Beilock et al. (2007)

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
- Effects of stereotype threat on working memory
- theories addressing working memory's organization
- The current work draws upon reserach demonstrating that the orientation of a presented math problem can alter the working memory resources it relies on to explore how stereotype threat impacts working memory.

Do authors report how the study was funded?

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

This research was supported by Institute of Education Sciences Grant R305H050004 to Sian L. Beilock and National Science Foundation Grant BCS-0601148 to Sian L. Beilock and Allen R. McConnell.

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

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

- effects of stereotype threat on working memory
- Knowledge regarding the locus of stereotype threat within the working memory system is then used to (a) design training regimens to alleviate unwanted performance decrements and (b) predict when such effects will persist - even when the task being performed is no longer related to the stereotype threat in question.

Experiment 1: We examined whether women, who received the information that they were participating in research investigating why men are generally better at math than women, would perform worse on a math problem-solving task than would women who did not receive this information.
Experiment 2: We set the stage to examine the types of problem most susceptible to stereotype threat. Individuals judged the validity of horizontally oriented and vertically oriented math problems under both a single-task and a phonological load condition. Our goal was to identify problems that depend heavily on verbal resources in order to test whether stereotype threat is most strongly revealed for such problems.
Experiment 3: Women performed either horizontally presented or vertically presented math problems in both a baseline and a subsequent stereotype threat condition and reported their thoughts and worries under stereotype threat. If stereotype threat taxes verbal working memory resources, then those problems that rely most heavily on this capacity should be most likely to fail.
Experiment 3b: Women performed horizontal or vertical math problems in a no stereotype threat control condition and reported the thoughts they had while performing the math.
Experiment 4: Explored ways to mitigate stereotype threat.
Experiment 5: Explored a novel implication of the hypothesis that stereotype threat harms math task performance via the consumption of working memory resources - and especially verbal resources.

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

- Women under math 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

- **Experiment 2**: Not specified which sex

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)

United States

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

Modular arithmetic: The object of modular arithmetic (MA) is to judge the validity of problems such as $51 = 19(mod4)$. To do this, the middle number is subtracted from the first number (i.e. $51 - 19$) and then this difference is divided by the last number (i.e. $32/4$). If the divided is a whole number, the problem is "true". MA is an advantageous math task because its working memory demands can be easily manipulated.
Working memory demand was determined by whether the first step of the MA problem involved numbers greater than 10 and a borrow operation (e.g. $45 = 27(mod4)$). Larger numbers and borrow operations involve longer sequences of steps and require maintenance of more intermediate products, placing greater demand on working memory.
Across all experiments, half of the MA equations presented to participants were "true", and the rest were "false". Additionally, each "true" problem had a "false" correlated that only differed as a function of the number involved in the mod statement.

Participants: Participants were undergraduate students. To ensure that all participants demonstrated reasonable performance on the MA task prior to the introduction of any experimental manipulations, only individuals whose problem-solving accuracy was greater than 75% in the practice and baseline blocks were retained as participants.

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)

See above

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)

- Stereotype threat impact on working memory

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**: Stereotype threat vs no threat
- **Experiment 2**: 1 group
- **Experiment 3**: horizontal vs. vertical MA condition
- **Experiment 3b**: horizontal vs vertical MA condition

- **Experiment 4**: 1 group
- **Experiment 5**: 2 groups, spatial two-back task vs verbal two-back task

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)

- **Experiment 1**: 2 groups

- **Experiment 2**: 1 group
- **Experiment 3**: 2 groups
- **Experiment 3b**: 2 groups
- **Experiment 4**: 1 group
- **Experiment 5**: 2 groups

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)

- **Experiment 1**: Random

- **Experiment 2**: Not applicable
- **Experiment 3**: Random
- **Experiment 3b**: random
- **Experiment 4**: not applicable
- **Experiment 5**: random

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)

- **Experiment 1**: yes

- **Experiment 2**: not applicable just one group
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: not applicable just one group
- **Experiment 5**: yes

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

- **Experiment 1**: yes

- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes

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.

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)

- for women under math 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)

- To ensure all participants demonstrated reasonable performance on the mA task prior to the introduction of any experimental manipulations, only individuals whose problem-solving accuracy was greater than 75 % in the practice and baseline block were retained as participants.
- **Experiment 1**: To be retained as participants, individuals had to have reported at least moderate levels of math skill and importance of these skills (an average rating greater than 5, the midpoint, of two 9-point math related questions "I am good at math" and "It is important to me that I am good at math".)
- **Experiment 2**: Each women met the criteria outlined in the experiment overview (the criteria that is applied to all experiments).
- **Experiment 3**: same criteria as Experiment 1
- **Experiment 3b**: same as Experiment 3
- **Experiment 4**: same as Experiment 1
- **Experiment 5**: Women qualified for participation by demonstrating adequate performance on the two-back tasks (i.e., at least 70% accuracy). 3 participants qualified but were not retained as study participants because they failed to spend an adequate amount of time reading the stereotype threat manipulation (i.e. <30 s).

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)

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)

- Female undergraduate students

Were any incentives provided to recruit people into the study?

☐ Not applicable (please specify)

☐ Explicitly stated (please specify)

☐ Not stated/unclear (please specify)

- **Experiment 1**: Not stated

- **Experiment 2**: Not stated
- **Experiment 3**: not stated
- **Experiment 3b**: not stated
- **Experiment 4**: not stated

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)

- **Experiment 1**: Following the posttest, participants were debriefed.

- **Experiment 3**: following the STAI, participants were debriefed
- **Experiment 3b**: following the STAI, participants were debriefed
- **Experiment 4**: following the completion of the stereotype threat block, participants were completely debriefed

- **Experiment 5**: After performing the critical 100 trails of the same version of the two-back task they had practised prior to the MA problems, they were thanked and debriefed.

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**: Thirty-one women. Seventeen individuals were randomly assigned to the control group and 14 participants to the ST groups.

- **Experiment 2**: Twenty-four individuals participated in this experiment.
- **Experiment 3**: Thirty-three women, eighteen individuals were randomly assigned to the vertical MA condition and 15 participants to the horizontal MA condition.
- **Experiment 3b**: Fourty-two women qualified for study participation using the aforementioned criteria and were evenly split between horizontal and vertical problem groups.
- **Experiment 4**: Thiry women
- **Experiment 5**: Thirty-three women

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)

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**: Undergraduate students, no further mention of age.

- **Experiment 2**: No information
- **Experiment 3**: Undergraduate students, no further mention of age.
- **Experiment 3b**: Undergraduate students, no further mention of age.
- **Experiment 4**: Undergraduate students
- **Experiment 5**: Undergraduate students

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**: not stated

- **Experiment 2**: not stated
- **Experiment 3**: not stated
- **Experiment 3b**: not stated
- **Experiment 4**: not stated
- **Experiment 5**: 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)

- **Experiment 1**: not stated

- **Experiment 2**: not stated
- **Experiment 3**: not stated
- **Experiment 3b**: not stated
- **Experiment 4**: not stated
- **Experiment 5**: not stated

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**: The control and ST groups did not differ in terms of their perception of their math skill.

- **Experiment 2**: Not stated

- **Experiment 3**: Individuals in the vertical and horizontal conditions did not differ in terms of their perception of their math skill, or the importance they assigned to this skill.
- **Experiment 3b**: Individuals in the horizontal and vertical conditions did not differ in terms of their perception of their math skill, or the importance they assigned to this skill.
- **Experiment 4**: Not stated
- **Experiment 5**: not stated

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)

- **Experiment 1**: high, for women under math stereotype threat.

- **Experiment 2**: unclear
- **Experiment 3**: high, for women under math stereotype threat
- **Experiment 3b**: high, for women under math stereotype threat
- **Experiment 4**: unclear
- **Experiment 5**: high, for women under math stereotype threat

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: - two 9-point math-related questions “I am good at math” and “It is important to me that I am good at math.” -> a - MA task -> b - response times -> b - baseline/posttest -> b
- MA task difficulty -> working memory demand -> b

Experiment 2: - MA task -> b - phonological secondary task -> b - dual-task block -> b
- RTs

Experiment 3: - two 9-point math-related questions “I am good at math” and “It is important to me that I am good at math.” -> a - Vertical/horizontal MA task -> b - response times -> b - MA working memory demand -> b - STAI

Experiment 3b: Same as Experiment 3 but we do not have a stereotype threat manipulation

Experiment 4: - baseline -> b - stereotype block -> b - RTs -> b - problem repetition > b - problem demand -> b

Experiment 5: - MA task -> b - two-back task -> b - stereotype threat manipulation
- RTs - MA task accuracy - MA task demand - experiment

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: - see above for (a) - MA task: e.g. low demand: $7 = 2(mod5)$, high demand: $43 = 16(mod3)$, intermediate capacity demand: $19 = 12(mod7)$

Experiment 2: - see above - MA task: Vertial and horizontal MA problem - Phonological secondary task: Nonwords (e.g. gib, lec, nup) were presented

Experiment 3: - see above - MA: horizontal and vertical MA problems, low and high demand - State-Trait Anxiety Inventory (STAI), individuals responded to items e.g. “I feel at ease” on a scale ranging from 1 (not at all) to 4 (very much so). - participants responded to a question (on a 7-point scale) regarding their perceptions on the importance of performing at a high level on the last block of MA problem, ranging from 1 (not at all important to me) to 7 (extremely important to me). - Verbal Thought Questionnaire, a questionnaire intended to elicit their thoguht during the stereotype threat block of problems, it stated: “We all have several thoughts that run through our mind at any given time. Please describe everything that you remember thinking about as you performed the last set of MA problems”.

Experiment 3b: - same as Experiment 3 but we do not have a stereotype threat manipulation

Experiment 4: - horizontal MA problems - Learning blocks (3 blocks, each with 212 problems) - ST manipulation

Experiment 5: - ST manipulation (this time directly after the general instructions) - MA task - two-back task

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)

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

□ Details

Experiment 1: - RT (response time) were log transformed, to reduce the positive skew of RTs and thus the impact of outliers. - Accuracy and corresponding RTs for MA problems to which responses were correct were compared in a 2 (group: control, ST) x 2 (block: baseline, posttest) x 2 (problem working memory demand: low demand, high demand) design, with group as a between-subjects variable, ANOVA

Experiment 2: - accuracy and corresponding RTs for MA problem to which responses were correct were compared in a 2 (block: single-task baseline, dual-task) x 2 (problem working memory demand: low vs. high) x 2 (problem orientation: horizontal vs. vertical) within subjects design. ANOVA - phonological secondary task: accuracy and corresponding RTs for phonological secondary task problems to which responses were correct were analyzed as a function of the working memory demands and orientation of the MA problem they were paired with in a 2 (MA problem working memory demand: low vs. high) x 2 (MA problem orientation: horizontal vs vertical) ANOVA

Experiment 3: - ANOVA - Verbal Thought Questionnaire data was coded by two experimenters, unaware of the hypotheses or experimental conditions. - mean calculation - MA: accuracy and corresponding (log transformed) RT measures for problem to which responses were correct were compared in a 2 (block: baseline, stereotype threat) x 2 (problem working memory demand: low demand, high demand) x 2 (problem orientation: vertical, horizontal) ANOVAs, which problem orientation as the between-subjects variable. - three-way ANOVA on RTs Block x Problem working memory demand x problem orientation

Experiment 3b: - same as Experiment 3 but we do not have a stereotype threat manipulation

Experiment 4: - ANOVAs

Experiment 5: - mean calculation - ANOVAs - regression analyses - multiple regression analyses

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

Experiment 3: - Verbal Thought Questionnaire data was coded individually by two experimenters, unaware of the hypotheses or experimental conditions. They had a 97.8% interjudge agreement

Experiment 3b: - same as Experiment 3 but we do not have a stereotype threat manipulation

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

Experiment 1: - figures (in text) - table (in text) - in text

Experiment 2: - table (in text) - in text

Experiment 3: - table (in text) - in text

Experiment 3b: - table (in text) - in text

Experiment 4: - table (in text) - in text - figure (in text)

Experiment 5: - in text - figures (in text)

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: - MA problems and their RTs that were not performed at least 65% correct across all participants in the baseline condition were removed from both the baseline and experimental blocks in all experiments to ensure that individual MA problems were not unduly difficult to solve - In terms of accuracy, a significant Group x Block x Problem Demand interaction obtained. - A 2 (block) x 2 (problem demand) ANOVA for the control group revealed only a main effect of problem difficulty - Accuracy was higher for the low-demand than the high-demand problems - Same ANOVA for the ST group revealed a Block x Difficulty interaction. - There was no difference between the ST group's low-demand problem performance from the baseline to the posttest. - High-demand accuracy was significantly lower in the posttest as compared with the baseline condition. - In terms of RTs, a 2 (group) x 2 (block) x 2 (problem working memory demand) ANOVA revealed main effects of block in which individuals performed the problems faster over time, and problem demand, in which high-demand problems RTs were slower than were low-demand RTs - All other main effects and interactions, including Group x Block x Problem Demand interaction were not significant.

Experiment 2: - No MA problems were performed below 65% correct across all participants in the baseline condition. - In terms of problem-solving accuracy, this analysis revealed a main effect of problem demand, and a Block x Working Memory Demand interaction. - Low-demand problems did not differ in accuracy from the single-task to the dual-task block - High-demand problems were performed less accurately in the dual-task in comparison to the single-task block - Analysis of RTs revealed a main effect of problem difficulty, in which the low-demand problems were performed faster than high-demand problems, and a Block x Difficulty interaction, in which high-demand problem RTs increased from the single-task baseline to the dual-task block, whereas low-demand RTs decreased. - No other main effect or interactions reached significance. - *phonological secondary task*: There were main effects of MA problem working memory demand, and MA problem orientation, which were qualified by a significant working memory demand by problem orientation interaction. - There was no difference in phonological accuracy when performing low-demand or high-demand vertical MA problems. - When performing horizontal MA problems, phonological task accuracy was significantly higher for low-demand in comparison to high-demand problems. - Phonological task accuracy while performing high-demand horizontal MA problems was also significantly worse than when performing low-demand and high-demand vertical MA problems. - Phonological secondary task accuracy was lowest when individuals performed high-demand horizontal MA problems, suggesting that horizontal high-demand problems and the phonological task are competing for the same verbal working memory resources. - In terms of phonological secondary task RTs a 2 (MA problem working memory demand) x 2 (MA problem orientation) ANOVA produced a main effect of difficulty, in which individuals were slower to respond to the phonological secondary task when it was performed with a high-demand in comparison to a low-demand MA problem and a marginal Difficulty x Direction interaction. - Although not significant, this interaction parallels the phonological accuracy data in that the slowest phonological task RTs were seen in association with the performance of the horizontal high-demand MA problems.

Experiment 3: - Participants performing horizontal and vertical MA problems did not differ in their perception of the importance of performing well under stereotype threat - Participants in the vertical orientation and the horizontal orientation condition reported that it was at least "moderately important" to perform well on these problems. - Participants in the vertical and horizontal orientation condition did not differ in reports of state anxiety. Thus, any differences in MA perforamnce under stereotype threat as a function of problem orientation cannot be accounted for by differences in anxiety or perceived importance between the two problem orientation conditions. - Verbal Thought Questionnaire interjudge agreement was extremely high (97.8%) - On average, participants reported about three thoughts in total. - With respect to specific thought categories, 14.5 % reflected worries or thoughts about confirming the stereotype threat manipulation, 34.9% were thoughts regarding monitoring their performance and its consequences, 32.4% were related to the steps involved in performing the math problems, and 18.3% were thoughts unrelated to the current experiment. - Participants in the horizontal and vertical MA conditions did not differ in the total number of thoughts reported or in the proportion of verbal reports across the categories. - Worries about the situation and its consequences accounted for about half of the participants' reported thoughts under stereotype threat. - Questionnaire data, when combined with MA performance, provide converging evidence that stereotype-induced consumption of working memory (especially phonological components of this system) is responsible for less-than-optimal performance in mathematical problem solving. - MA: Analysis of accuracy revealed the anticipated three-way interaction - THe impact of stereotype threat was quite different depending on the working memory demand and the orientation of the problems being performed. - For vertical problems there was no Block x Problem demand interaction. - There was a significant Block x Problem demand interaction for the horizontal problems. - The horizontal low-demand problems did not significantly differ in accuracy from the baseline to stereotype threat block, the horizontal high-demand problems were performed significantly less accurately in the stereotype threat block than in the baseline block. - This pattern of data supports the prediction that stereotype threat targets the working memory resources on which horizontal high-demand problems rely for successful execution. - A three-way ANOVA on RTs also revealed a Block x Problem Working Memory Demand x Problem Orientation interaction. - A 2 (block) x 2 (problem demand) ANOVA on vertical problem RTs revealed only a main effect of problem demand in which high-demand RTs were slower than low-demand problem RTs - A ANOVA on horizontal problem RTs revealed a significant Block x Problem Demand interaction. - While horizontal low-demand problem RTs decreased from the baseline to stereotype threat block, horizontal high-demand problem RTs increased, albeit not significantly.

Experiment 3b: - Regardless of problem orientation, individuals did not differ in terms of their perceptions of the importance of performing well on the last block of problems, both reporting that it was at least “moderately important” to perform well on these problems. - Horizontal and vertical problem participants did not differ in their reports of state anxiety. - In terms of state anxiety, there was no main effect of experiment, or orientation, and no Experiment x Orientation interaction. - In terms of importance, there were no main effects of experiment orientation, and no Experiment x Orientation interaction. *Verbal Thought Questionnaire:* - Interjudge agreement was extremely high (98%) - On average participants reported about four thoughts in total, which was a somewhat larger total than reported in Experiment 3 - A significantly larger proportion of individuals’ reported thoughts in Experiment 3b as compared with Experiment 3 were unrelated to the task at hand - The proportion of participants’ worries and thoughts related to the performance situation and its consequences was significantly greater under stereotype threat than no threat conditions. - Statistically controlling for such worries eliminated differences in math task performance under threat, supporting a causal role of verbal thoughts and worries in stereotype threat-induced failure. - 4.2% of these reports reflected worries about the task, 29.7% were worries regarding monitoring their performance and its consequences, 30.5% were related to the steps involved in performing the math problems, and 35.5% were thoughts unrelated to the current experiment - Participants in the horizontal and vertical conditions did not significantly differ in the total number of thoughts reported or in the proportion of verbal reports across the categories. - Worries about the task accounted for only 4% of participants’ reported thoughts and together with monitoring performance and its consequences, these thoughts accounted for roughly one third of what was reported - In terms of percentage of reported worries, this analysis (2 [experiment: 3 vs 3b] x 2 [problem orientation group]) revealed a main effect of experiment - Individuals under stereotype threat reported a significantly greater proportion of their thoughts being devoted to worrying than those under the no

threat condition. - There was neither a main effect of problem orientation nor an Orientation x Experiment interaction. - A similar pattern of results was seen for the proportion of thoughts regarding monitoring performance and its consequences., although the main effect of experiment as well as the main effect of problem orientation, and their interaction, was not significant - Experiment x Problem Orientation ANOVA on the percentage of reported worries together with monitoring performance and its consequences also produced a main effect of experiment - The main effect of problem orientation and the Problem Orientation x Experiment interaction was not significant - Individuals performing MA problems under the no stereotype threat control condition of Experiment 3b devoted a significantly lower portion of their thoughts to worrying about the situation and monitoring performance and its consequences than those performing the same problems under stereotype threat
MA: - A 2 (block) x 2 (problem demand) x 2 (problem orientation) ANOVA on accuracy revealed only a main effect of problem demand in which the low-demand problems were performed more accurately than the high-demand problems. - A similar ANOVA on RTs produced main effect of problem demand, which were qualified by a Problem Demand x Block interaction - The low-demand problems were performed faster than the high-demand problems, this difference was greater in the posttest than in the baseline block - If one compares RTs on the types of problems shown to be impacted by stereotype threat across Experiment 3 and 3b in a 2 (block) x 2 (problem orientation) x 2 (experiment) ANOVA, a significant three-way interaction obtains - For the vertical high-demand problem RTs, a 2 (block) x 2 (experiment) ANOVA revealed no main effect of block, or experiment and no Block x Experiment interaction. - A analysis of horizontal high-demand problem RTs, revealed a significant Experiment x Block interaction - Horizontal high-demand RTs decreased from the baseline to the posttest in Experiment 3b, these same RTs increased from the baseline to the stereotype threat block in Experiment 3. The simple effects did not reach significance - A 2 (block) x 2 (problem orientation) x 2 (experiment) ANOVA on high-demand problem accuracy also revealed a significant three-way interaction - Vertical high-demand problems, a 2 (block) x 2 (experiment) ANOVA revealed no main effects - The same ANOVA for horizontal high-demand problems revealed a significant Block x Experiment interaction - Horizontal high-demand problems accuracy significantly decreased in Experiment 3 from the baseline block to the stereotype threat block, accuracy for the same problems in Experiment 3b did not. - We performed the same 2 (block) x 2 (experiment) ANOVA on horizontal high-demand problem accuracy presented above and added as a covariate the proportion of reported worries together with monitoring performance and its consequences. - To the extent that worries and thoughts about performance consequences underlie stereotype threat, covarying out these thoughts should render the significant Block x Experiment interaction reported above non significant. This is exactly what was found.

Experiment 4: - Accuracy and RTs for correct problems were analysed in separate ANOVAs with a 2 (block: baseline, stereotype threat) x 2 (problem repetition: no repeat problems, multiple repeat problems) x 2 (problem working memory demand) design. - A significant Block x Problem Repetition x Problem Working Memory demand interaction obtained - This three-way interaction was examined by analysing the heavily practiced (multiple-repeat) problems and novel (no repeat) problems separately - A 2 (block) x 2 (problem demand) ANOVA on the multiple repeat problems revealed no Block x Problem

Demand interaction - The same ANOVA on the no repeat problems revealed a significant Block x Problem Demand interaction - Accuracy for the no repeat low-demand problems did not differ between the baseline and stereotype threat block - Accuracy for the no repeat high-demand problems significantly declined from the baseline to the stereotype threat block - Analysis of RT data did not alter teh conclusions supported by the accuracy analysis - 2 (block) x 2 (problem repetition) x 2 (problem working memory demand) ANOVA on RTs revealed main effects of problem repetition, and problem demand, which were qualified by the significant Repetition x Demand interaction - No Block x Repetition x Problem demand interaction was found - Problem demand level had more of an effect on RTs for no repeat problems than for the multiple repeat problems - Repeated high-demand problems did yield longer RTs than did repeated low-demand problems, suggesting that there was at least some degree of nonautomatic answer retrieval - As accuracy for the high-demand no repeat problems declined from the baseline to the stereotype block, RTs increased, although not significantly

Experiment 5: - We compared horizontal high-demand accuracy in Experiment 5 with accuracy in the same type of problems in the other experiments in which stereotype threat was manipulated at low levels of practice. This analysis revealed no difference in horizontal high-demand problem accuracy as a function of experiment - We next compared performance on these horizontal high-demand problems under stereotype threat with performance on the same type of problems under no threat conditions. A significant main effect of stereotype threat was found. - Across experiments, performance on the horizontal high-demand problems under stereotype threat was significantly lower than performance on the same problems under no threat conditions
Two-back task: - Accuracy and RT measures for correct trails were analysed, revealing faster and more accurate spatial than verbal two-back task performance - The RT difference was significant. The accuracy difference was not - We examined two-back Rt and accuracy in a 2 (task: verbal/spatial) x 2 (experiment: control pilot vs stereotype threat) design. Task x Stereotype threat interaction obtained for RT. - The verbal two-back task was performed significantly slower than the spatial two-back task following stereotype threat in MA - This did not occur when the two-back task was not preceded by stereotype threat performance. There were no significant effects for the accuracy analyses. - We examined the relation between MA and two-back task performance as a function of the type of two-back task individuals performed. In this context, better performance can be revealed by greater accuracy, faster RTs, or both. Thus, we conducted three sets of multiple regressions in which two-back task performance (where spillover was exhibited) was regressed on MA performance, the type of two-back task (dummy coded), and their interaction (the key prediction). The three regression analyses examined performance on MA and the two-back task using standardized accuracy, standardized RTs, and a composite of the two (standardized accuracy minus standardized RTs). - With RT as an index of performance, there was a main effect of two-back task type, a marginal main effect of MA RT, and the predicted interaction between the two - The relation between RTs for MA and the two-back task was significant for those completing the verbal two-back tas, but not for those completing the spatial two-back task. - When using accuracy as an index of performance, there was a main effect of MA performance, and a marginal interaction of MA performance and two-back

task type - Although not reliable at conventional levels, this latter outcome reflects that the relation between MA accuracy and two-back task was, as expected, significant for the verbal two-back task but not for the spatial two-back task. - We conducted a multiple regression analysis by using composite measures that captured both accuracy and RTs. In many way, the composite reflects the best index of performance because it simultaneously takes into account both accuracy and latency. - In this analysis, there was a main effect of two-back task, which was qualified by its predicted interaction with MA performance. - As found above for both accuracy and RT separately, the relation (now with the composite approach) between performance for MA and the two-back task was significant for the verbal two-back task but not for the spatial two-back task. - Regardless of whether performance was defined as accuracy, latency, or a composite of the two, those who performed worse on the MA task under stereotype threat performed more poorly on the subsequent two-back task - however, this relation only held up for verbal two-back task performance.

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

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

☒ Yes

☐ No

☐ Can't tell

- **Experiment 1**: yes

- **Experiment 2**: yes
- **Experiment 3**: no
- **Experiment 3b**: no
- **Experiment 4**: no
- **Experiment 5**: no

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

☐ Yes (please specify)

☐ No

- **Experiment 1**: no

- **Experiment 2**: no
- **Experiment 3**: no
- **Experiment 3b**: no
- **Experiment 4**: no

- **Experiment 5**: 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

- **Experiment 1**: yes

- **Experiment 2**: yes
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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

☐ Yes (please specify)
☒ No

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: Participants assigned to a control or ST group performed horizontally presented MA problems that varied as a function of the demands they placed on working memory. Only MA problems heavily dependent on working memory (i.e. horizontal high demand problems) failed under stereotype threat, suggesting that stereotype threat exerts its impact by co-opting working memory resources needed for the successful execution of such problems.

Experiment 2: Adding a phonological memory load to MA execution led to performance decrements (primarily reflected in a decrease in secondary task accuracy) only when the MA problems being performed were high in working memory demand and presented in a horizontal orientation. Because participants were instructed to perform both the MA and the phonological secondary tasks equally well, errors in either task are evidence of disruption in working memory.
This finding suggests that high-demand horizontal (more so than vertical) MA problems and the phonological secondary task were competing for the same pool of verbal resources. More important, these findings establish the condition to test whether stereotype-threat-induced failure is strongest for problems that rely most heavily on verbal working memory resources.

Experiment 3: There were no differences as a function of block for vertical problem performance - regardless of problem working memory demand. This was not the case for the horizontal problems. The horizontal low-demand problems were not impacted by the introduction of a negative performance stereotype, the horizontal high-demand problems were performed significantly worse under stereotype threat in comparison to baseline conditions. Individuals were also asked to report their thoughts during the stereotype threat block. Approximately half of these reported thoughts related to worries about the stereotype threat situation and to monitoring performance and its consequences. Unfortunately, off-line measures such as these cannot capture the intensity, duration, or precise timing of participants' thoughts. What the verbal reports do reveal is that participants did indeed report worries and performance concerns while under stereotype threat and, furthermore, that the prevalence of these thoughts did not differ as a function of problem orientation. This suggests that although all individuals experienced worries and verbal thoughts related to their performance under stereotype threat, these thoughts were only problematic for those individuals performing horizontally presented problems - problems that rely heavily on verbal working memory resources. Nonetheless, one might note that we have not demonstrated that individuals worry more under stereotype threat than in a no threat situation.

Experiment 3b: Under the no stereotype threat conditions of Experiment 3B, women performed at a high level on the MA tasks, regardless of problem orientation or demand. Moreover, in comparison to women performing the same problems under stereotype threat in Experiment 3A, a significantly lower proportion of individuals' reported thoughts in Experiment 3B were related to worries and thoughts about the situation and its consequences. Finally, the critical interaction of experiment and problem block for the type of performance shown to be most strongly impacted by stereotype threat across the first several studies in the current work (i.e., horizontal high-demand MA accuracy) was rendered nonsignificant when worries and thoughts about performance and its consequences was taken into account.

Experiment 4: Performance of horizontally presented MA problems practised 48 times each (multiple repeats), and thus not heavily reliant on working memory, did not fail under stereotype threat. Problems presented only once (no repeats) did. Furthermore, these failures were limited to the no repeat problems that placed the heaviest demand on verbal working memory.
The current findings reaffirm the adage that "practice makes perfect", and further, they suggest and addendum to this statement. Practice not only makes perfect, but practice (provided that it creates less reliance on working memory) makes skills robust to stereotype threat effects. Practice designed to alleviate the working memory demand of the sub-components of the problems one encounters should be an efficacious training strategy.

Experiment 5: To our knowledge, Experiment 5 is the first demonstration that following underperformance on a stereotype-relevant task, subsequent task performance in a different domain is also negatively impacted - as long as the subsequent task depends heavily on the same type of working memory resources that stereotype threat also consumes. This stereotype threat spillover occurred despite the subsequent task being unrelated to the stereotype in question. That is, a math-related stereotype should not apply to verbal task performance. If anything, women might anticipate doing better in a verbal domain. In summary, performance decrements were observed in a task performed subsequent to the stereotyped task, demonstrating how stereotype threat can spill over onto other activities not implicated by the stereotype in question.

General Discussion: The current work examined how negative performance stereotypes impact cognitive resources necessary to successfully execute working memory intensive tasks.

Results revealed that stereotype threat exerts its impact by co-opting working memory resources - especially *phonological aspects* of this system - needed for the successful performance of some types of math problems more than others.

We demonstrated that stereotype threat may not only impact performance in the domain implicated by the stereotype, but it can spill over onto subsequent, unrelated tasks that depend on the same processing resource that stereotype threat consumes.

These novel findings not only provide insights into the cognitive underpinnings of stereotype threat but also reveal new circumstances when its effects are attenuated and propagated. Such knowledge contributes to our theoretical understanding of stereotype threat and speaks to how environmental factors can influence the working memory system. This is an issue that has not yet received adequate attention in the working memory literature but is of import for researchers interested in developing models of working memory that capture the complexity of real-world performance.

Our work also provides evidence that stereotype threat induced task-related thoughts and worries that target *phonological aspects* of working memory. Using Baddeley's multicomponent model as a framework, one could unpack verbal memory into a *phonological* store capable of holding speech-based information and an articulatory control process based on inner speech mechanism.

Regardless of the specific subcomponentson which such tasks rely, the current work's demonstration of a heavy involvement of *verbal resources* in stereotype threat impairment **does not exclude other subcomponents of the working memory system** from being implicated in stereotype threat related failure. Rather, stereotype threat likely **affects a combination of phonological loop functioning** (via verbal thoughts and worries) and probably **some central executive functioning** (via attempts to suppress such thoughts and to focus on the task at hand). This leaves open the possibility that tasks with spatial components, but that also tax central executive or phonological resources, may show sings of stereotype threat as well - although we would argue that such failures should not be as pronounced as in tasks that depend more so on phonological aspects of the working memory system.

Is it possible that the **current results** could be *accounted for* **solely** *by stereotype threat's impact on* **general executive control resources**? There is a number of reasons why this notion seems **highly unlikely**. First, previous research has shown that although horizontally presented problems are impacted most heavily by a phonological load, vertical problems are impacted more heavily by a spatial load, suggesting that horizontal problems' stronger reliance on phonological (rather than executive) resources that makes them susceptible to stereotype threat effects. Second, the vertical and horizontal problems presented in the current work were exactly the same - only orientation differed. And indeed, there was no difference in horizontal and vertical problem performance under single-task baseline condition. Thus, it seems unlikely that one problem type would place heavier demands on central executive resources than another. Moreover, both types of high-demand problems involved carry operation that have been shown to implicate central executive resources. Thus, to extend that stereotype threat or the phonological task used in Experiment 2 solely taxed executive resources, then both types of problems should have failed. Finally, not only did the verbal (but not the spatial) two-back task in Experiment 5 show signs of stereotype threat induced spillover, the verbal two-back task was the only one that correlated with MA performance under stereotype threat. If general resource consumption could solely explain stereotype threat effects and their spillover, then a correlation between MA performance under stereotype threat and spatial two-back performance should exist, but there was not. In summary, an explanation for the current work's stereotype threat effects based exclusively on the taxing of general executive control resources does not seem tenable.

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)

- **Experiment 1**: yes
- **Experiment 2**: yes
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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)

- **Experiment 1**: yes
- **Experiment 2**: no

- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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)

- **Experiment 1**: yes
- **Experiment 2**: no
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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)

- **Experiment 1**: yes
- **Experiment 2**: yes
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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)

- **Experiment 1**: yes
- **Experiment 2**: yes
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes

Is the study replicable from this report?

☐ Yes (please specify)

☐ No (please specify)

- **Experiment 1**: yes
- **Experiment 2**: yes
- **Experiment 3**: yes
- **Experiment 3b**: yes
- **Experiment 4**: yes
- **Experiment 5**: yes

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)

- **Experiment 1**: unclear
- **Experiment 2**: unclear
- **Experiment 3**: unclear
- **Experiment 3b**: unclear
- **Experiment 4**: unclear
- **Experiment 5**: unclear

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

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

- only toward females under math stereotype threat

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)

- **Experiment 1**: medium trustworthiness
- **Experiment 2**: unclear, questions not clearly stated
- **Experiment 3**: medium trustworthiness
- **Experiment 3b**: medium trustworthiness
- **Experiment 4**: high trustworthiness
- **Experiment 5**: medium trustworthiness

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

- **Experiment 1**: not applicable

- **Experiment 2**: not applicable
- **Experiment 3**: not applicable
- **Experiment 3b**: not applicable
- **Experiment 4**: not applicable
- **Experiment 5**: not applicable

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

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

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

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.

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*.