Wulandari and Hendrawan (2020)

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)

- Gender stereotype threat
- Stereotype threat situations favouring women
- self-efficiency beliefs
- clustering and switching

Do authors report how the study was funded?

☐ Explicitly stated (please specify)

☐ Implicit (please specify)

☒ Not stated/unclear (please specify)

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

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

- letter fluency performance and strategy in stereotype threat situations using different gender stereotypes

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

- men and women

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)

- Indonesia

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)

- This study aimed to address how different gender stereotype threat-activation cues and different levels of task difficulty would affect men and women on letter fluency task performances in the aspects of number of correct words, errors, clustering, and switching.

- We hypothesized that different types of gender-stereotype activation cue would affect male and female performance differently, and increasing difficulty levels will decrease their letter fluency 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)

- Task difficulty

- Stereotype threat activation

- Letter fluency performance -> switching -> clustering -> errors -> number of correct words

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)

- Stereotype-activation -> blatant vs moderately explicit vs sublte vs no activation

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)

- 4, see above

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)

- was a single-blind study

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.

Procedures: 1. pre-recorded stereotype activation cues 2. pre-recorded letter fluency test -> consisted of 3 parts: test instructions, sample item, and test items -> experimenter took notes of the words generated 3. gender-stereotype questionnaire 4. self-rating

Methods - Sampling strategy

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

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

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

- individuals under gender-stereotype threat

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

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

☐ Not applicable (please specify)

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

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

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

☐ Not applicable (no sampling frame)

☒ Explicitly stated (please specify)

☐ Implicit (please specify)

☐ Not stated/unclear (please specify)

- Participants were excluded from this study if they did not speak Bahasa as their first language for daily communication, had a history of trauma and/or surgical operation related to the brain and central nervous system, suffered from neurological and psychiatric disorders, had a history of substance and/or alcohol abuse, or were left-handed or ambidextrous.

- Given that letter flfuency is a neuropsychological test, these criteria were applied to ensure that language barrier and neurological or psychicatric conditons did not affect participants' letter fluency performance.

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)

- undergraduate students from a state university in Depok, Indonesia.

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)

- The experimental procedure was carried out in accordance with the Declaration of Helsinki.

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)

- 168 undergraduate students (91 female)

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)

- 168 students were selected, 168 participated (100%)

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)

- Indonesia

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)

- undergraduates

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)

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)

- it is a homogeneous group, as intended by the study design

- the age is not stated

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

- Gender stereotype-activation cues -> b

- Letter fluency performance -> b

- Gender-stereotype questionnaire -> b

- Self-rating -> b

- manipulation check -> 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)

- manipulation check: questionnaire asking whether the recorded instruction for activating gender stereotype threat was perceived as having a different degree of stereotype across the three manipulation groups (blatant, moderately explicit, subtle) and a control group.

- self rating: translated from Hausmann et al (2009); "To what extent does the following item describe you? I can generate many words beginning with the same letter within one minute". On a 7-point scale, the participant was asked how applicable that statement was to them, with '1' being 'not descriptive at all and '7' indicating 'highly descriptive'

- Gender-stereotype questionnaire: Translated into Bahasa from gender-stereotype questionnaire by Hausmann et al. (2009), permission was obtained. Only used one item that refers to verbal fluency.

- Letter Fluency Test: K, R, W; each letter represents easy, moderate, and hard level of difficulty (Hendrawan & Hatta, 2010; Hendrawan et al., 2015).

- Gender stereotype-activation cues: pre recorded using Audacity. Adapted from Nguyen and Ryan (2009) and Hirnstein et al, (2012)

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)

- was a single blind design

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

- the experimenter noted the words generated in the Letter Fluency Test

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

- Manipulation check: Kruskal-Wallis Test

- Kolmogorov-Smirnov Test: manipulation across gender

- effect of activation types, gender, and level of task difficulty on letter fluency performance, number of correct words and error from each letter on the letter fluency test were subjected to 4 x 3 x 2 mixed anova.

- The total correct words, total errors, mean cluster size, and number of switches from the letter fluency test were subjected to a 4 x 2 ANOVA to examine the effect of activation types and gender on letter fluency overall performance and its component processes

- One sample t-test was performed on the probability estimate from the male column of the gender-stereotype questionnaire to examine whether the probability estimation was significantly different from a value of 50 (an indication of neutral appraisal)

- One sample t-test was also performed on the self-rating score to examine whether the score was significantly different from a value of 4 (an indication neutral appraisal of their performance)

- To further examine whether gender stereotype and performance appraisal varied across activation type and gender, probability estimates from the gender-stereotype questionnaire and scores from self-rating were subjected to 4 x 2 ANOVA

- Pearson correlations were performed on the scores from the letter fluency test (number of correct words, number of errors, mean cluster size, number of switches), gender-stereotype questionnaire, and self-rating to examine whether the scores from these instruments correlated with each other.

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

- figures

- in text

- table

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

Manipulation check: - median score from the male column on the manipulation check showed a gradual increase across the four stereotype threat activation groups, with the lowest mean produced by the blatant group, followed by the moderately explicit, subtle and control group. - Median score from the male column on manipulation check showed a comparable value between the 2 gender groups. - Instruction was significantly different across activation groups, but not significantly different across gender - Pairwise comparisons with adjusted p-values showed that there were significant differences in the score from the male column between the blatant activation group compared to the rest.

Letter Fluency Test: - Analysis of correct words showed that the level of task difficulty was the only variable with a significant effect on the letter fluency performance, as the difficulty increased the number of correct words decreased. - Neither gender stereotype-threat activation nor gender showed a significant effect and no interaction effect existed between the three variables. - When number of error words for each letter stimulus was entered as a dependent variable, analysis showed that gender-stereotype threat, gender, or task difficulty had no significant effect. - Analysis of mean cluster size showed that gender had a significant effect on mean cluster size, with male participants having higher mean value than female participants - Gender stereotype-threat activation did not have a significant effect on mean cluster size - Analysis of the number of switches showed that neither gender nor gender stereotype-threat activation had a significant effect on the number of switches.

Gender-stereotype Questionnaire and Self-rating: - The mean probability estimate from the male column across all participants was M = 46.16 (SD = 12.74) and differed significantly from a value of 50, t(167) = -3.91, p < 0.0d, d = 0.30, indicating that in general, participants held a gender stereotype favouring women. - The mean probability estimates from male and female participants were also calculated separately. - The mean probability estimate from men indicated that men believed that the hypothetical person with the letter fluency ability was more likely to be a woman. - In other words, they believed that letter fluency task was more likely to have a gender-stereotype favouring women. - The mean probability estimate from women indicated that women did not believe a gender stereotype favouring women existed in the letter fluency task. - In other words, women believed that letter fluency task was more likely to be gender-neutral - The 4 x 2 ANOVA for the probability estimate showed that there was a significant effect of gender, but there was no significant effect of stereotype threat-activation type. - Even though in general participants rated themselves lower than the average score, one sample t-test showed that the overall self-rating score and gender-based self-rating scores (men and women scores calculated separately) did not differ significantly from a value of 4, which was the average score. - The 4 x 2 ANOVA for the self-rating score showed that gender stereotype activation type and gender had no significant effects.

Correlations between Total Correct Words, Total Errors, Mean Cluster Size, Number

of Switches, Gender-stereotype Score, and Self-rating Score: - Significant positive correlations were obtained between total correct words and number of switches, total correct words and the self-rating score, total errors and the gender-stereotype score, total errors and the self-rating, mean cluster size and number of switches, and number of switches and the self-rating score.

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

- Overall, this study found that, although the results of the gender stereotype questionnaire showed that the participants held certain beliefs about the presence of a gender stereotype in the letter fluency test, participants' actual performance was not affected by the stereotype threat.

- This result was still the case even when we accounted for the variation in the stereotype threat activation cues and the level of task difficulty.

- Regarding the variations in the threat activation, our analysis of the manipulation check showed that participants perceived that the cues that activated the gender stereotype threat were gradually increased in terms of emphasizing women's superiority.

- In terms of task difficulty, the analysis showed that the level of task difficulty affected the production of the correct words but did not affect the errors for each letter stimulus.

- Even after the cognitive demand of the test was manipulated the participants' performance under the stereotype threat was not affected.

- It was found that the participants' performance in this test was related to their self-rating, which measured their perception about how well they performed on the letter fluency test.

- The scores from the participants' self-ratings accurately reflected the participants' actual performance on the letter fluency test.

- When the variations in gender and stereotype-threat activation were considered, the scores from the participants' self-ratings showed that these scores were not affected by the variations in gender and stereotype-threat activation cues.

- Collectively, our results showed that neither the participants' perceptions about women's superiority in relation to the activation cues nor their beliefs about the gender stereotypes attached to the task necessarily caused differing performances among the participants during the letter fluency test.

- The participant's own judgement about their ability was more important for their performance than the above mentioned factors.

- This study also examined the effect of gender-stereotype threat on clustering and switching as the underlying cognitive processes of the letter fluency test.

- We found that clustering and switching were not affected by the gender-stereotype threat.

- Based on our findings, we propose that self-efficacy, educational level, and individual moderators may explain on why gender-stereotype threat did not affect letter fluency performance even when the task difficulty had been increased.

- We argue that this positive belief (referring to the self-rating scores) enabled participants to block the negative influence and simultaneously enabled them to employ more efficient cognitive strategies to sustain their performance even when the test became more demanding for their working memory due to increased level difficulty.

- Another important aspect is the participants' educational level as a university student. According to Andreoletti and Lachmann (2004), level of education buffers the effect of the gender-stereotype threat on cognitive performance.

Our findings contribute to the body of literature on stereotype threat by showing that belief about one's ability, namely, selfefficacy, is an important buffer to the effect of a gender-stereotype threat on both overall performance and underlying cognitive process of the letter fluency task. Other personal characteristics that also need to be considered are educational level, stigma consciousness, domain identification, and group identification. Taken together, our findings suggested the importance of taking into account psychological

factors that determine variation in individuals' susceptibility in studying the highly situational phenomenon of stereotype threat.

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)

- no age ranges etc.

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

- not a lot

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

- The conclusions drawn do not make a lot of sense.

- Beliefs about the performance were not triggered until after the letter fluency test.

- Furthermore, using the education level as a buffer of the effect of gender-stereotype threat might be in line with a single paper by Andreoletti and Lachman (2004), however, the vast majority of research is on gender-based stereotypes (e.g. women are bad at math), and done in a university setting with undergraduate students.

- A more reasonable explanation would be that the participants were not aware of the stereotype that was being activated. As in, it is a more common stereotype to say "women are bad at math" than to say "women are better at verbal fluency". Furthermore, maybe Indonesian students' gender-stereotypes are not comparable with the classic "western examples".

- Also, the study did not mention the majors the students were in, their explanation that a different identity might be activated, might be in line with the participants' major. As in, a literature major would be more relevant than the gender, thus protecting the participant from the stereotype threat.

- Furthermore, the age of the participants was not mentioned, verbal fluency might be more correlated with age than other stereotypes - as in, older people are more likely to have a better verbal fluency than younger people. Additionally, the gender identification might be more "fragile" if the participants were to be younger, e.g. in their late teens, compared to their mid twenties.

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.

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*.

Wulandari, S. W., & Hendrawan, D. (2020). Trust your abilities more than the stereotype: Effect of gender-stereotype threat and task difficulty on word production, clustering, and switching in letter fluency. *Pertanika Journal of Social Sciences and Humanities*, *28*(4), 2567–2588. https://doi.org/10.47836/pjssh.28.4.05