Schmader and Johns (2003)

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 impact on performance
- Working memory capacity

Do authors report how the study was funded?

☒ Explicitly stated (please specify)

□ Implicit (please specify)

□ Not stated/unclear (please specify)

This research was supported in part by National Science Foundation Grant BCS-0112427 and by a faculty small grant awarded to Toni Schmader from the University of Arizona Foundation.

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

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

- Primary goal of this research was to investigate the impact of ST manipulations on WM capacity
- WM capacity has not been examined as an explanation for gender differences in maths performance, let alone differences that are produced by experimentally manipulated conditions of ST>

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

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)

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: - Tested the hypothesis that women (but not men) would show reduced WM capacity when a testing situation was framed as measuring maths ability.

We hypothesized that if stereotype threat interferes with working memory, then women who complete a working memory test described as a measure of mathematical ability—a stereotyperelevant ability domain—would show lower working memory scores compared with men and compared with women in a control condition.

Experiment 2: - Was a conceptual replication of Experiment 1, comparing the WM capacity of Caucasian and Latino participants when the task was said to be related to general intelligence - Having to contend with negative group stereotypes places an additional cognitive burden on people that can interfere with their ability to perform up to their potential on complex cognitive tasks

We expected that Latino students would show evidence of reduced WM capacity compared with Whites when the test was described as a measure related to intelligence. We did not expect any performance differences when the test was described as a measure of working memory.

Experiment 3 - Using a different manipulation of ST and measure of WM that does not involve maths to directly test the hypothesis that a reduction in WM mediates the effects of ST on women's performance on a standardized maths exam.

We expected that if WM capacity mediates ST effects, then scores on the WM test should account for performance decrements on a standardized maths test.
We expected that women would exhibit a reduction in WM capacity after ST was primed but before beginning the stereotype-relevant task and that this reduction would mediate the effect of ST on performance.

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:
- WM capacity - maths performance - ST manipulation - anxiety, perceived difficulty - gender identity

Experiment 2:
- same as Experiment 1 but gender identity questionnaire was swapped for ethnicity.

Experiment 3:
- Experiment 1

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:

- 2 (male or female) x 2 (ST or control)

Experiment 2:

- 2 (Latino or White) x 2 (ST vs control)

Experiment 3:

- ST vs control

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 and 2:

- four

Experiment 3:

- two

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 & 2:

1. 2 female research assistants ran the experimental sessions in same-sex gender groups

ranging from 2 to 4 participants 2. participants seated in individual rooms 3. prerecorded description of the study (read along and listen) -> served as the ST manipulation 4. general overview of testing procedure 5. practice set of three equation/word trails 6. test 7. test experience questionnaire 8. debriefing

Experiment 3:
1. in large room: overview of the experimental session + consent form 2. ST manipulation, then moved to different room (individual room) 3. WM task 4. WM questionnaire 5. back to large room 6. ST manipulation, then maths test 7. posttest questionnaire 8. probed for suspicion 9. debriefing

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)

Experiment 1, 3:
- women under ST

Experiment 2:
- ethnic minorities under ST

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)

Experiment 1, 2:
- psychology students

Experiment 3:
- female undergraduates

-### 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)

Experiment 1:
- mass survey asking about SAT scores on quantitative section and ST knowledge

Experiment 2:
- basis of their self-reported ethnicity in a mass survey

Experiment 3:
- see Experiment 1

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)

- survey

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 & 3:
- course credit or $10

Experiment 2:
- course credit for a research requirement

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)

Experiment 1 and 2: - not stated

Experiment 3: - consent form

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:
- 40 male and 35 female undergraduate psychology students - 9 men and 7 women excluded, leaving a final sample of 31 men and 28 women.

Experiment 2:
- 33 Latino (20 women, 13 men) and 40 White (27 women, 13 men) psychology students.

Experiment 3:
- 31 female undergraduates - 3 participants in the control condition were excluded, leaving a final sample of 28.

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)

Experiment 1:
- 31 out of 40 male and 28 out of 35 female undergraduate psychology students

Experiment 2:
- Data of 1 white participant was lost, leaving a final sample of 72.

Experiment 3:
- 28 out of 31

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)

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 2:
- 33 Latino and 39 White (in the final sample)

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:
- we selected participants who indicated in a previous mass survey session that they had scores 500 or higher on quantitative section of the SAT (or equivalent converted American College Test score). - In the same mass survey, we assessed stereotype knowledge; only those who responded at or above the scale midpoint of 4 were recruited. - participants were randomly assigned to one of two test description conditions in a 2 (male or female) x 2 (ST or control) factorial design

Experiment 2:
- participants were only selected on the basis of their self-reported ethnicity in an earlier mass survey - Participants were randomly assigned to one of the two conditions in a 2 (Latino or White) x 2 (ST or control) factorial design

Experiment 3:
- The participants were selected on basis of the same criteria as in Experiment 1 and completed the study in exchange for course credit or $10. - Data from 3 participants in the control condition were lost, 1 because of a computer malfunction and 2 because of procedural errors.
- Participants were randomly assigned to complete two tasks (the WM test and a standardized maths test) under ST or control conditions.

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, 2:
- demographic information -> a - WM test -> b - Test experience questionnaire -> b

Experiment 3:
- same as experiment 2 - additionally, maths test -> 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:

WM test: - adapted dual-processing test called the operation-span task that has been developed and used extensively by Engle et al., to assess WM
ST manipulation: participants were told that gender differences in maths performance might stem from underlying gender differences in quantitative capacity. Test was described as a reliable measure of quantitative capacity.
test experience questionnaire: contained anxiety scale and items related to perceptions of the test and testing situation. Anxiety was measured with questions adapted from the Spielberger State Anxiety Scale

Experiment 2:

- same as Experiment 1, apart from ST manipulation and test experience questionnaire (identity was changed from gender to ethnicity)
*ST manipulation**: - WM test was described as a reliable measure of WM capacity in both the control and ST condition. In ST condition, researcher also said that research had shown that performance on the WM test is “highly predictive” of the performance on intelligence tests and that their performance on the test would be used to “help establish norms for different groups”. Participants in both conditions were informed that they would receive feedback about their performance

Experiment 3:

- created version of the WM test that did not involve mathematical equations. Instead, the processing component of the test required participants to count the number of vowels contained within a given sentence - Standardized maths test we used consisted of 30 multiple-choice word problems taken from the quantitative section of practice GREs. - ST manipulation: adapted procedures used by Inzlicht and Ben-Zeev (2000) to test the effects of solo genders status on women's maths performance - post-test questionnaire: see experiment 1 and 2 + asked to rate how they thought the researcher expected men and women to do relative to each other on the maths test using a 7-point-scale.

- WM questionnaire: rate the extent to which they thought performance on the test would be related to maths ability and memory ability, on a 7-point scale + rate on an 11-point scale how important they thought each part of the task (counting vowels and remembering words) was for determining their overall performance + indicated how they allocated their attention while completing the task using an 11-point scale + asked to rate how well they expected to perform on the upcoming maths test/problem-solving exercise using a 9-point scale

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:
- Gender x ST ANOVA on participants' quantitative SAT scores - controlled for correlation between operation-span task and SAT scores by conducting 2 (gender) x 2 (ST) ANCOVA
WM capacity:
Absolute span score: - derived by summing the total number of words from only those sets of words where all the words in the set were recalled correctly. ANCOVA on absolute span score - simple main effects tests
Equation evaluation: - Although working memory is assessed as a function of performance on the word recall, we also analyzed performance on the equations to assess whether there were any significant differences due to the stereotype threat manipulation.
Test experience questionnaire: - ANOVA - simple effects analyses

Experiment 2:
- Because of the quasi-experimental design of this study, we conducted initial 2 (ethnicity) x 2 (gender) x 2 (ST) ANOVAs on participants' quantitative and verbal SAT scores to discern whether there were any group differences in these variables that should be controlled in our

analyses
WM capacity: - see experiment 1, gender swapped for ethnicity
Test experience questionnaire: - see experiment 1, gender swapped for ethnicity

Experiment 3:
- see experiment 1 and 2
Perceptions of the WM task: - perceptions of what performance on the WM task would be related to (maths ability and memory ability) were analysed with 2 (ST) x 2 (ability type) mixed-factors ANOVA
Maths test performance: - maths test was analysed as women's accuracy on the test, that is, the number of problems answered correctly divided by the number of problems attempted
Mediational analyses: - To test whether the reductions of working memory mediated the effect of stereotype threat on test performance, we computed a series of regression equations as prescribed by Baron and Kenny (1986). According to this approach, three relationships between the target variables must be demonstrated to establish a basis for testing mediation. The independent variable must predict both the dependent and the mediator variable and the mediator must predict the dependent variable. Once these conditions are established, the dependent variable is regressed onto the independent variable and mediator in a final regression analysis. Support for mediation is obtained by demonstrating that the effect of the independent variable (stereotype threat) on the dependent variable (math test performance) is significantly reduced when accounting for the effect of the hypothesized mediator (working memory capacity).
- to rule out other explanations for relationships (e.g., model misspecification), we also conducted a reverse mediation analysis with maths test performance serving as the mediator and WM serving as the dependent variable.

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
- figure

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:
- Gender x ST ANOVA on participants' quantitative SAT scores revealed that men's SAT scores were higher than women's
WM capacity:
Absolute span score: - ANCOVA significant main effect of gender, and ST, and the predicted two-way interaction, was a significant covariate - Simple main effects tests revealed that women in the ST condition recalled fewer words than men in the ST condition and than women in the control condition - Span score for men in the control condition was not significantly difference from the score of men in the ST condition and women in the control condition
Equation evaluation: - There were no significant effects of gender or ST on percentage of equations solved correctly - Analysis of the average amount of time (in seconds) spent on

each equation revealed a marginal main effect of ST - SAT was a significant covariate, but no other effects were significant - Regardless of gender, participants in the ST condition tended to spend more time evaluating each equation than did participants in the control condition
Test experience questionnaire:
Anxiety: - Analysis of the anxiety measure did not yield any significant main effects or interactions
Perceived difficulty: - analysis of the difficulty ratings revealed a significant two-way interaction - SAT was a significant covariate, neither main effect was significant - Simple effects analyses showed that within the ST condition, women rated the test as more difficult than men did - Difficulty ratings of women and men in the control condition were not significantly different.
Gender identity threat: - Analysis yielded a main effect for the ST manipulation - Both women and men in the ST condition expressed greater concern that the researcher would evaluate their performance in terms of their gender identity compared with participants in the control condition - No other effects were significant - Mean pattern reveals that the manipulation did produce some conscious awareness that the researcher might use their gender as a factor to evaluate their performance, but this concern was not unique to women.

Experiment 2:
- 2 x 2 x 2 ANOVA: no significant differences in quantitative SAT but there was a marginal ethnic difference in verbal SAT scores - No other significant effects
WM capacity:
Absolute span score: - Analysis of the absolute span score produced the predicted Ethnicity x ST interaction - Verbal SAT was a significant covariate - Simple effects testing revealed that Latinos in the ST condition recalled significantly fewer words than did Whites in the ST condition and Latinos in the control condition - Recall by Whites in the control condition was equivalent to that of Latinos in the control condition and to the recall of Whites in the ST condition - Analysis yielded main effect of gender (men recalled more words than women) but no other effects were significant.
Equation evaluation: - Accuracy of responses to the equations did not vary as a function of ethnicity, ST, or the interaction of the two - Significant gender difference in equation accuracy (men were more accurate than women), that was qualified by an unexpected Gender x ST interaction - No other effect were significant - Verbal SAT was a marginally significant covariate - Simple effects tests of this interaction showed that although men and women were equally accurate in their evaluation of the problems in the control condition, women were less accurate than men in their evaluations of the problems when the test was described as predictive of intelligence - No significant differences in the average time participants spent on each equation
Test experience questionnaire:
Anxiety: - Yielded marginal Gender x Ethnicity interaction and a significant Ethnicity x ST interaction - Simple effects analysis indicated that Latinos in the ST condition reported significantly more anxiety compared with Latinos in the control condition, whereas the ST manipulation had no effect on self-reported anxiety of Whites.
Test difficulty: - Yielded a gender main effect (women perceived the test to be more difficult

than did men), and a marginal Ethnicity x ST interaction - No other effects were significant - Simple effects tests revealed that Whites and Latinos saw the test as equally difficult in the control condition, whereas Latinos rated the test to be more difficult than did Whites in under ST

Ethnic identity threat: - Yielded only a marginal Ethnicity x ST interaction - Latinos reported slightly more ethnic identity threat in the ST condition than in the control condition, whereas Whites reported slightly less ethnic identity threat in the ST condition than in the control condition, although none of the simple effects were significant - No other effects were significant

Experiment 3:

WM capacity:

Absolute span score: - women in the ST condition recalled fewer words on the WM test than did women in the control condition

Vowel counting: - Accuracy of number of vowels counted correctly did not vary as a function ST, and there were no significant differences in the average time participants spent counting the vowels - Amount of time participants spent counting vowels was not substantially different than the average amount of time participants spent evaluating the difficult equations in Experiment 1 and 2.

Perceptions of the WM task: - ANOVA revealed only a main effect of ability type - Participants in both conditions indicated that they perceived performance on the WM test to be more related to memory ability than maths ability

- Additional analyses revealed that participants in the control condition and the ST condition did not differ significantly in their perceptions of which part of the WM task was more important (vowel counting or word recall); and participants in both conditions generally viewed word recall to be the more important portion of the test - No significant differences between conditions in what part of the test participants reported focusing on, and participants focused more on remembering the words than counting the vowels - Results reduce the plausibility of the alternative explanation that lowered word recall in the ST condition results from participants simply shifting their focus to the processing task and away from memorizing and recalling words.

Maths test performance: - Women in the ST condition were less accurate on the maths test than women in the control condition - No significant differences in the number of maths problems attempted by women in the ST condition and women in the control condition, suggesting that women in both condition expended comparable effort on the test

Test experience questionnaire:

Anxiety, perceived difficulty, and gender identity threat: - no significant differences in self-reported anxiety between the two condition - average anxiety rating for both conditions was the midpoint of the scale - whereas ST had a significant effect on women's difficulty ratings and perceptions of gender identity threat in Experiment 1, the ST manipulation in Experiment 3 did not affect either their difficulty ratings for the maths test or their ratings of gender identity threat - inconsistency in the conscious experience of ST across the two studies with women probably reflects the fact that our manipulation of threat was more explicit in the first experiment in which we explicitly told women that there were gender differences on the test.

Performance expectancies on the maths test/problem-solving exercise: - No significant differences in the expectancies of participants in the ST condition and the control condition. - Women in the ST condition were no more likely to believe that the researcher expected gender differences than were women in the control condition - Women in both conditions tended to believe that the researcher expected men to outperform women).
Mediational analyses: - ST had a significant negative effect on women's WM capacity (the mediator) and maths test performance (the dependent variable) - Third regression analysis established that WM capacity was a significant predictor of accuracy on the maths test - when performance on the maths test was regressed onto both ST and WM capacity, ST was no longer a significant predictor of maths test performance, whereas WM capacity remained significant in the equation - Sobel test of the reduction in the direct ST effect was significant providing support for our hypothesis that ST interferes with women's maths test performance by reducing their WM capacity. - Reverse mediation analysis: In contrast to the primary mediation analysis, St remained a marginally significant predictor of WM when controlling for maths test performance. - Sobel test confirmed that reduction in ST effect when controlling for performance on the maths test was not significant - These analyses provide greater support for the hypothesis that WM mediates effects on maths test performance than for an alternative model in which maths test performance mediates ST effects on WM capacity.

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

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:
- provide initial support for the hypothesis that a ST manipulation can lead to a measurable decrease in cognitive resources. - As predicted, women completing a working memory test described as a test related to mathematical ability showed reduced cognitive capacity, as measured by the number of words they were able to recall within the task. Men, and women in a nonthreat control condition, did not exhibit this working memory decrement and were able to recall an equal number of words.

Experiment 2:
- The results of the second study provide further evidence for the negative effects of stereotype threat on working memory capacity. - When the working memory test was described as a measure related to intelligence, Latinos recalled fewer words compared with Whites and compared with Latinos in the nonthreat control group. These results demonstrate that the working memory task we chose is not uniquely relevant to stereotypes about math performance but can be used to assess the capacity deficits experienced by other stereotyped groups during a cognitively taxing test. This study also provided a stronger test of our hypothesis because, although participants were told that their performance on the test would be used to establish norms for different groups, the issue of group differences in the stereotyped performance domain was never mentioned explicitly.

General Discussion:
The experiments reported here were designed to test the hypothesis that stereotype threat reduces an individual's performance on a complex cognitive test because it reduces the individual's working memory capacity. Results of three experiments provide support for this hypothesis. Experiments 1 and 2 demonstrated that manipulations of stereotype threat led to lower working memory scores among individuals who are targeted by the stereotype (women and Latinos) but had no effect on those who are not targeted by the stereotype (men and Whites). Results of the third experiment reveal that the reductions in working memory capacity observed under stereotype threat mediate the reductions in performance on a standardized test. Taken together, these findings suggest that members of stigmatized groups perform poorly on cognitive tests when negative stereotypes have been primed because this added information interferes with their attentional resources.
Although the pattern of results on the primary variable of interest (i.e., working memory capacity) was quite consistent across three studies using different groups, different manipulations, and different measures, the patterns of data on participants' self-reported experiences were more variable.
We want to state explicitly that our emphasis on the cognitive deficits associated with stereotype threat is not meant to imply that negative affect does not contribute to the effect of stereotype threat on performance.

The primary purpose of the research reported here is to advance our knowledge of the ways in which negative stereotypes exert their influence on the individuals they target. Following the work of Steele et al., we approached this issue from the perspective that negative social stereotypes can create an added psychological burden in situations where one's behavior might be interpreted as evidence for the validity of such belief systems. Across three studies, we provide converging evidence that performing under the specter of a negative stereotype can deplete the cognitive resources of stigmatized group members and impair performance on challenging academic tasks.

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)

- can't tell

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

- Experiment 1 and 2 not mentioned if consent was obtained

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

- pretty generalisable, as results were consistent acorss different groups and ST manipulations/tests

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.

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

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*.