Pennington et al. (2019)

EPPI-Centre (2003) & Critical Appraisal Skills Programme (2018)

If the study has a broad focus and this data extraction focuses on just one component of the study, please specify this here

☒ Not applicable (whole study is focus of data extraction)

☐ Specific focus of this data extraction (please specify)

Study aim(s) and rationale

Was the study informed by, or linked to, an existing body of empirical and/or theoretical research?

Please write in authors' declaration if there is one. Elaborate if necessary, but indicate which aspects are reviewers' interpretation.

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

- Stereotype threat
- Mere effort account of stereotype threat
- multi-threat framework of stereotype threat (Shapiro & Neuberg, 2007)
- Stereotype threat and working memory inference account

Do authors report how the study was funded?

☐ Explicitly stated (please specify)

☐ Implicit (please specify)

☒ Not stated/unclear (please specify)

Study research question(s) and its policy or practice focus

What is/are the topic focus/foci of the study?

Utilising an anti-saccade eye-tracking paradigm, Experiment 1 examined the influence of a negative self- and group-relevant stereotype threat on women's inhibitory control (termed 'visiospatial performance' in Jamieson & Harkins, 2007).

What is/are the population focus/foci of the study?

- women under either self-as-target or group-as-target stereotype threat

What is the relevant age group?

☐ Not applicate (focus not learners)

☐ 0 - 4

☐ 5 - 10

☐ 11 - 16

☐ 17 - 20

☐ 21 and over

☒ Not stated/unclear

What is the sex of the population focus/foci?

☐ Not applicate (focus not learners)

☒ Female only

☐ Male only

☐ Mixed sex

☐ Not stated/unclear

What is/are the educational setting(s) of the study?

☐ Community centre

☐ Correctional institution

☐ Government department

☐ Higher education institution

☐ Home

☐ Independent school

☐ Local education authority

☐ Nursery school

☐ Other early years setting

☐ Post-compulsory education institution

☐ Primary school

☐ Residential school

☐ Secondary school

☐ Special needs school

☐ Workplace

☐ Other educational setting

In Which country or cuntries was the study carried out?

☒ Explicitly stated (please specify)

☐ Not stated/unclear (please specify)

- UK

Please describe in more detail the specific phenomena, factors, services, or interventions with which the study is concerned

What are the study reserach questions and/or hypotheses?

Research questions or hypotheses operationalise the aims of the study. Please write in authors' description if there is one. Elaborate if necessary, but indicate which aspects are reviewers' interpretation.

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

Experiment 1:
Experimental predictions were two-tailed, allowing us to pit the mere effort account against the working memory inference account of stereotype threat.
The mere effort account predicts that participants primed with a negative stereotype should make more reflexive eye movements towards the target (incorrect saccades) relative to the control condition because increased motivation facilitates the dominant response. Additionally, it predicts that this heightened motivation will influence stereotype-threat participants to launch quicker correct saccades (i.e. eye movements directed correctly away from the target) and quicker corrective saccades (i.e. eye movements directed to the correct location after an incorrect response) compared to participants who are not subject to evaluation.
The working memory inference account also predicts that participants who are primed with a negative stereotype will launch more incorrect saccades towards the target relative to control participants. However, in contrast to the mere effort account, this theory predicts that stereotype-threatened participants should launch slower correct saccades and be less likely to correct for incorrect responses owning to diminished working memory capacity (Rydell et al., 2014).

Experiment 2:
In line with a welth of previous research, it was predicted that women primed with a negative group stereotype would solve fewer mathematical problems compared to the control condition. This is particularly the case for difficult mathematical problems presented horizontally relative to vertically because such problems have shown to place greater demands on verbal working memory.
Furthermore, it was hypothesised that a positive group stereotype threat might facilitate women's performance on simple problems becaue they are motivated to perform well, but diminish their performance on difficult problems because this heightened expectation for success influences them to 'choke under pressure'.

Methods - Design

Which variables or concepts, if any, does the study aim to measure or examine?

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

Experiment 1:
- correct response time (RT) - correct % - corrective % - corrective RT

Experiment 2:
- MA taks performance - pro- and anti-saccade trails - eye-trakcing data

Study timing

Please indicate all that apply and give further details where possible.

If the study examines one or more samples, but each at only one point in time it is cross-sectional.
If the study examines the same samples, but as they have changed over time, it is retrospective, provided that the interest is in starting at one timepoint and looking backwards over time.
If the study examines the same samples as they have changed over time and if data are collected forward over time, it is prospective provided that the interest is in starting at one timepoint and looking forward in time.

☒ Cross-sectional

☐ Retrospective

☐ Prospective

☐ Not stated/unclear (please specify)

If the study is an evaluation, when were measurements of the variable(s) used for outcome made, in relation to the intervention?

If at least one of the outcome variables is measured both before and after the intervention, please use the before and after category.

☐ Not applicable (not an evaluation)

☐ Before and after

☐ Only after

☐ Other (please specify)

☐ Not stated/unclear (please specify)

Methods - Groups

If comparisons are being made between two or more groups, please specify the basis of any divisions made for making these comparisons.

Please give further details where possible.

☐ Not applicable (not more than one group)

☒ Prospecitive allocation into more than one group (e.g. allocation to different interventions, or allocation to intervention and control groups)

☐ No prospective allocation but use of pre-existing differences to create comparison groups (e.g. receiving different interventions, or characterised by different levels of a variable such as social class)

☐ Other (please specify)

☐ Not stated/unclear (please specify)

How do the groups differ?

☐ Not applicable (not more than one group)

☒ Explicityly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

Experiment 1:
- self-as-target stereotype threat, group-as-target stereotype threat and non-threat control

Experiment 2:
- negative group-as-target stereotype; (ii) positive group-as-target stereotype; and (iii) a non-threat control condition

Number of groups

For instance, in studies in which comparisons are made between groups, this may be the number of groups into which the dataset is divided for analysis (e.g. social class, or form size), or the number of groups allocated to, or receiving, an intervention.

☐ Not applicable (not more than one group)

☐ One

☐ Two

☒ Three

☐ Four or more (please specify)

☐ Other/unclear (please specify)

Was the assignment of participants to interventions randomised?

☐ Not applicable (not more than one group)

☐ Not applicate (no prospective allocation)

☒ Random

☐ Quasi-random

☐ Non-random

☐ Not stated/unclear (please specify)

Where there was prospective allocation to more than one group, was the allocation sequence concealed from participants and those enrolling them until after enrolment?

Bias can be introduced, consciously or otherwise, if the allocation of pupils or classes or schools to a programme or intervention is made in the knowledge of key characteristics of those allocated. For example: children with more serious reading difficulty might be seen as in greater need and might be more likely to be allocated to the 'new' programme, or the opposite might happen. Either would introduce bias.

☐ Not applicable (not more than one group)

☐ Not applicable (no prospective allocation)

☒ Yes (please specify)

☐ No (please specify)

☐ Not stated/unclear (please specify)

Apart from the experimental intervention, did each study group receive the same level of care (that is, were they treated equally)?

☒ Yes
☐ No
☐ Can't tell

Study design summary

In addition to answering the questions in this section, describe the study design in your own words. You may want to draw upon and elaborate the answers you have already given.

Experiment 1:
1. recruited for a study that examined ostensibly factors relating to problem solving 2. seated in front of the eye-tracker + calibrations 3. Task information that corresponded to their assigned condition 4. instructions for anti- and pro-saccade trials 5. manipulation check 6. verbal and written debrief

Experiment 2:
- Equivalent to Experiment 1, with the exception of the addition of the MA task, which was presented in a counterbalanced order with the anti-saccade task - Manipulation check, this time pertaining to the MA task

Methods - Sampling strategy

Are the authors trying to produce findings that are representative of a given population?

Please write in authors' description. If authors do not specify please indicate reviewers' interpretation.

☐ Explicitly stated (please specify)

☒ Implicit (please specify)

☐ Not stated/unclear (please specify)

- individuals under stereotype threat

Which methods does the study use to identify people or groups of people to sample from and what is the sampling frame?

e.g. telephone directory, electoral register, postcode, school listing, etc. There may be two stages – e.g. first sampling schools and then classes or pupils within them.

☐ Not applicable (please specify)
☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☒ Not stated/unclear (please specify)

Which methods does the study use to select people or groups of people (from the sampling frame)?

e.g. selecting people at random, systematically - selecting for example every 5th person, purposively in order to reach a quota for a given characteristic.

☐ Not applicable (no sampling frame)
☒ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Not stated/unclear (please specify)

Planned sample size

If more than one group please give details for each group separately.

☐ Not applicable (please specify)
☒ Explicitly stated (please specify)
☐ Not stated/unclear (please specify)

The power analysis indicated that a sample sizeo f 66 participants was required to detect this lowest reported effect size of 80% power.
Bayesian analyses were also utilised in addition to NHST to overcome limitations posted by inferences of statistical power.

Methods - Recruitment and consent

Which methods are used to recruit people into the study?

e.g. letters of invitation, telephone contact, face-to-face contact.

☐ Not applicable (please specify)

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

- recruited via a university in the UK

Were any incentives provided to recruit people into the study?

☐ Not applicable (please specify)

☒ Explicitly stated (please specify)

☐ Not stated/unclear (please specify)

- Course credit or monetary remuneration for their time.

Was consent sought?

Please comment on the quality of consent if relevant.

☐ Not applicable (please specify)
☐ Participant consent sought
☐ Parental consent sought
☐ Other consent sought
☐ Consent not sought
☒ Not stated/unclear (please specify)

Are there any other details relevant to recruitment and consent?

☒ No
☐ Yes (please specify)

Methods - Actual sample

What was the total number of participants in the study (the actual sample)?

If more than one group is being compared please give numbers for each group.

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Not stated/unclear (please specify)

Experiment 1:
- Sixty-four females were successfully recruited (M_{age} = 22 years, SD = 5.53; 87.5% White British) from a university in the United Kingdom and received course credits or monetary remuneration for their time.
- Of the sample, 95.3% were university students, with the majority studying Psychology (40.6%) or Health and Social Sciences (54.7%), and all spoke English as their first language.
- They were assigned randomly to one of three stereotype conditions: (i) self-as-target stereotype threat (n = 21); (ii) group-as-target stereotype threat (n = 23); and (iii) a non-threat condition (n = 20).

Experiment 2:
- Sixty female participants ($M_{age} = 21 years, SD = 5.87; 98.3% White British) were successfully recruited from the same UK university (66.7 Psychology students) and received course credit or monetary remuneration for their time - They were assinged equally to one of three conditions (n = 20 in each): (i) negative group-as-target stereotype threat; (ii) positive group-as-target stereotype threat; and (iii) a non-threat control condition.

What is the proportion of those selected for the study who actually participated in the study?

Please specify numbers and percentages if possible.

☐ Not applicable (e.g. study of policies, documents, etc)
☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☒ Not stated/unclear (please specify)

Which country/countries are the individuals in the actual sample from?

If UK, please distinguish between England, Scotland, N. Ireland, and Wales if possible. If from different countries, please give numbers for each. If more than one group is being compared, please describe for each group.

☐ Not applicable (e.g. study of policies, documents, etc)

☐ Explicitly stated (please specify)

☒ Implicit (please specify)

☐ Not stated/unclear (please specify)

- United Kingdom

What ages are covered by the actual sample?

Please give the numbers of the sample that fall within each of the given categories. If necessary, refer to a page number in the report (e.g. for a useful table). If more than one group is being compared, please describe for each group. If follow-up study, age at entry to the study.

☐ Not applicable (e.g. study of policies, documents, etc)

☐ 0 to 4
☐ 5 to 10
☐ 11 to 16
☐ 17 to 20
☒ 21 and over
☐ Not stated/unclear (please specify)

Experiment 1:
- $M_{age} = 22 years, SD = 5.53

Experiment 2:
- $M_{age} = 21 years, SD = 5.87

What is the socio-economic status of the individuals within the actual sample?

If more than one group is being compared, please describe for each group.

☐ Not applicable (e.g. study of policies, documents, etc)
☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☒ Not stated/unclear (please specify)

What is the ethnicity of the individuals within the actual sample?

If more than one group is being compared, please describe for each group.

☐ Not applicable (e.g. study of policies, documents, etc)
☐ Explicitly stated (please specify)
☒ Implicit (please specify)
☐ Not stated/unclear (please specify)

Experiment 1:
- 87.5% White British, rest not stated

Experiment 2:
- 98.3% White British, rest not stated

What is known about the special educational needs of individuals within the actual sample?

e.g. specific learning, physical, emotional, behavioural, intellectual difficulties.

☐ Not applicable (e.g. study of policies, documents, etc)
☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☒ Not stated/unclear (please specify)

Is there any other useful information about the study participants?

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Explicitly stated (please specify no/s.)
☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

Experiment 1:
- Of this sample, 95.3% were university students, with the majority studying Psychology (40.6%) or Health and Social Sciences (54.7%), and all spoke English as a first language

Experiment 2:
- 66.7% of the sample were Psychology students

How representative was the achieved sample (as recruited at the start of the study) in relation to the aims of the sampling frame?

Please specify basis for your decision.

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Not applicable (no sampling frame)
☐ High (please specify)
☐ Medium (please specify)
☐ Low (please specify)
☐ Unclear (please specify)

If the study involves studying samples prospectively over time, what proportion of the sample dropped out over the course of the study?

If the study involves more than one group, please give drop-out rates for each group separately. If necessary, refer to a page number in the report (e.g. for a useful table).

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Not applicable (not following samples prospectively over time)
☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Not stated/unclear

For studies that involve following samples prospectively over time, do the authors provide any information on whether and/or how those who dropped out of the study differ from those who remained in the study?

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Not applicable (not following samples prospectively over time)
☐ Not applicable (no drop outs)
☐ Yes (please specify)
☐ No

If the study involves following samples prospectively over time, do authors provide baseline values of key variables such as those being used as outcomes and relevant socio-demographic variables?

☐ Not applicable (e.g. study of policies, documents, etc)
☒ Not applicable (not following samples prospectively over time)
☐ Yes (please specify)

☐ No

Methods - Data collection

Please describe the main types of data collected and specify if they were used (a) to define the sample; (b) to measure aspects of the sample as findings of the study?

☐ Details

Experiment 1:
- perceived maths ability and domain identification -> a - anti-saccade eye tracking task -> b - manipulation check -> b

Experiment 2:
- same as Experiment 1 - MA task -> b

Which methods were used to collect the data?

Please indicate all that apply and give further detail where possible.

☐ Curriculum-based assessment
☐ Focus group
☐ Group interview
☐ One to one interview (face to face or by phone)
☐ Observation
☐ Self-completion questionnaire
☐ Self-completion report or diary
☐ Exams
☐ Clinical test
☐ Practical test
☐ Psychological test
☐ Hypothetical scenario including vignettes
☐ School/college records (e.g. attendance records etc)
☐ Secondary data such as publicly available statistics
☐ Other documentation
☐ Not stated/unclear (please specify)

Details of data collection methods or tool(s).

Please provide details including names for all tools used to collect data and examples of any questions/items given. Also please state whether source is cited in the report.

☒ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Not stated/unclear (please specify)

Experiment 1:
Anti-saccade eye tracking task: - developed using Experiment Builder (SR Research Ltd) - eye movements were recorded using an EyeLink 1000 desktop eye-tracker, with a sampling

rate of 1,000 Hz
Manipulation check: - 'To what extent are there gender differences in visiospatial performance?' - 'Who do you believe performs better on this task?' -> both 1 to 10 scale -> Jamieson & Harkins, 2007
Condition manipulation:
- Non-threat control: Participants were informed that their anti-saccade performance would not be evaluated (cf. Steele & Davies, 2003) and that the experiment was investigating the role of working memory (Schmader & Johns, 2003) - Self-as-target stereotype threat: Participants were informed of the negative gender-related stereotype based on research suggesting that participants should be knowledgeable of a negative stereotype in order to be susceptible to stereotype threat (Shapiro & Neuberg, 2007). They were informed that the task would be diagnostic of their personal ability (cf. Shapiro & Neuberg, 2007) - Group-as-target stereotype threat: Participants were primed with the identical manipulation employed by Jamieson and Harkins (2007, p. 548). They were informed that their task performance would be a diagnostic indicator of gender-related ability (cf. Aronoson et al., 1999; Shapiro & Neuberg, 2007)

Experiment 2:
Manipulation check: - same as in Experiment 1 pertained to the MA task - 'To what extent are there gender differences in maths performance' - 'Who do you believe performs better on this task?'
MA task: - MA task (in accordance with previous research; see Beilock et al., 2007; Beilock & Carr, 2005; Seitchik & Harkins, 2015)
Condition manipulation: - negative group-as-target prime and the control prime were identical as those used in Experiment 1 - Positive group-as-target stereotype: Prime suggested that women typically outperform men on tests of visuospatial and mathematical ability.
Rest: - equal to Experiment 1

Who collected the data?

Please indicate all that apply and give further detail where possible.

- ☐ Researcher
- ☐ Head teacher/Senior management
- ☐ Teaching or other staff
- ☐ Parents
- ☐ Pupils/students
- ☐ Governors
- ☐ LEA/Government officials
- ☐ Other education practitioner
- ☐ Other (please specify)
- ☐ Not stated/unclear

Do the authors describe any ways they addressed the reliability of their data collection tools/methods?

e.g. test-retest methods (Where more than one tool was employed please provide details for each.)

☐ Details

Do the authors describe any ways they have addressed the validity of their data collection tools/methods?

e.g. mention previous validation of tools, published version of tools, involvement of target population in development of tools. (Where more than one tool was employed please provide details for each.)

☐ Details

Was there concealment of study allocation or other key factors from those carrying out measurement of outcome – if relevant?

Not applicable – e.g. analysis of existing data, qualitative study. No – e.g. assessment of reading progress for dyslexic pupils done by teacher who provided intervention. Yes – e.g. researcher assessing pupil knowledge of drugs - unaware of pupil allocation.

☐ Not applicable (please say why)
☐ Yes (please specify)
☐ No (please specify)

Where were the data collected?

e.g. school, home.

☐ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Unclear/not stated (please specify)

Are there other important features of data collection?

e.g. use of video or audio tape; ethical issues such as confidentiality etc.

☐ Details

Methods - Data analysis

Which methods were used to analyse the data?

Please give details e.g. for in-depth interviews, how were the data handled? Details of statistical analysis can be given next.

☒ Explicitly stated (please specify)
☐ Implicit (please specify)
☐ Not stated/unclear (please specify)

Experiment 1:
- In accordacne with trimming and exclusion criteria reported by Jamieson and Harkins (2007; Experiment 3, pp. 553), filters were used prior to data analysis to ensure that eye movements recorded by the eye tracker represented responses to the stimuli presented - Specifically, the initial four practice trails were removed from analyses, resulting in a total

of 160 trails for each participants - Eye movements were categorised as valid if participants' initial eye position did not vary by more than 2.82° (50 pixels) from the central fixation cross - Eye movements more than 2.82° were considered invalid and were removed from the analysis - A total of 3% of pro-saccade and 3% of anti-saccade trails across all participants were excluded using this criterion - Eye movements were classed as anticipatory if participants initiated saccades in less than 80 ms and saccades beginning at 1,000 ms or greater were excluded because they could not have been initiated in response to the target (Crevit & Vandierendonck, 2005; Jamieson & Harkins, 2007). - This criterion resulted in the exclusion of an additional 3% of anti-saccade trails and 6% of pro-saccade trails - As a total, 9% of pro-saccade and 6% of anti-saccade trails were removed from the analysis - Data from four participants were excluded from the overall analysis because of invalid centre starts and calibration errors on the anti-saccade task (resulting in n = 60 participants)

Experiment 2:
- trimming and exclusion criteria followed those reported by Jamieson and Harkins (2007) - total of 4% of pro-saccade and 5% of anti-saccade trails were excluded because initial saccades exceeded 2.82° - an additional 6% of pro-saccade trails and 3% of anti-saccade trails were excluded because participants initiated saccades less than 80 ms or greater than 1,000 ms. - Eye tracking data from three participants were removed due to excessive invalid central starts and calibration error - Mathematical accuracy data from four participants were excluded from analyses because they responed with below chance performance.

Which statistical methods, if any, were used in the analysis?

☐ Details

Experiment 1:
Self-reported maths ability and domain identification: - Means and SD calculations
Analysis strategy: - NHST and Bayesian statistics - ANOVA with Bonferroni-corrected pairwise comparisons - p-values and associated 95% CI comparisons (partial-eta squared and Cohen's d) - Bayes factors (B)
Anti-saccade task: - Two separate analyses were conducted on correct saccades and corresponding saccadic reaction time (SRT) as a function of trail-type (pro and anti-saccade)
Anti-saccade trails: - series of analyses were conducted for percentage accuracy and ART of reflexive, corrective and correct saccades as a function of stereotype threat condition

Experiment 2:
- see above - see Experiment 1
Mathematical performance: - 3 (condition) x (difficulty) x (orientation) mixed-design ANOVA was conducted on Ma accuracy scores
Bayesian meta-analysis: - fixed-effects meta-analysis was conducted using Dienes' (2008) calculator to test the main experimental hypotheses that priming a negative group stereotype has a detrimental impact on women's inhibitory control performance. - Internal meta-analyses provide a measure of the total weight of evidence across studies - Only direct comparisons between the conditions matching those in Jamieson and Harkins (2007, Experiment 3) were included in the meta analysis

What rationale do the authors give for the methods of analysis for the study?

e.g. for their methods of sampling, data collection, or analysis.

☐ Details

For evaluation studies that use prospective allocation, please specify the basis on which data analysis was carried out.

'Intention to intervene' means that data were analysed on the basis of the original number of participants as recruited into the different groups. 'Intervention received' means data were analysed on the basis of the number of participants actually receiving the intervention.

☐ Not applicable (not an evaluation study with prospective allocation)
☐ 'Intention to intervene'
☐ 'Intervention received'
☐ Not stated/unclear (please specify)

Do the authors describe any ways they have addressed the reliability of data analysis?

e.g. using more than one researcher to analyse data, looking for negative cases.

☐ Details

Do the authors describe any ways they have addressed the validity of data analysis?

e.g. internal or external consistency; checking results with participants.

☐ Details

Do the authors describe strategies used in the analysis to control for bias from confounding variables?

☐ Details

Please describe any other important features of the analysis.

☐ Details

Please comment on any other analytic or statistical issues if relevant.

☐ Details

Results and Conclusions

How are the results of the study presented?

e.g. as quotations/figures within text, in tables, appendices.

☐ Details

- in text
- table(s)

What are the results of the study as reported by authors?

Please give details and refer to page numbers in the report(s) of the study where necessary (e.g. for key tables).

☐ Details

Experiment 1:
self-reported maths skills and domain identification: - these two factors were not found to moderate stereotype threat effects in any of the forthcoming analyses
Stereotype threat manipulation check: - No significant main effect of stereotype condition on participants' responses to the first manipulation check - Bayes factors indicated weak support for the alternative hypothesis when evaluating whether participants in the self-as-target and group-as-target conditions endorsed that there were gender differences in visuospatial performance relative to the control condition - Significant main effect of stereotype condition on participant's responses to the second manipulation check - Participants in the group-as-target stereotype threat condition were more likely to report that men outperformed women relative to the control condition - Moderate evidence for the difference between the self-as-target condition compared to the control condition on this measure *Anti-saccade task*: - Significant main effect of accuracy, with participants responding more accurately to pro-saccade relative to anti-saccade trials - Significant main effect of response time, with participants expectedly faster at responding to pro-saccade relative to anti-saccade trails
Anti-saccade trails: - No significant differences on any of the dependent variables as a function of the stimuli used (i.e. numbers vs. shape) and therefore this variable was collapsed within all analyses.
Correct saccades: - No significant main effect of stereotype condition on the percentage of correct anti-saccades - Bayes factors indicated moderate evidence for the null hypothesis when comparing the self-as-target and group-as-target conditions to the control - No significant main effect of stereotype condition on SRT for correct saccades - Bayes factors indicate noteworthy evidence for the null hypothesis when comparing SRT for the self-as-target and group-as-target condition to the control
Reflexive saccades (incorrect responses): - no significant main effect of stereotype condition on the percentage of reflexive saccades - Bayes factors indicated strong evidence for the null hypothesis when comparing data for the self-as-target and group-as-target conditions to the control - No significant main effect of stereotype condition on SRT of reflexive saccades - Bayes factors indicated weak evidence for the null hypothesis when comparing data for the self-as-target and group-as-target condition to the control condition
Corrective saccades: - significant main effect of stereotype condition of the proportion of corrective saccades, pairwise comparisons between conditions were non-significant - Bayes factors indicated strong evidence for the null hypothesis when comparing the self-as-target and group-as-target conditions to the control - No significant main effect of stereotype condition on SRT for corrective saccades - Bayes factors indicated weak evidence for the null hypothesis when comparing the self-as-target and group-as-target conditions to the control

Experiment 2:

Stereotype threat manipulation check: Anti-saccade task: - marginally significant main effect of stereotype condition on the first manipulation check, with Bayesian analyses providing support for the alternative hypothesis - Participants in the negative group-as-target stereotype condition appeared to endorse gender differences in visuospatial performance to a greater extent than the control condition - Participants in the positive stereotype condition also seemingly endorsed gender differences in visuospatial performance to a greater extent than the control condition - Significant main effect of stereotype condition on the second manipulation check - Participants in the negative group-as-target stereotype condition perceived that men would outperform women on the anti-saccade task relative to the control - There was a substantial evidence for the null hypothesis when comparing judgements in the positive stereotype to the control condition.

MA task: - significant main effect of stereotype condition on the third manipulation check - Participants in the negative group-as-target stereotype condition were more likely to endorse gender differences in mathematical performance relative to the control condition - Strong evidence for the null hypothesis when comparing the positive stereotype to the control condition - Significant main effect of stereotype condition on the fourth manipulation check - Participants in the negative group-as-target stereotype condition were more likely to report that men would outperform women on the MA task relative to the control condition - There was moderate evidence for the null hypothesis when comparing the positive stereotype to the control condition

Anti-saccade task - significant main effect of anti-saccade accuracy, with participants responding more accurately on pro-saccade relative to anti-saccade trails - Significant main effect of SRT, with participants responding faster to pro-saccade relative to anti-saccade trails

Anti-saccade trails:

Correct saccades: - No significant main effect of stereotype condition on correct saccades - Bayes factors indicated weak evidence for the null hypothesis when comparing the negative group-as-target, and moderate evidence for the null hypothesis when comparing the positive stereotype condition to the control condition - No significant main effect of SRT for correct saccades - Bayes factors indicated moderate evidence for the null when comparing data between the negative stereotype, and weak evidence for the null when comparing the positive stereotype to the control condition

Reflexive saccades: - no significant main effect of stereotype condition on incorrect saccades - Bayes factors indicated weak support for the null when comparing data between the negative stereotype condition, and moderate support for the null when comparing the positive stereotype to the control condition - No significant main effect of stereotype condition on reflexive saccade SRT - Bayes factors indicated strong evidence for the null hypothesis when comparing the negative stereotype, and weak support for the alternative hypothesis when comparing the positive stereotype to the control.

Corrective saccades: - No significant main effect of stereotype condition on the percentage of corrective saccades - Bayes factors indicated moderate support for the null when comparing the negative and positive stereotype conditions to the control condition - No significant main effect of stereotype condition on corrective saccade SRT - Bayes factors indicated weak evidence for the null hypothesis when comparing the negative and positive stereotype

condition to the control condition
Mathematical performance: - significant main effect of problem difficulty on accuracy scores, with participants solving fewer difficult compared to simple problems - Significant main effect of problem orientation, with participants solving fewer horizontally relative to vertically oriented problems - significant two-way interaction between problem difficulty and orientation, with participants solving fewer difficult horizontal compared to vertical problems - Weak evidence in favour of the null hypothesis when comparing simple problems as a function of horizontal and vertical orientation - no significant main effect of stereotype condition on mathematical performance - evidence for non-significant effect sizes consistent with those reported by Pennington and Heim (2016) when comparing the negative stereotype to the control condition - non-significant effect when comparing the positive stereotype to the control condition - No significant interactions between experimental condition, problem demand, and orientation
Bayesian meta-analysis: - Raw effects and single study Bayes factors are shown in Table 7 - individually, the level of evidence in support for the null hypothesis in both Study 1 and Study 2 varies from weak to strong - Meta Bayes factors, calculated by combining the two datasets and using this data to test the expected effect sized specified using the results reported by Jamieson and Harkins (2007; Experiment 3), revealed that the overall body of evidence indicated substantial support for the null relative to the experimental hypothesis.

Was the precision of the estimate of the intervention or treatment effect reported?

- CONSIDER:
  - Were confidence intervals (CIs) reported?

☒ Yes
☐ No
☐ Can't tell

Are there any obvious shortcomings in the reporting of the data?

☐ Yes (please specify)
☒ No

Do the authors report on all variables they aimed to study as specified in their aims/research questions?

This excludes variables just used to describe the sample.

☒ Yes (please specify)
☐ No

Do the authors state where the full original data are stored?

☒ Yes (please specify)
☐ No

https://osf.io/mdwyv/

What do the author(s) conclude about the findings of the study?

Please give details and refer to page numbers in the report of the study where necessary.

☐ Details

Experiment 1:
Findings indicate that priming a negative self- or group-relevant stereotype did not hamper participants' correct, corrective or reflexive saccadic accuracy or associated SRT compared to the control condition. Bayesian analyses corroborated these findings, offering substantial support for the null compared to the alternative hypotheses. This contrasts with the findings reported by Jamieson and Harkins (2007; Experiment 3), who found that participants under stereotype threat launched quicker correct and corrective saccades relative to the control condition; a finding they interpret as support for the mere effort motivational account of stereotype threat. They also report that participants launched reflexive saccades (incorrect eye movements towards a peripherally placed cue) on a greater proportion of anti-saccade trails. As such, findings from Experiment 1 offer little support for the mere effort account of stereotype threat when using the anti-saccade task with an overlap paradigm. In addition, our findings do not lend support ot a working memory inference account, which suggests that participants will launch slower correct and corrective saccades because of diminished working memory capacity.
Resultantly, participants may not have perceived this particular task to be a valid indicator of their mathematical ability, which may explain why both the self-as-target and group-as-target primes did not influence anti-saccade performance. Furthermore, the simplicity of this task may have obscured stereotype threat effects by not evoking sufficient working memory demand. In order to corroborate the findings of Experiment 1, we therefore conducted a second experiment using the same anti-saccade task, but also included a measure of mathematical performance.

General Discussion: Despite having these two-tailed predictions, the current studies were unable to provide support for either the mere effort or working memory explanations of stereotype threat. Specifically, priming a negative self- or group-relevant stereotype (Experiment 1) did not appear to influence reflexive saccades launched incorrectly towards a peripherally placed target, nor the time it took to generate correct and corrective saccades. Moreover, the saliency of a negative or positive group stereotype (Experiment 2) did not influence significantly women's inhibitory control or mathematical performance. These findings garnered from Null Hypothesis Significance Testing (NHST) were augmented by a Bayesian meta-analysis, which proffered substantial evidence in favour of the null over the alternative hypothesis specified using the results of Jamieson and Harkins (2007; Experiment 3).
Overall, the current findings run contrary to a wealth of studies demonstrating that priming negative gender related stereotypes impairs women's mathematical performance (Beilock et al., 2007; Rydell et al., 2014; Spencer et al., 1999). They also contrast with prior studies indicating that women underperform on mathematical tests (Beilock & Carr, 2005; Cheryan & Bodenhausen, 2000; Rosenthal & Crisp, 2006; Tagler, 2012), but perform better on spatial tasks (e.g. Moè, 2009; Wraga et al., 2007) when they are primed with a positive

stereotype. It is worth noting, however, that recent research suggests that the stereotype threat literature may be subject to publication bias; a phenomenon whereby significant findings are published and disseminated at a substantially greater rate than non-significant findings (Flore & Wicherts, 2015). Whilst this could have stemmed from the desirable implication that stereotype threat might partly explain real-world achievement outcomes (see seminal papers by Spencer et al., 1999; Steele & Aronson, 1995), the sheer amount of positive findings published in the literature is problematic because it disproportionately inflates effect size estimates and biases meta-analyses. The results reported here suggest that the null hypothesis is a substantially better predictor of the data than the alternative hypothesis specified by previous findings (Jamieson & Harkins, 2007; Experiment 3), with none of the 95% credible intervals of the replicated effects excluding values around zero. As such, the magnitude of the effects that negative gender-related stereotypes exert on women's inhibitory control performance (and other task performance) may be smaller than that reported in original studies and may be inflated by small sample sizes and publication bias (see Flore & Wicherts, 2015).

Quality of the study - Reporting

Is the context of the study adequately described?

Consider your answer to questions: Why was this study done at this point in time, in those contexts and with those people or institutions? (Section B question 2) Was the study informed by or linked to an existing body of empirical and/or theoretical research? (Section B question 3) Which of the following groups were consulted in working out the aims to be addressed in the study? (Section B question 4) Do the authors report how the study was funded? (Section B question 5) When was the study carried out? (Section B question 6)

☒ Yes (please specify)
☐ No (please specify)

Are the aims of the study clearly reported?

Consider your answer to questions: What are the broad aims of the study? (Section B question 1) What are the study research questions and/or hypotheses? (Section C question 10)

☒ Yes (please specify)
☐ No (please specify)

Is there an adequate description of the sample used in the study and how the sample was identified and recruited?

Consider your answer to all questions in Methods on 'Sampling Strategy', 'Recruitment and Consent', and 'Actual Sample'.

☒ Yes (please specify)
☐ No (please specify)

Is there an adequate description of the methods used in the study to collect data?

Consider your answer to the following questions in Section I: Which methods were used to collect the data? Details of data collection methods or tools Who collected the data? Do the authors describe the setting where the data were collected? Are there other important features of the data collection procedures?

☒ Yes (please specify)
☐ No (please specify)

Is there an adequate description of the methods of data analysis?

Consider your answer to the following questions in Section J: Which methods were used to analyse the data? What statistical methods, if any, were used in the analysis? Who carried out the data analysis?

☒ Yes (please specify)
☐ No (please specify)

Is the study replicable from this report?

☒ Yes (please specify)
☐ No (please specify)

Do the authors avoid selective reporting bias?

(e.g. do they report on all variables they aimed to study as specified in their aims/research questions?)

☒ Yes (please specify)
☐ No (please specify)

Quality of the study - Methods and data

Are there ethical concerns about the way the study was done?

Consider consent, funding, privacy, etc.

☒ Yes, some concerns (please specify)

☐ No concerns

- no mention of ethical approval or consent

Were students and/or parents appropriately involved in the design or conduct of the study?

☐ Yes, a lot (please specify)
☒ Yes, a little (please specify)
☐ No (please specify)

Is there sufficient justification for why the study was done the way it was?

☒ Yes (please specify)
☐ No (please specify)

Was the choice of research design appropriate for addressing the research question(s) posed?

☒ Yes (please specify)
☐ No (please specify)

To what extent are the research design and methods employed able to rule out any other sources of error/bias which would lead to alternative explanations for the findings of the study?

e.g. (1) In an evaluation, was the process by which participants were allocated to or otherwise received the factor being evaluated concealed and not predictable in advance? If not, were sufficient substitute procedures employed with adequate rigour to rule out any alternative explanations of the findings which arise as a result? e.g. (2) Was the attrition rate low and if applicable similar between different groups?

☐ A lot (please specify)
☒ A little (please specify)
☐ Not at all (please specify)

How generalisable are the study results?

☐ Details

Weight of evidence - A: Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)?

In some studies it is difficult to distinguish between the findings of the study and the conclusions. In those cases please code the trustworthiness of this combined results/conclusion. ***Please remember to complete the weight of evidence questions B-D which are in your review specific data extraction guidelines.***

☐ High trustworthiness (please specify)
☒ Medium trustworthiness (please specify)
☐ Low trustworthiness (please specify)

Have sufficient attempts been made to justify the conclusions drawn from the findings so that the conclusions are trustworthy?

☐ Not applicable (results and conclusions inseparable)
☒ High trustworthiness
☐ Medium trustworthiness
☐ Low trustworthiness

Wells et al. (2014)

CASE CONTROL STUDIES

Note: A study can be awarded a maximum of one star for each numbered item within the Selection and Exposure categories. A maximum of two stars can be given for Comparability.

Selection

Is the case definition adequate?

- a) yes, with independent validation
- b) yes, e.g., record linkage or based on self reports
- c) no description

Representativeness of the cases

- a) consecutive or obviously representative series of cases *
- b) potential for selection biases or not stated

Selection of Controls

- a) community controls *
- b) hospital controls
- c) no description

Definition of Controls

- a) no history of disease (endpoint) *
- b) no description of source

Comparability

Comparability of cases and controls on the basis of the design or analysis

- a) study controls for _____ (Select the most important factor.) *
- b) study controls for any additional factor * (This criterion could be modified to indicate specific control for a second important factor.)

Exposure

Ascertainment of exposure

- a) secure record (e.g., surgical records) *
- b) structured interview where blind to case/control status *
- c) interview not blinded to case/control status
- d) written self report or medical record only
- e) no description

Same method of ascertainment for cases and controls

- a) yes *
- b) no

Non-Response rate

- a) same rate for both groups *
- b) non respondents described
- c) rate different and no designation

COHORT STUDIES

Note: A study can be awarded a maximum of one star for each numbered item within the Selection and Outcome categories. A maximum of two stars can be given for Comparability.

Selection

Representativeness of the exposed cohort

- a) truly representative of the average _____ (describe) in the community *
- b) somewhat representative of the average _____ in the community *
- c) selected group of users, e.g., nurses, volunteers
- d) no description of the derivation of the cohort

Selection of the non exposed cohort

- a) drawn from the same community as the exposed cohort *
- b) drawn from a different source
- c) no description of the derivation of the non exposed cohort

Ascertainment of exposure

- a) secure record (e.g., surgical records) *
- b) structured interview *
- c) written self report
- d) no description

Demonstration that outcome of interest was not present at start of study

- a) yes *
- b) no

Comparability

Comparability of cohorts on the basis of the design or analysis

- a) study controls for _____ (select the most important factor) *
- b) study controls for any additional factor * (This criterion could be modified to indicate specific control for a second important factor.)

Outcome

Assessment of outcome

- a) independent blind assessment *
- b) record linkage *
- c) self report
- d) no description

Was follow-up long enough for outcomes to occur

- a) yes (select an adequate follow up period for outcome of interest) *
- b) no

Adequacy of follow up of cohorts

- a) complete follow up - all subjects accounted for *
- b) subjects lost to follow up unlikely to introduce bias - small number lost - > _____ % (select an adequate %) follow up, or description provided of those lost) *
- c) follow up rate < _____% (select an adequate %) and no description of those lost
- d) no statement

University of Glasgow (n.d.)

DOES THIS REVIEW ADDRESS A CLEAR QUESTION?

Did the review address a clearly focussed issue?

- Was there enough information on:
  - The population studied
  - The intervention given
  - The outcomes considered

☐ Yes
☐ Can't tell
☐ No

Did the authors look for the appropriate sort of papers?

- The 'best sort of studies' would:
  - Address the review's question
  - Have an appropriate study design

☐ Yes
☐ Can't tell
☐ No

ARE THE RESULTS OF THIS REVIEW VALID?

Do you think the important, relevant studies were included?

- Look for:
  - Which bibliographic databases were used
  - Follow up from reference lists
  - Personal contact with experts
  - Search for unpublished as well as published studies
  - Search for non-English language studies

☐ Yes
☐ Can't tell
☐ No

Did the review's authors do enough to assess the quality of the included studies?

- The authors need to consider the rigour of the studies they have identified. Lack of rigour may affect the studies results.

☐ Yes
☐ Can't tell
☐ No

If the results of the review have been combined, was it reasonable to do so?

- Consider whether:
  - The results were similar from study to study
  - The results of all the included studies are clearly displayed
  - The results of the different studies are similar
  - The reasons for any variations are discussed

☐ Yes
☐ Can't tell
☐ No

WHAT ARE THE RESULTS?

What is the overall result of the review?

- Consider:
  - If you are clear about the review's 'bottom line' results
  - What these are (numerically if appropriate)
  - How were the results expressed (NNT, odds ratio, etc)

How precise are the results?

- Are the results presented with confidence intervals?

☐ Yes
☐ Can't tell
☐ No

WILL THE RESULTS HELP LOCALLY?

Can the results be applied to the local population?

- Consider whether:
  - The patients covered by the review could be sufficiently different from your population to cause concern
  - Your local setting is likely to differ much from that of the review

☐ Yes
☐ Can't tell
☐ No

Were all important outcomes considered?

☐ Yes
☐ Can't tell
☐ No

Are the benefits worth the harms and costs?

- Even if this is not addressed by the review, what do you think?

☐ Yes
☐ Can't tell
☐ No

References

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

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.

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

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*.