while it was lower for pro-saccade trails, $F(1, 32) = 8.30$, p =

.010, $d = 1.01$.

In Experiment 2, RTs were, again, faster under threat in both saccade tasks, $F(1, 32) = 19.52$, $p < .001$, $d = 1.58$). In Experiment 3, RT was lower in the stereotype threat condition, $F(1, 32) = 30.74$, $p < .001$, $d = 1.96$. Further, contrasts showed that this was true for both the anti- and pro-saccade tasks, $F(1, 32) = 29.53$, $p < .001$, $d = 1.91$ and $F(1, 32) = 43.47$, $p < .001$, $d = 2.34$, respectively - the same can be said for the adjusted reaction times from the eye movements, $F(1, 31) = 10.06$, $p = .010$, $d = 1.12$.

In Experiment 4, *n*-back task performance did not differ between controls and participants under threat, neither did accuracy. Previous results for reaction time under threat were only replicated for the 0-back condition, $F(1, 64) = 13.67$, $p = .010$, $d = 0.93$, while the 2-back condition showed a reduction in speed for participants under threat, $F(1, 64) = 12.15$, $p = .010$, $d = 0.87$. H3 is mostly not supported by this paper, only the slower reaction times under high cognitive load indicate decrease working memory speed (it is not a direct measure of working memory speed), while the other results rather support the mere effort account, which is an alternative hypothesis to the working memory account.

Johns et al. (2008) conducted a series of experiments to examine the effects of stereotype threat on working memory capacity, attention allocation to anxiety-related stimuli, and maths performance. The study used condition (stereotype threat vs. control), emotion regulation strategy, anxiety measure description, and ethnicity as independent variables, while working memory capacity, attention allocation, maths performance, and self-reported anxiety served as dependent variables. In Experiment 1, eighty-one Caucasian female students participated In the working memory task, women were able to recall fewer words under threat, resulting in a main effect of threat, $F(1, 77) = 9.53$, $p < .010$. A significant Condition $\times$ Anxiety measure description interaction was found for attention allocation, $F(1, 77) = 6.41$, $p = .010$, with more attention given to anxiety-related stimuli under threat when the task was framed as a measure of perceptual focus. Experiment 3 had sixty-one Caucasian women participating. Working memory performance did suffer under stereotype threat; however, just in the threat only condition,

$t(55) = 2.31$, $p < .050$, $d = 0.62$, the same results were found for maths test performance, $t(55) = 2.11$, $p < .050$, $d = 0.64$. Working memory performance was found to mediate the relationship between stereotype threat and maths performance, $\beta = .300$, $p < .050$. In Experiment 4, thirty-four Latino (22 women) and forty-seven Caucasians (28 women) participated. In the threat only condition, anxiety-related words received less attention by Latinos, $t(71) = 2.47$, $p < .050$, $d = 0.70$. Working memory performance was significantly impaired for Latinos under only stereotype threat, $t(71) = 2.18$, $p < .050$, $d = 0.55$. Further, Caucasians in the threat only condition recalled the most words, while Latinos and Caucasians did not significantly differ in performance in the anxiety reappraisal condition. H3 is supported by this paper.

Pennington et al. (2019) conducted two experiments to investigate the effects of positive and negative stereotypes on cognitive performance. The study used stereotype condition as the independent variable, with levels varying between experiments. Experiment 1 used self-as-target stereotype threat, group-as-target stereotype threat, and no-threat control, while Experiment 2 employed negative/positive group-as-target stereotype threat and non-threat control. The dependent variables included anti-saccade task performance measures and, in Experiment 2, maths task performance. In Experiment 1, participants were sixty-four female undergraduates. One of the manipulation checks was significant, while the other was not. For the anti-saccade task, neither SRT for correct saccades, nor their accuracy revealed a significant main effect of stereotype threat, $F(2, 58) = 0.30$, $p = .750$, $\eta_p^2 = .010$), and $F(2, 57) = 0.03$, $p = .970$, $\eta_p^2 < .001$, respectively. The same can be said for reflexive saccades and their SRTs, $F(2, 57) = 0.03$, $p = .970$, $\eta_p^2 < .001$, and $F(2, 56) = 0.25$, $p = .780$, $\eta_p^2 = .009$, respectively. While the SRTs for corrective saccades, again, did not reach significance for stereotype threat, $F(2, 53) = 0.30$, $p = .750$, $\eta_p^2 = .010$, the percentage of corrective saccades did, $F(2, 57) = 3.57$, $p = .004$, $\eta_p^2 = .110$.

The sample in Experiment 2 consisted of sixty female undergraduates. The manipulation checks were successful. On the anti-saccade task, no significant main effects were found for any of the performance measures, the same can be said for the MA task.

This paper does not support H3. In Experiment 1 there is weak evidence of a few impairments; however, overall stereotype threat did not seem to significantly alter performance.

Rydell et al. (2009) hypothesized that different social identities can mitigate the performance impairments of stereotype threat. Further, a positive social identity can lighten the depletion of working memory capacity under stereotype threat.

Gender stereotype (present vs. absent) and college student stereotype (present vs. absent) were the independent variables, while working memory capacity alongside maths performance and vowel counting accuracy were the dependent variables. Fifty-seven ($N = 57$) female undergraduates participated in the study. A significant two-way interaction was found between both stereotype conditions, $F(1, 53) = 6.01$, $p = .020$, $\eta_p^2 = .102$, without gender-stereotypes, performance did not differ between the college student stereotype conditions. Further, under gender-stereotype threat, maths performance did suffer to a significant degree. Working memory showed the same pattern, with, the two-way interaction being significant, $F(1, 53) = 4.91$, $p = .030$, $\eta_p^2 = .080$, the without gender stereotypes conditions showing no significant differences, $F(1, 27) = 2.33$, $p = .140$, $\eta_p^2 = .008$. The number of words recalled was lower, for those under gender-stereotype threat, who did not have the college student stereotype, $F(1, 26) = 31.41$, $p < .001$, $\eta_p^2 = .547$. Working memory capacity mediated the relationship of both stereotypes and maths performance, as analysed with a Sobel test, $z = 1.96$, $p = .050$. H3 is supported by this paper.

Schmader et al. (2009) conducted two experiments to examine the relationship between anxiety, stereotype threat, and working memory capacity. The study used prime condition (confidence vs. doubt) and self-reported anxiety as independent variables in both experiments, with additional variables of ethnicity in Experiment 1 and task frame in Experiment 2. Working memory performance, measured by a modified Reading Span Test, served as the dependent variable. The final sample in Experiment 1 consisted of 37 minorities (17 Hispanics, 16 African Americans, and 4 American Indian) and 40 Whites

($N = 77$). While the Prime × Ethnicity interaction did not reveal a main effect on anxiety, a marginal one was found for ethnicity, $F(1, 73) = 2.92$, $p = .090$. For working memory, the sentence reading times did not differ significantly between conditions.

In Experiment 2, the final sample consisted of 111 females (79 White, 10 Hispanic, 7 African American, 7 Asian American, 1 American Indian, and 7 unidentified; $N = 111$).

A prime × task frame interaction was found for anxiety, $F(1, 107) = 3.83$, $p = .050$, with anxiety being higher in the maths test condition, compared to the problem-solving task; however, significance was not reached, while also not reaching significance, the opposite pattern was found for women in the confidence condition. For working memory the three-way interaction among task frame, prime, and anxiety level was significant, $\beta = -0.20$, $p = .050$. If the task was framed to be diagnostic (stereotype threat), the anxiety × prime interaction was significant, $\beta = -0.30$, $p < .040$, with higher anxiety and doubt resulting in significantly lower working memory performance, while lower anxiety showed no effects, independently of the prime. H3 is partially supported by this paper, as Study 1 does not support it, while Study 2 does.

Schmader and Johns (2003) conducted three experiments to investigate the effects of stereotype threat on working memory capacity and maths test performance. The study used condition (stereotype threat vs. control) as an independent variable across all experiments, with additional variables of gender in Experiment 1 and ethnicity in Experiment 2. maths test performance and working memory capacity served as dependent variables. In Experiment 1, the final sample consisted of 51 (28 female) undergraduates. Working memory capacity under threat was also significantly lower than in the control condition, $F(1, 54) = 23.84$, $p < .001$. Women under threat recalled fewer words than any other condition, $F(1, 54) = 15.69$, $p < .001$. Further, women under threat spent longer on each equation, showing a marginal main effect of stereotype threat, $F(1, 54) = 3.44$, $p < .100$; however, no significant differences were found for accuracy (equations solved correctly) between the conditions.

Experiment 2 had a final sample of 72 ($N = 72$) undergraduates, 33 of which were Latino (20 women), and 39 White (27 women). Latinos under threat were the lowest

performing group, reaching significantly lower scores than Whites under threat, $F(1, 58) = 6.45$, $p < .050$, $d = 0.66$ and Latinos in the control condition, $F(1, 58) = 4.19$, $p < .050$, $d = 0.55$. A Gender × Stereotype Threat interaction was also found, $F(1, 58) = 5.21$, $p < .050$. In Experiment 3, the final sample consisted of 28 female undergraduates ($N = 28$). Under stereotype threat, while absolute span scores were lower, $t(26) = 3.13$, $p < .010$, $d = 1.19$. Under threat, accuracy was impaired, $t(26) = 2.38$, $p < .050$, while effort (number of attempted problems) was not. Regression analysis revealed that working memory capacity significantly predicted maths performance , $\beta = .580$, $t(26) = 3.26$, $p < .010$, while the direct effect of threat became non-significant when controlling for working memory, $\beta = -0.12$, $t(26) = -0.66$, $p = .500$. A Sobel test confirmed the mediating role of working memory ($z = 2.26$, $p < .020$). H3 is supported by this paper.

Tine and Gotlieb (2013) examined how different types of stereotype threat (gender-, race-, and income-based) affect maths performance and working memory, and whether these threats can have cumulative effects. The study used gender (male vs. female), race (White vs. racial/ethnic minority), income level (low, middle, high), and number of stigmatized aspects of identity (0 to 3) as independent variables, while post-test scores of both maths performance and working memory performance served as dependent variables. Seventy-one undergraduates ($N = 71$) participated, 46 female, 24 racial/ethnic minority (17 Black/African American, 7 Hispanic/Latino), and 15 low-income, all of these were considered to have a stigmatized aspect of identity, with some having multiple.

For gender-based stereotype threat, working memory performance was affected significantly, $F(1, 68) = 4.91$, $p < .050$, $\eta^2_p = .067$, but not maths performance Under race-based stereotype threat, both maths and working memory performance were impaired, $F(1, 68) = 16.73$, $p < .001$, $\eta^2_p = .197$ and $F(1, 68) = 7.41$, $p < .001$, $\eta^2_p = .098$, respectively. Both, maths, $F(2, 67)$[1] $= 5.92$, $p < .010$, $\eta^2_p = .150$ and working memory performance, $F(2, 67) = 4.92$, $p < .050$, $\eta^2_p = .128$, were also lower under income-based

[1] corrected to plausible values for F, since it was reported as $F(12\ 67)$

threat. The number of stigmatized aspects significantly predicted both maths and working memory performance, with participants having three stigmatized aspects performing significantly worse than those with zero, particularly in working memory, $F(3, 66) = 6.82$, $p < .001$, $\eta^2_p = .227$. The amount of effort did not differ under gender- or race-based threat; however, low-income participants reported to have put in more effort into the task, so did the two stigmatized aspects group; each compared to their counterparts. H3 is supported by this paper.

Van Loo and Rydell (2013) investigated how power priming might shield women from the negative effects of stereotype threat on maths performance, hypothesizing that this protection would occur through maintained working memory capacity. The study employed power prime (high, low, control) and stereotype threat condition (stereotype threat vs. no threat) as independent variables, while working memory capacity, maths performance, and threat-based concern served as dependent variables. The Positive and Negative Affect Schedule (PANAS) was used as a control variable.

The final sample consisted of 131 ($N = 131$) female undergraduates.

Both, stereotype threat, $F(1, 125) = 39.46$, $p < .001$, $\eta^2_p = .240$, and power prime, had a significant main effect on working memory capacity, the two-way interaction qualified these results, $F(2, 125) = 13.38$, $p < .001$, $\eta^2_p = .180$). Under threat, working memory was lower for women with low power or in the control condition, $F(1, 125) = 50.05$, $p < .001$, $\eta^2_p = .270$, $F(1, 125) = 15.50$, $p < .001$, $\eta^2_p = .110$, respectively. Maths performance showed similar results across the board. Working memory capacity mediated the relationship between power and stereotype threat, $\beta = -0.36$, $p < 0.00$, further a positive and significant correlation between maths performance and working memory was found, $\beta = .620$, $p < 0.00$. Using a Sobel test, it was found that, in the relationship between maths performance and the power $\times$ stereotype threat interaction, working memory capacity was a significant mediator, $z = -3.53$, $p < .001$. H3 is mostly supported by this hypothesis, only when primed with high power, stereotype threat did not impair working memory capacity.

An overview of the included papers for Hypothesis 2 can be found in Table 4.

Table 4

Overview of the Included Papers for Hypothesis 3

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Bedyńska et al. (2020)	Cross-sectional	319 male secondary school students	Working memory tasks	Chronic stereotype threat, working memory, language achievement	Mediation analysis	Stereotype threat negatively impacted working memory capacity	Yes
Bedyńska et al. (2018)	Cross-sectional	624 female secondary school students	Working memory tasks	Chronic stereotype threat, working memory, maths achievement	Mediation analysis	Working memory mediated stereotype threat and maths achievement	Yes

Table 4 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Beilock et al. (2007)	Experimental	Female college students in US	Modular Arithmetic task	Stereotype threat, working memory efficiency	ANOVA	Reduced performance on high-demand problems under threat	Yes
Brown and Harkins (2016)	Experimental	73 female undergraduates	SART, maths test	Stereotype threat, SART framing, mind-wandering	ANOVA	Support for mere effort account, not working memory impairment	No
Guardabassi and Tomasetto (2020)	Cross-sectional	176 primary school children	N-back task	BMI, stereotype threat, working memory	Mixed-effects models	zBMI negatively correlated with working memory under threat	Yes

Table 4 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Hutchison et al. (2013)	Experimental	187 men	Stroop task, OSPAN	Working memory capacity, stereotype threat, Stroop performance	Regression analysis	Stroop effect larger under threat for low WMC individuals	Partially
Jamieson and Harkins (2007)	Experimental	224 undergraduates across 4 experiments	Saccade tasks, N-back task	Stereotype threat, task type, cognitive load	ANOVA	Support for mere effort account in most conditions	Mostly No
Johns et al. (2008)	Experimental	176 participants across 3 experiments	Working memory task, maths test	Stereotype threat, emotion regulation, working memory	ANOVA, mediation analysis	Working memory impaired under threat, mediated maths performance	Yes

Table 4 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Pennington et al. (2019)	Experimental	124 female university students	Anti-saccade task, maths task	Stereotype condition, task performance	ANOVA	No significant effects of threat on performance	No
Rydell et al. (2009)	Experimental	57 female undergraduates	Vowel counting task, maths problems	Gender stereotype, college student stereotype, working memory	ANOVA, mediation analysis	Working memory capacity mediated stereotype effects on maths performance	Yes
Schmader et al. (2009)	Experimental	188 participants across 2 experiments	Reading Span Test	Stereotype threat, anxiety, working memory	Regression analysis	Anxiety predicted lower working memory under stereotype threat	Partially

Table 4 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Schmader and Johns (2003)	Experimental	159 undergraduates across 3 experiments	OSPAN, maths test	Stereotype threat, working memory capacity, maths performance	ANCOVA, mediation analysis	Working memory capacity reduced under threat, mediated maths performance	Yes
Tine and Gotlieb (2013)	Experimental	71 undergraduates	Maths test, working memory test	Multiple stereotype threats, maths and working memory performance	ANOVA	Working memory impaired under various stereotype threats	Yes

Table 4 continued

Citation	Study Design	Population	Research Questions	Variables	Methods of Data Analysis	Results	Hypothesis confirmed
Van Loo and Rydell (2013)	Experimental	131 female undergraduates	Letter-memory task, maths test	Power prime, stereotype threat, working memory	ANOVA, mediation analysis	High power prime protected working memory from stereotype threat effects	Mostly

Note. This table outlines studies investigating working memory impairment under stereotype threat. The 'Variables' column focuses on working memory measures and associated performance indicators. 'Methods of Data Analysis' details specific working memory tasks employed, such as complex span tasks, operational span tasks, or reading span tests. 'Results' highlight changes in working memory capacity and performance under stereotype threat.

Discussion

References

6.36 Decimal Fractions. (2020). In *Publication manual of the American Psychological Association, 7th ed.* (pp. 179–180). American Psychological Association.

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

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

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

* Bedyńska, S., Krejtz, I., Rycielski, P., & Sedek, G. (2020). Stereotype threat is linked to language achievement and domain identification in young males: Working memory and intellectual helplessness as mediators. *Psychology in the Schools*, *57*(9), 1331–1346. https://doi.org/10.1002/pits.22413

* Bedyńska, S., Krejtz, I., & Sedek, G. (2018). Chronic stereotype threat is associated with mathematical achievement on representative sample of secondary schoolgirls: The role of gender identification, working memory, and intellectual helplessness. *Frontiers in Psychology*, *9*, 428. https://doi.org/10.3389/fpsyg.2018.00428

* 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

* Brown, A. J., & Harkins, S. G. (2016). Threat does not make the mind wander: Reconsidering the effect of stereotype threat on mind-wandering. *Motivation Science*, *2*(2), 85–96. https://doi.org/10.1037/mot0000032

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.

* Guardabassi, V., & Tomasetto, C. (2020). Weight status or weight stigma? Obesity stereotypes—not excess weight—reduce working memory in school-aged children. *Journal of Experimental Child Psychology*, *189*, 104706. https://doi.org/10.1016/j.jecp.2019.104706

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). *PRISMA2020* : An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

* Hirnstein, M., Coloma Andrews, L., & Hausmann, M. (2014). Gender-stereotyping and cognitive sex differences in mixed- and same-sex groups. *Archives of Sexual Behavior*, *43*(8), 1663–1673. https://doi.org/10.1007/s10508-014-0311-5

* Hutchison, K. A., Smith, J. L., & Ferris, A. (2013). Goals can be threatened to extinction: Using the stroop task to clarify working memory depletion under stereotype threat. *Social Psychological and Personality Science*, *4*(1), 74–81. https://doi.org/10.1177/1948550612440734

* Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology*, *93*(4), 544–564. https://doi.org/10.1037/0022-3514.93.4.544

* 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

* 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

* Jordano, M. L., & Touron, D. R. (2017). Priming performance-related concerns induces task-related mind-wandering. *Consciousness and Cognition*, *55*, 126–135. https://doi.org/10.1016/j.concog.2017.08.002

* Krendl, A. C., Richeson, J. A., Kelley, W. M., & Heatherton, T. F. (2008). The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math. *Psychological Science*, *19*(2), 168–175. https://doi.org/10.1111/j.1467-9280.2008.02063.x

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

* Lin, Y., Zhang, B., Jin, D., Zhang, H., & Dang, J. (2023). The effect of stereotype threat on females' spatial perspective taking and the mediating role of executive functions. *Current Psychology*, *42*(6), 4979–4990. https://doi.org/10.1007/s12144-021-01849-7

* Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., & Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*, *7*(2), 230–241. https://doi.org/10.1093/scan/nsq100

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

* 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

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/

* Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology*, *96*(5), 949–966. https://doi.org/10.1037/a0014846

* Rydell, R. J., Van Loo, K. J., & Boucher, K. L. (2014). Stereotype threat and executive functions: Which functions mediate different threat-related outcomes? *Personality and Social Psychology Bulletin*, *40*(3), 377–390. https://doi.org/10.1177/0146167213513475

* Schmader, T., Forbes, C. E., Shen Zhang, & Berry Mendes, W. (2009). A metacognitive perspective on the cognitive deficits experienced in intellectually threatening environments. *Personality and Social Psychology Bulletin*, *35*(5), 584–596. https://doi.org/10.1177/0146167208330450

* Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440

* Ståhl, T., Van Laar, C., & Ellemers, N. (2012). The role of prevention focus under stereotype threat: Initial cognitive mobilization is followed by depletion. *Journal of Personality and Social Psychology*, *102*(6), 1239–1251. https://doi.org/10.1037/a0027678

* Tine, M., & Gotlieb, R. (2013). Gender-, race-, and income-based stereotype threat: The effects of multiple stigmatized aspects of identity on math performance and working memory function. *Social Psychology of Education*, *16*(3), 353–376. https://doi.org/10.1007/s11218-013-9224-8

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

* Van Loo, K. J., & Rydell, R. J. (2013). On the experience of feeling powerful: Perceived power moderates the effect of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *39*(3), 387–400. https://doi.org/10.1177/0146167212475320

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*.

* Wister, J. A., Stubbs, M. L., & Shipman, C. (2013). Mentioning menstruation: A stereotype threat that diminishes cognition? *Sex Roles*, *68*(1-2), 19–31. https://doi.org/10.1007/s11199-012-0156-0

* Wu, X., & Zhao, Y. (2021). Degree centrality of a brain network is altered by stereotype threat: Evidences from a resting-state functional magnetic resonance imaging study. *Frontiers in Psychology*, *12*, 705363. https://doi.org/10.3389/fpsyg.2021.705363

* Wulandari, S. W., & Hendrawan, D. (2020). Trust your abilities more than the

stereotype: Effect of gender-stereotype threat and task difficulty on word production, clustering, and switching in letter fluency. *Pertanika Journal of Social Sciences and Humanities*, *28*(4), 2567–2588. https://doi.org/10.47836/pjssh.28.4.05

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