Mangels et al. (2012)

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
- its effects on learning
- underlying cognitive and emotional mechanisms

Do authors report how the study was funded?

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

This work was funded by the Institute for Educational Sciences (IES) Cognition and Student Learning (CASL) Grant (to J.A.M, C.G., and C.S.D.)

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

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

The present study focused on the emotional and cognitive processes that differentiate those females who rebound from math errors under stereotype threat from those who succumb.

Our primary questions concerned whether ST influences females' emotional responses to negative feedback, and whether these emotional responses predicted: (i) use of the tutor (i.e. engagement with learning) and (ii) the ability to correct the errors on a retest (i.e. success in learning).

Greater vigilance and initial sensitivity to negative feedback may undermine learning by motivating individuals to regulate their arousal by distracting themselves from the tutorial learning opportunity. In contrast, attempts to regulate ones emotional response by focusing on the response itself, through either rumination or suppression, will likely divert cognitive resources from processing task-relevant information.

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

- females who rebound from math errors under stereotype threat from those who succumb.

What is the relevant age group?

☐ Not applicable (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)

United States, Columbia University

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)

We focused our analysis on three ERP waveforms: the medial frontal feedback-related negativity (FRN), the anterior P3 (P3a) and the posteriorally-maximal late positive potential (LPP). Drawing upon previous research, we relate these waveforms, respectively, to the relatively automatic appraisal of feedback valence, the subsequent orienting to the feedback as a function of its motivational salience, and the sustained attentional processes that are sensitive to the subjective level of emotional arousal the feedback elicits.
We expect that, together, the FRN and P3a will primarily measure the effects of stereotype threat on individuals' initial sensitivity to feedback valence, whereas the LPP should track the dynamics of arousal up-regulation, as might be produced by rumination or suppression.

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)

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)

- whether they ended up in the stereotype threat condition or control condition

Number of groups

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

☐ Not applicable (not more than one group)

☐ One

☒ Two

☐ Three

☒ Four or more (please specify)

☐ Other/unclear (please specify)

maybe even 4 -> stereotype threat vs no stereotype threat, and within these 2 conditions there was a separation between better and poorer learners but this might be a pre-existing difference, i'm not sure to be honest.

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)

The text does not mention how the participants were assigned to the stereotype threat or control condition.

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)

- participants did not know whether or not they ended up in the stereotype threat or no threat condition, they didn't even know these conditions existed.

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.

Three days. First day: pre-measures about gender, math identification, perception of environmental stereotype threat. Second day: more questionnaires assessing mood and confidence in math ability, afterwards preparation for EEG, first round of math tasks, then received either stereotype threat or no threat framing. Thrid day (approximately 24hours later): surprise retest of math problems that wre isomorphic to those from the first test.

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)

- females under stereotype threat struggling with math

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

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

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

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

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

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

Planned sample size

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

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

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)

Were any incentives provided to recruit people into the study?

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

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)

Asians were not included given the potential for conflict between gender- and racebased stereotypes.

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)

68 female undergraduate students (32 stereotype threat, 36 non-threat)

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)

began with 71, ended up with 68 but 4 subjects were excluded in total -> does not add up

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)

paper describes the subjects as English-speaking undergraduate females at Columbia University.

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)

Mean age was between 19.9 and 20.7, with the SD being 0.43 and 0.46, however lowest and highest age are not mentioned.

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)

Out of the initial 71 participants 61 were Caucasian and 10 were non-Caucasian, none however, were Asian. It is not stated whether or not the three excluded participants are Caucasian or non-Caucasian.

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)

- Participants were English-speaking, right-handed, neurologically healthy, consenting undergraduate females at Columbia University.
  Three subjects with EEG recording difficulties were excluded.

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

None.

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

- demographics questionnaire -> a
- PEST -> b
- Math Identification -> b
- mood and confidence in math ability questionnaire -> b
- EEG -> b
- math tasks -> b
- retest of math problems -> 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)

- 48 multiple-choice GRE-like math questions
- 48 isomorphic 'yokes'
- tutor (computer)
- EEG
- PEST (Perceived Environmental Stereotype Threat)
- Stereotype threat manipulation through questionnaire and vocal instructions
- math SAT

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)

- analysis of tutor use took a hierarchical approach

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

☐ Details

- univariate and multivariate approaches
- all univariate analyses were conducted with 2 (threat) x 2 (learning: better learners vs poorer learners) mixed-measures ANOVAs/ANCOVAs

- t-tests
- structural equation modeling (SEM) of the paths to retest success
- AMOS was used to conduct maximum likelihood parameter estimations on resulting variance-covaraince matrices.

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
- tables
- figures

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

Gender and math-related individual differences

Gender Identification did not differ as a function of either ST or subsequent learning success. Perceptions of others' stereotypes about females' abilities relative to males' in amth (PEST) were also similar across groups, as were students' Math SAT and confidence in math ability. Math identification demonstrated a significant interaction between ST and learning. Under NT, better learners reported marginally higher math identification than poorer learners. In addition, although under ST, better and poorer learners did not differ from each other, the better learners under ST reported significantly lower math identification than better learners in the NT condition.

First-test and retest performance:

ST suppressed females' overall performance on the initial test. There was no effect of ST either on overall retest performance or our measure of tutor-based error correction (i.e. proportion corrected at retest with greater than guessing level of confidence, following some use of the tutor).

These results affirm that some females were able to take advantage of tutorial opportunities and rebound from math failures even under threat. Within the ST group, some individuals achieved more successful outcomes than others, despite starting from nearly identical levels of first-test impairment.

ERP responses to performance feedback:
We focused our analyses on the difference waves created by subtracting positive from negative feedback. This subtraction highlights any differential salience associated with negative feedback by treating the positive feedback response as a baseline against which the negative feedback response is compared.
These difference waves clearly illustrate the negative-going FRN and positive-going P3a. Yet, neither ST nor learning success were associated with significant modulations of these waveforms. Despite the appearance of an enhances $P3a_{diff}$ in poorer learners under ST, exploratory pair-wise comparisons between better and poorer learners as a function of threat condition confirmed that there were no significant differences as a function of learning success.
Learning success was directly related to the more sustained attention and emotional arousal processes as indexed by the LPP. An enhanced LPP_{diff} was found in poorer learners only, supporting the general hypothesis that sustained emotional processing of negative feedback would interfere with learning success. Although no significant effects of ST nor interactions between threat and learning were found, exploratory pair-wise comparisons indicated that ST was largely driving this effect, even though a similar trend could be found under NT. This finding suggests that stereotypes may indeed selectively increase the emotional burden of negative feedback on learning.

Engagement with tutor following errors:
First, we asked whether the frequency with which students made some attempt to investigate the tutor differed as a function of threat or learning success. There were no significant group differences in the basic measure of tutor entry. However, when we probed the depth of tutor engagement for tutors that were entered, we found that deeper exploration within the tutor led to greater overall success in solving isomorphic retest problems. This effect was moderated by the precence of threat, in that the quantity of exploration was related to error correction only under NT. Under ST, better and poorer learners appeared to make similar efforts to investigate the tutor, at a level that was marginally shallower than the better learners under NT, but did not differ from the poorer learners under NT. This suggests that poorer learners in the ST condition were retaining less information from their investigation of the tutor than were the better learners in this condition.

Structural equation modelling of retest success:
Figure 3B and C.
As can be seen by comparing the optimized models of ST and NT, the response to negative feedback played a much greater role in determining error correction success under ST than under NT conditions.
Under ST, the initial detection of negative outcome significantly predicted how deeply females attempted to explore the tutor, with those who had a more negative-going FRN_{diff} withdrawing effort from the tutor earlier. Tutor engagement then marginally predicted error correction. Taken together, these results lend some support to the disengagement hypothesis, in that they revealed a mediated relationship between FRN_{diff} and error correction that was not apparent in our univariate analyses. In contrast, the significant direct path from the LPP_{diff} to error correction under ST is more consistent with the interference hypothesis, and suggests that sustained emotional processing interfered with the quality of information

extracted from the tutor.
These results suggest that ST can undermine learning through either reduced quality or quantity of tutor engagement, and furthermore, that these effects are linked to different components of the emotional response to negative feedback. Although we acknowledge that the FRN_{diff} and $P3a_{diff}$ also demonstrated marginally significant direct paths to learning outcome under ST, it is also important to note that these paths accounted for as much less variance than the direct path between the LPP_{diff} and error correction.
Perhaps even stronger evidence that the FRN_{diff} and LPP_{diff} were related to different maladaptive learning strategies, however, is the finding that under ST, a more negative-going FRN_{diff} was associated with a smaller positive-going LPP_{diff}. Thus, ST did not simply increase all aspects of emotional processing of negative feedback uniformly Rather, it was those individuals who demonstrated the largest initial response to the negative feedback who appear to have been the most motivated to disengage further attention from either this information, as indicated by their smaller LPP_{diff}, or any other reminders of their poor performance on that problem.
In contrast to the effects of the response to negative feedback on learning under ST, the only significant predictor of error correction under NT was the quantity of exploration within the tutor. This suggests that whatever parts of the tutor students in this condition opted to explore were encoded deeply, without the interference experienced by students in the ST condition. Furthermore, when these students opted to withdraw 'earlier' from the tutor it did not appear to be negatively related to their emotional response to the error, at least as measured by the ERPs of interest.
We note that Math SAT was unrelated to learning success under NT conditions, but was the best predictor of outcome under ST. In the presence of ST, better Math SAT scores were also associated with less tutor use.

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

Our main finding was that the paths to this successful learning differed as a function of ST, with threat increasing the dependence between females' emotional response to negative performance feedback and their error correction success. Specifically, females who demonstrated evidence of enhanced initial detection of negative outcomes, perhaps stemming from greater overall vigilance for ability-impugning information, disengaged earliest from their exploration of subsequent learning opportunities, whereas those who demonstrated poorer regulation of their attention to and arousal from the negative feedback failed to receive significant learning benefits from any tutorial information they may have explored. Neutralizing threat not only liberated learning from dependence on these emotional responses, but also from prior measures of ability, leaving learning success to be determined by other factors, such as intrinsic motivation to investigate new approached to math problems, which may have been rooted in higher identification with the math domain.
These results are boradly consistent with recent fMRI findings that females' threat-induced performance deficits on math-relevant tasks are related to increased activity in neural regions associated with emotional conflice and its regulation, including the ventral anterior cingulate cortex.
Importantly, however, we demonstrate for the first time that these emotional responses can disrupt not only initial performance but also learning. These findings also lend support to general feedback theories that model negative feedback as harmful rather than helpful to learning if it directs attention away from the task and toward processing of the individual's affective experience.
In addition our findings begin to unpack the different mechanisms by which emotion might impact learning under ST.

Quality of the study - Reporting

Is the context of the study adequately described?

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

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

Are the aims of the study clearly reported?

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

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

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

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

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

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

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

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

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

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

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

Is the study replicable from this report?

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

Do the authors avoid selective reporting bias?

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

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

Quality of the study - Methods and data

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

Consider consent, funding, privacy, etc.

☐ Yes, some concerns (please specify)
☒ No concerns

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

generalisable within females under 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)

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.

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

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*.