Hirnstein et al. (2014)

IMPORTANT NOTE: PUT AS FOOTNOTE!

UNCLEAR WHETHER OR NOT THE PAPER FITS THE INCLUSION/EXCLUSION CRITERIA.

DESCRIPTION OF THE PARTICIPANTS IS TOO VAGUE!

EVERYTHING ELSE FITS THE INCLUSION/EXCLUSION CRITERIA SET BY THE PREREGISTRATION SO IT IS INCLUDED ANYWAYS

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
- sex/gender differences in cognitive tasks
- Testosterone (T) and stereotype threat

Do authors report how the study was funded?

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

- This work was supported by Grant HA3285/4-1 and HI1496/1-1 of the Deutsche Forschungsgemeinschaft (DFG).

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

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

- Testosterone (T) in particular has been shown to have oranizational effects during prenatal neural development with consequences for cognitive abilities later in life.
- The present study focused on activational effects that occur throughout life by its non-genomic, direct neuromodulatory effects on brain functions and cognitive abilities.
- The aim of the present study was to investigate whether gender stereotypes affect cognitive sex differences only in mixed-sex groups or whether they also apply to same-sex settings which amy have important implications for the ongoing debate on co-education.

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

- Males and females in an educational setting

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)

- Germany

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)

H1: We hypothesized that the action of gender stereotypes increases sex differences in all tasks (Sex by Condition interaction) by either enhancing performance of positively stereotyped participants (e.g., females' performance in verbal fluency) or by reducing performance of negatively stereotyped participants (e.g. males' performance in verbal fluency).
H2: We further hypothesized that cognitive sex differences will be largest if gender-stereotypes are activated in mixed-sex setting (Sex by Condition by Group Sex Composition interaction). By measuring participants' T levels as a consequence of the competitive testing situation related to gender stereotypes and group sex composition.
H3: Based on Hausmann et al. (2009) and Josephs et al. (2003) and the general assumption that T levels are related to cognitive performance, we hypothesized that if cognitive performance is enhanced after gender stereotype activation, there will be a rise in T levels, particularly in mixed-sex groups. In contrast, stereotype threat might be associated with a T drop.
H4: In addition, we investigated whether T levels in general were correlated with cognitive performance and whether cognitive performance was correlated with the magnitude of the corresponding gender stereotype. That is, the more strongly females are convinced that females in general excel in verbal fluency, the higher their verbal fluency performance (and the more strongly males are convinced, the lower their 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)

- Gender Stereotype Questionnaire
- Testosterone Assays
- Cognitive Tests

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)

- Mixed vs same-sex group

- Gender stereotype activated vs control
- gender
- it is Gender stereotype activated vs control -> Males vs Females -> Same-sex vs Mixed-sex

Number of groups

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

☐ Not applicable (not more than one group)

☐ One

☐ Two

☐ Three

☐ Four or more (please specify)

☐ Other/unclear (please specify)

Was the assignment of participants to interventions randomised?

☐ Not applicable (not more than one group)

☐ Not applicate (no prospective allocation)

☒ Random

☐ Quasi-random

☐ Non-random

☐ Not stated/unclear (please specify)

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

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

☐ Not applicable (not more than one group)

☐ Not applicable (no prospective allocation)

☒ Yes (please specify)

☐ No (please specify)

☐ Not stated/unclear (please specify)

- All participants were naive to the study's hypotheses.

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.*

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)

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)

- 12 participants were excluded because of neurological conditions, cognitive performance of more than two SDs below average in all five cognitive tests, or missing data.

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)

- 148 adults (78 women, 70 men) recrited at the Department of Psychology, Ruhr-University Bochum, Germany.

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)

- 12 participants were excluded
- 70 women participated
- 66 men participated
- in total 136 participants

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)

- Mean age of 24.40 years (SD=4.9) for women
- Mean age of 25.56 years (SD=4.3) for men.

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)

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)

- Gender stereotypes activated: n = 66
- Control = 70
- Gender stereotypes activated -> males: 32
- Gender stereotypes activated -> females: 34
- Control -> males: 34
- Control -> females: 36
- Gender stereotypes activated -> males -> same-sex: 13
- gender stereotypes activated -> males -> mixed-sex: 19
- gender stereotypes activated -> females -> same-sex: 17
- gender stereotypes activated -> females -> mixed-sex: 17
- control -> males -> same-sex: 16
- control -> males -> mixed-sex: 18
- control -> females -> same-sex: 20

- control -> females -> mixed-sex: 16

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)

- sample was not described

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

- ages -> a
- gender -> a
- gender stereotype questionnaire -> b
- gender neutral stereotype questionnaire -> b
- Redrawn Vandenberg and Kuse Mental Rotation Test (Version A) [MRT-3D] -> b
- MP-2d -> b (is a subtest of the WILDE-Intelligenz-Test)
- Word Fluency Test (WF) -> b
- 4-Word Sentences Test (4W) -> b
- Perceptual Speed Tst (PS) -> b
- Testosterone Assays -> b (salvia samples)

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)

- see above

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)

- ANOVA
- alpha-level pretest and posttest
- t-test
- multiple linear regressions

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

☐ Details

- see above

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

- Tables
- Figures
- in text

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

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

☐ Details

Gender stereotypes: - Females were generally associated with higher verbal and males with higher spatial skills - 2 (Sex) x 2 (Condition: Gender-stereotyped vs control) vs 2 (Group sex composition: same- vs mixed-sex) ANOVA, only a main effect of Sex, for item 8 and a main effect of Group Sex Composition, for Item 15 emerged. No further main effects or interactions were significant across all 16 items. - Examined whether gender stereotypes changed from before to after cognitive testing by subjecting probability estimates for all 16 items in the gender-stereotyped group to a 2 (sex) x 2 (group: same vs mixed-sex) vs 2 (Pre-/post cognitive testing) ANOVA, a main effect of Pre-/Post cognitive testing was found for Item 5. - Before stereotype threat manipulation, participants were significantly more convinced that a person who "can draw a map of the area where he/she lives" was more likely to be male than after testing. - No further main effects or interactions involving Pre-/Post cognitive testing were significant. - Overall, the analyses of individual questionnaire items indicated pronounced gender stereotypes, which were relatively stable over the time of testing and differed only marginally between groups and conditions.

Cognitive Test Performance: - Test socres of the MRT-3D, MP-2D, WF, 4W, and PS were subjected to 2 (Sex) x 2 (Condition) x 2 (Group sex composition) ANOVAS - The alpha-level was set at .05 post hoc t-tests were carried out with Bonferroni adjustment, and effect sized are given as the proportion of variance accounted for. - For sex differences, Cohen's d is additionally provided to facilitate comparison with previous studies. Means and SEs are shown in *Table 2*.

Mental Rotation: - For the MRT-3D, the ANOVA revealed a significant main effect of Sex with higher scores for men than women. There were no other significant main effects or interactions. - In the MP-2D, the main effect of Condition was significant, indicating that participants in the control condition achieved higher scores than those in the gender-stereotyped condition. No further main effects or interactions were significant.

Verbal Fluency: - For WF, the ANOVA revealed a main effect of SEX, indicating that women obtained higher verbal fluency scores than men. - Moreover, there was an interaction between Condition and Group Sex Composition. - Controls in mixed-sex groups achieved the highest scores, differing significantly from controls in same-sex groups. - No other post hoc t-test reached significance - The higher performance of controls in mixed-sex groups led to a significant main effect of Group Sex Composition, with participants in mixed-sex groups outperforming participants in same-sex groups. No other main effects or interactions were significant.

- For 4W, there was a Sex x Condition Interaction.
- As predicted in H1, there was no significant difference between men and women in the control condition, but verbal fluency was significantly lower in men than women in the gender stereotype condition.
- Gender-stereotyped men also had significantly lower scores than non-stereotyped men and non-stereotyped women.
- As in WF, a significant interaction between Condition and Group Sex Composition emerged, with controls in mixed-sex groups achieving the highest 4W score. They outperformed controls in the same-sex groups and gender-stereotyped participants in mixed-sex groups.

- The main effect of Condition was significant with lower scores in the gender stereotype condition than in the control condition, and a trend towards higher scores in women was observed. No other main effects or interactions were significant.

Perceptual Speed: - For PS, the ANOVA revealed a significant interaction between Condition and Group Sex Composition. - Controls in the mixed-sex groups obtained the highest score, they significantly outperformed controls in same-sex groups as well as gender-stereotyped participants in same-sex and mixed-sex groups. - Paired comparisons between the three same-sex groups did not reach significance. As a result of the higher performance of controls in the mixed-sex groups, there was a significant main effect of Condition, and Group Sex Composition, with higher performance in controls and mixed-sex groups compared to gender-stereotyped and same-sex participants, respectively. There were no other significant main effects or interactions.

- Despite a relatively large sample size (N = 136), the predicted 3-way interaction between Condition, Group Sex Composition, and Sex (h2) did not approach significance in any cognitive test.
- A power analysis revealed a power of .82 to detect a three way interaction.

Testosterone Levels: - To test whether T changes occurred with respect ot stereotype boost or stereotype threat (h3), T levels from saliva samples were subjected to a 2 (Sex) x 2 (Condition) x 2 (Group Sex Composition) x 2 (pre-/post stereotype manipulation) mixed-design ANOVA - A main effect of Sex revealed significantly higher T levels in men than women. No other effects were significant.

Relationship between Gender Stereotypes, Testosterone, and Cognitive Performance: - We investigated whether there was a general association between T, gender stereotypes, and cognitive performance (h4) - Multiple linear regressions were computed for each cognitive test (MRT-3D, MP-2D, 4W, WF, and PS) with the specific test score as the dependent variable and T levels (after the experiment) as well as gender stereotypes as predictors. - We focused on Items 10 and 12 from the gender stereotype questionnaire because they directly relate to mental rotation (MRT-3D and MP-2D) and verbal fluency (WF and 4W) tasks, respectively. - Item 10 was used as a predictor for MRT-3D and MP-2D, while Item 12 was used as a predictor for WF and 4W. - No gender stereotype item was used for PS - Multiple regressions were conducted separately for men and women because it has previously been suggested that the relationship between T and cognitive performance is sex-specific. - Given that T levels are highly correlated with sex, this procedure additionally avoids multicollinearity.

- In men, none of the cognitive tasks showed a significant model. Only the MRT-3D, gender stereotype was as significant predictor, indicating a positive correlation between MRT-3D score and the probability that somebody who was good at mental rotation was male (i.e., the better men performed on the MRT-3D the stronger was their gender stereotype that males excel in mental rotation).

- T levels were not significantly correlated with cognitive performance.

- In women, a significant model only emerged in 4W, accounting for 13 % of variance

- Only the gender stereotype significantly predicted the 4W score, indicating a negative correlation between 4W scores and the probability that somebody who was good at verbal fluency was male (i.e., the better women performed on the 4W the stronger was their gender stereotype that women excel in verbal fluency).
- T levels did not significantly predict any of the cognitive task scores
- Since T levels before and after the experiment were highly correlated, regressions did not include both predictors because of multicollinearity.
- The results did not change if the predictor T levels after the experiment, as reported above, was replaced by pre-test T levels.

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

The gender stereotype questionnaire revealed that participants of both sexes believed that males, rather than females, were more likely to do well on spatial tasks and that females, rather than males, were more likely to do well in verbal tasks. These findings are consistent with two previous studies which used the same gender stereotype questionnaire. The gender stereotypes remained stable across cognitive testing and were very similar across men and women, across participants in the control and gender stereotype condition, and across participants in mixed and same-sex groups. The observed differences in cognitive performance were thus unlikely to arise from differences in pre-existing gender stereotypes. Finally, in accordance with Hypothesis 7, we found that men and women who performed better on mental rotation (MRT-3D) and verbal fluency (4W), respectively, also held stronger gender stereotypes regarding spatial and verbal skills. As gender stereotypes for all participants were measured after cognitive testing, this shows that cognitive performance may strengthen gender stereotypes.

Overall, there was a female advantage in verbal fluency, which was consistent with meta-analyses and comprehensive reviews on sex differences in verbal abilities. IN the control condition, however, men generated as many four-word sentences as women whereas men's performance was significantly reduced when gender stereotypes were. This is a typical stereotype threat effect consistent with Hypothesis 1. No stereotype threat emerged in the other verbal fluency test. This discrepancy between tasks can be explained by the fact that stereotype threat emerges particularly in cognitively more demanding tasks. Since 4W is considered to be more demanding than WF, because whole sentences instead of only single words need to be generated, this might explain the emergence of stereotype threat in 4W only.

In mental rotation, the typical male advantage emerged on the MRT-3D. The sex differences in the MP-2D were nonsignificant and negligible in size and was consistent with previous findings showing that 3-dimensional objects yield stronger sex differences than 2-dimensional objects. Unexpectedly, the sex difference in MRT-3D was unaffected by the gender stereotype manipulation. Most importantly, we did not replicate the enhanced MRT-3D scores in gender-stereotyped men as reported in Hausmann et al. (2009) although the gender stereotype manipulation was identical and participants in both studies showed similar pronounced gender stereotypes with respect to spatial abilities. This cannot be attributed to a failure of the stereotype manipulation, since our stereotype intervention successfully induced stereotype threat and group sex composition effects in other tasks. It is also rather unlikely that we recruited an unusual sample, since well-known sex differences in mental rotation and verbal fluency were found. The results of the present study rather suggest that it is difficult to induce stereotype threat and boost simultaneously when the test battery includes tasks favouring men and women.

In addition, the perceptual speed test neither revealed significant sex differences nor any gender stereotype effects.

Hypothesis 2 was that stereotype boost or threat effects, such as men's reduced performance in verbal fluency, are particularly pronounced in mixed-sex groups. However, no three-way interaction emerged.

...

Taken together, the present study showed that the cognitive performance of men and women was affected by gender stereotypes and group sex composition. First, the present study was one of the very few that found a stereotype threat in men's cognitive performance (i.e., verbal fluency). Second, the present study demonstrated that an interaction of gender stereotyping and group sex composition affected the performance in sex-sensitive cognitive tasks. This probably occurs when the test environment is appraised as challenging, thereby raising the arousal level close to its optimum. However, when gender stereotypes are additionally activated, the testing situation might be evaluated as threatening and performance is likely to be reduced. This is a strong argument against proponents of single-sex schooling who argue that mixed-sex settings have generally detrimental effects on performance. Finally, the present study did not find any interaction between gender-stereotyping and T levels: Gender-stereotyping neither affected T levels nor were baseline T levels related to the susceptibility to stereotype threat. In fact, the present study did not find any evidence for a relationship between baseline T and cognitive performance.

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

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.

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

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*.