Stereotypes, It Is All in Your Head - An EEG Study Investigating Gender Stereotypes in The Brain Through a Lexical Decision Task

Søren Søndergaard Meiner (SM), 202205445@post.au.dk

Marie Vestergaard Thomsen (MVT), 202208819@post.au.dk

School of Communication and Culture, Aarhus University

Jens Chr. Skous Vej, 8000, Aarhus

Supervisor: Andreas Højlund Lorenzen

10th June 2024

AARHUS UNIVERSITY

Keywords: EEG, lexical decision task, gendered stereotype activation, N400, P600

Abstract (MVT)

Stereotypes inevitably impact our perception of the world and ourselves. Gender stereotypes can have a large impact on societal factors such as leadership roles and wage gaps which is why it is important to understand the inherent neurological basis of gender biases and where these stem from. This research paper aims to investigate whether gender stereotypes can be detected in a lexical decision task using electroencephalography (EEG). This was investigated through the event-related potentials N400 and P600 which are respectively related to semantic processing and syntactic anomalies and are typically found in gender stereotype activation. The experiment presented participants with a prime word followed by a target word being either masculine, feminine or a non-word and made them determine whether the target word was a non-word. The study hypothesised that incongruent word pairs would elicit larger event-related responses as measured through EEG. However, the study did not find any significant effects from either of the conditions. The effects of the small sample size are discussed throughout the paper, along with highlighting flaws of the experimental design such as the presence of unwanted semantic spreading activation. Though no significant results were found, the research paper still provides insights into how gender stereotypes are involved in brain processes.

Table of Contents

Motivation

(SM)

Stereotypes are thought to define how we perceive both ourselves and the world which raises the question of whether stereotypes neurologically alter our perception? This sparked the interest for the current study. Stereotypes as a phenomenon can be thought of as a societal construct and affect our self-perception and how we evaluate the world around us (Ellemers, 2018). Gendered stereotypes in particular have led to some of the equality issues that society is facing today such as the wage gap as well as the unequal distribution of gender in board positions. These issues stem back to how we perceive gender which has typically been formed by others' stereotypical views early in life (Fiske, 2017). Everyone has their inherent stereotypes which serve as an energy-saving device for the brain to make faster inferences of the world. Looking at the neurological basis of stereotypes allows us to explore how ingrained these stereotypes are in our worldview. Reaction time has previously been used to look into underlying cognitive processes of decisions related to stereotypes; however, it shows the cumulative time it takes for sensory processing, decision making, and motor response whereas EEG allows to distinguish between these three processes. This results in more valuable insight into underlying cognitive processes compared to when using reaction time. The current paper aims to discover whether there are differences in brain activity in the perception of stereotypically gendered words as measured through event-related potentials (ERPs).

This led us to investigate the following research question: can EEG responses during a lexical decision task reveal the presence of gender stereotypes in language?

Lexical Decision Task and Spreading Activation (MVT)

The Lexical Decision Task (LDT) is commonly used in psycholinguistic experiments. In a typical LDT, participants are placed in front of a screen where they are presented with a pair of words and have to judge whether these words are real words or non-words (Fischler, 1977). Researchers are typically interested in the time it takes to make this judgement (reaction time) when it comes to behavioural data; however, in neuroscientific research, experimenters typically look at differences in brain activity after the presentation of the stimulus (Bermúdez-Margaretto et al., 2019). Meyer & Schvaneveldt (1971) were the first to introduce this type of task and were initially interested in how long-term memory is categorised, and why some information is retrieved more easily than others. They found that when the two words presented were associated, then participants had a faster reaction time compared to when they were not associated. Associated word-pairs could be words such as "bread-butter" and non-associated word-pairs could be "bread-tree". From these results, they concluded that the human brain organises memories in a network and that memories that are more closely related can hence be retrieved faster. Meyer and Schvaneveldt (1971) presented two words at a time placed on top of each other, however; variations of the tasks have been developed and many researchers use a sequential priming paradigm where a prime word is presented followed by a target word in order to "nudge" the participant to think about a certain concept (Neely et al., 1989).

Collins & Loftus (1975) formalised this idea of a semantic network as the theory of spreading activation. They proposed that information is organised in a semantic network with concepts identified as nodes. These nodes are linked together, with associated nodes being closer to each other. This means that a word such as "dog" will be closer to "wolf" compared to "cup". When one node is activated, the activation will spread bilaterally across the conceptual network to nodes that are more closely related to the initial node. There are also variations in how strong some links are, "pillow-bed" will for example have a stronger relation compared to "pillow-room" even though both pairs are semantically associated (Foster et al., 2016).

Stereotype Activation (SM)

Every individual holds different stereotypes about groups of people, and the strength of these stereotypes varies. Kunda and Spencer (2003) presented a theoretical framework for identifying when individuals apply stereotypes to others. They defined stereotype activation as "the extent to which a stereotype is accessible in one's mind" (Kunda & Spencer, 2003). Additionally, they looked at when people apply the activated stereotypes. One of the main distinguishments was that there are both automatic and controlled processes, with controlled processes indicating a stereotype that one can suppress. Since the emergence of stereotype activation as a concept, scholars have looked at it from a neuroscientific perspective, especially through ERP studies. Researchers are typically interested in the automatic processes of stereotype activation to investigate how these are stored in the brain and at what stage of processing they are apparent (Jia et al., 2012; Wang et al., 2017). Behavioural results have supported the existence of stereotype activation by using conceptual priming. The results show that participants display a faster reaction time when a target word is stereotypically consistent with a preceding prime word, compared to if they are stereotypically inconsistent (Tsamadi et al., 2020; Wang et al., 2016).

Previous neuroscientific research has looked at social categorisation and how humans categorise social information, leading up to the formation of stereotypes (Dickter & Bartholow, 2007; Ito & Urland, 2003). These have primarily looked into the categorisation of race and have found differences in out- and -ingroup racial processing. Processing out-group racial faces, for example, a black person seeing a photo of a white person, has been shown to elicit large amplitudes around the P200 component (Dickter & Bartholow, 2007). This suggests that the brain does form social categorisations, which may lead to stereotypes further along in processing. Furthermore, neuroscientific research supports the notion of stereotype activation. Research indicates that once participants are presented with a stimulus provoking a type of stereotype, the brain is then surprised when another element is incongruent with the initial stereotype due to post-perceptual processing (Proverbio et al., 2018).

Electroencephalography (EEG)

(MVT)

Electroencephalography (EEG) is a non-invasive neuroimaging method measuring the electrical activity of the brain. Electrodes are placed on the scalp to measure electrical activity in the form of small voltage fluctuations as a result of neuronal firing. These signals stem from populations of pyramidal cells located under the scalp with dendrite connections to the cerebral cortex, and EEG can therefore be used to get a glimpse of the activity present in the cerebral cortex. If the pyramidal cells synchronise and the number of inputs fire within a narrow time frame, the EEG signal will be larger in amplitude and can be measured. The electrical currents measured are quite small in amplitude and by means of this, the signal is amplified by the EEG recording system which is typically measured in the unit μV. EEG therefore provides insights into cortical activity with excellent temporal resolution but poor spatial resolution, as it cannot be used to pinpoint exact spatial activities (Bear et al., 2016, p.646-658). Due to the high temporal resolution of EEG, the paradigm of using event-related potential (ERP) measures has emerged.

ERP signals are time-locked EEG responses to specific sensory, cognitive or motor events, characterised by their polarity (positive or negative) and their timing (latency). The polarity of ERP signals does not need to be negative or positive in absolute terms but is a measure of the congruency effect (difference between congruent- and incongruent conditions). ERPs are obtained by averaging EEG signals over multiple stimuli which then results in a common signal emerging if there is a brain response to the related event (Kappenman et al., 2021). For instance, the negative 170 (N170) response is an ERP related to the processing of visual stimuli such as faces or words. For studying the effect of the N170 component, one would average trials that contained a stimulus related to the signal and then select only the EEG electrodes related to occipitotemporal sites for analysing the effect (Bentin et al., 1999). The N170 component is therefore typically used to perform sanity checks on EEG data where a visual stimulus has been presented.

The N400 Component (SM)

The N400 component is an ERP that is particularly sensitive to the processing of meaning in language. The response typically appears as a negative deflection 400 ms after a stimulus in the form of a word has been presented (it can also appear with pictures). The N400 is most prominently elicited when the stimulus is either semantically incongruent or unexpected in a given context (e.g. “bread” and “tree”). This supports the theory of spreading activation as it indicates that the amount of processing is dependent on how closely associated the words are.
Though called the N400 component, related brain activity to the N400 component can be measured as early as 200ms after stimulus presentation and until 600ms. The effect of the N400 component is typically largest around centro-parietal sites and the effect is typically a little larger for the right hemisphere (Kutas & Federmeier, 2011).

In a study done by White et al., (2009), using an LDT, they studied the effects of gender stereotype activation (GSA) where they primed participants with either the gender category “Male” or “Female” and then a following target word, which was either consistent with stereotypes related to the gender (e.g Male: Strength) or inconsistent (e.g. Male: Adorable). The study found a significant N400 effect in the incongruent word pairs and slower reaction times compared to the congruent task. In another study done by Wang et al., (2017), utilising an LDT, they compared the effects of the N400 response of lexical semantic activation (LSA; e.g. bread and butter) and GSA. They found that both the reaction times and the congruency effect at N400 were generally larger and started earlier for the GSA than the LSA. This indicates that stereotype activation is detected more quickly and has a greater effect compared to semantic spreading activation.

The P600 Component (MVT)

The P600 is another ERP component that is typically associated with the processing of syntactic anomalies or other processes where a reanalysis or repair of a sentence is required. The effect is a positive deflection typically observed at an onset at 500ms lasting until 800ms. The P600 may seem similar to the N400 component, but syntactic errors (e.g. the cat will eating the food) will elicit a P600 response, and semantic anomalies (e.g. the cat will be eating the chair) will elicit an N400 response. The effect is typically observed in frontal, central, and parietal regions (Kaan et al., 2000; Osterhout & Holcomb, 1992). The effect has also been observed in stereotype-incongruent sentences. Such a stereotypic incongruent sentence could be “The nurse prepared her equipment” and “The nurse prepared his equipment” where the sentence with “his” would elicit a larger P600 effect, reflecting the brain’s effort to process and integrate the unexpected gender information (Osterhout et al., 1997).

In the paper by Wang et al., (2017) they also compared the effects of P600 in the LSA task compared to the GSA task. They found that the GSA task elicited larger amplitudes in P600 responses which suggests that the brain engages more deeply when processing socially relevant information such as gender stereotypes.

Conflicting Theories and Findings (SM)

Both EEG and behavioural findings have supported the notion that stereotype activation is equivalent to semantic activation (Macrae et al., 1994). This has been supported as both theories are based on a conceptual network where stronger associations, either semantic or stereotypical, are more readily available in one's mind. Contreras and colleagues (2011) investigated whether semantic categories and social categories are processed in the same areas of the brain through an fMRI study. Well-established research has previously found that general semantic knowledge is processed in the left inferior frontal gyrus and inferotemporal cortex (Binder & Desai, 2011). Contreras et al. (2011) found that non-social categories were indeed processed in areas typically related to semantic knowledge; however, social categories were processed in areas typically associated with social cognition such as the medial prefrontal cortex, posterior cingulate, and anterior temporal cortex. These findings suggest that social categories of stereotypes are processed differently from non-social categories and semantic activation and stereotype activation should therefore be analysed contrastingly rather than equivalently.

Whilst the theory of semantic networks and spreading activation is well-researched, others have proposed theories indicating how information is stored and retrieved in the brain. One of these theories is the prototype theory by Rosch (1975). She proposed that humans have a prototype for different categories and that new concepts are compared to the prototype for the category. Those concepts that share enough features with the prototype will be thought of as part of the associated category. This theory poses a different way to think about semantic relations compared to the semantic networks proposed by Meyer and Schvaneveldt (1971).

The present paper has used Wang et al. (2017) as a base for the experimental design. They performed a lexical decision task with gendered stereotypes where they used either “male” or “female” as the prime words. The study was performed using Chinese participants and words were presented in Chinese characters. This study uses the same experimental design; however, with slight variations as the prime words are not limited to two words. We therefore aim to investigate whether the findings by Wang et al. (2017) can be seen in Western participants as well by using an experimental design with slight variation from theirs.

Hypotheses (MVT)

Based on previous research investigating gender stereotypes through a lexical decision task, the N400 and P600 components were deemed important components when examining ERPs. Furthermore, the effects of reaction times form a foundation of the overall processes involved in the processing of stimuli. This resulted in the following hypotheses:

Behavioural: It is hypothesised that participants will have a slower reaction time in stereotypically gendered incongruent word pairs compared to stereotypically congruent word pairs when determining whether the word presented is a real word or non-word.

H1: It is hypothesised that through the use of EEG, stereotypically incongruent word pairs will elicit larger N400 amplitudes compared to stereotypically congruent word pairs as measured by a lexical decision task.

H2: It is hypothesised that through the use of EEG, stereotypically incongruent word pairs will elicit larger P600 amplitudes compared to stereotypically congruent word pairs as measured by a lexical decision task.

Methods

Participant (SM)

The participant pool was limited to a single participant due to time and monetary constraints. The participant was a 40-year-old male who completed the experiment five times over two separate days. The participant's native language was Danish but was a fluent English speaker. He reported corrected-to-normal vision and signed a consent form prior to participation.

Materials (Stimuli) (MVT)

The primary database used for the experiment was a database of words with a gendered score from 1-7 developed by Roberts & Utych (2020). They developed the database by having 175 participants determine whether words were masculine or feminine and had to rate them on a scale from 1-7 with 1 being very feminine and 7 being very masculine. The database consists of 700 words and each word was rated by at least 15 participants with a mean of 25 participants. In the database, a score of 4 was considered neutral words, scores below 3 feminine words, and scores above 5 as masculine words. A list of non-words was generated by the use of ChatGPT (OpenAI, 2024) which was prompted to generate words that do not exist but follow phonotactic rules for them to resemble real words.

An electrode cap by ActiCap systems (Brain Products GMBH, Gilching, Germany) was used to ensure standardised placement of the electrodes across the scalp of the participant. The cap contained pre-placement of the electrodes which corresponded to the international 10-20 standardised system of electrode placement with 32 pre-placed electrodes. The electrode cap consists of electrodes that are dispersed across the entire brain to measure activity. The Cz electrode, which is the central electrode placed on top of the head, was used as the common reference electrode. The EEG signal was recorded across the brain scalp using Ag/AgCi electrodes, a BrainAmp DC amplifier with a sampling rate of 1000hz. Additionally, for some of the sessions, an electrooculography (EOG) signal was recorded by the use of electrodes placed on the forehead, used to measure movements and eye blinks, which potentially can be used to correct artefacts (Bear et al., 2016, p. 646-658; Elbert et al., 1985; Klem et al., 1999).

The experiment was conducted in an electrically shielded room at Skejby Universitetshospital, Aarhus, Denmark, designed to minimise electromagnetic interference, ensuring the accuracy of the EEG signal. Participants were placed in front of a monitor with a set distance of approximately 60cm from their eyes. The monitor had a width of 20cm, a resolution of 1200x1000 pixels, and a frame rate of 60hz.
The experiment was made using PsychoPy 2, an open-source framework for building and running psychological experiments using Python (Peirce et al., 2019; Van Rossum & Drake, 2009). All reading and writing of data was done using the Python software Pandas (McKinney & others, 2010).

Experimental design

Cognitive Task (SM)

The trial sequence consisted of first showing the participants a fixation cross for a duration randomly selected between 0.4 and 0.6 seconds to make it less predictable. Following the fixation cross, a prime word was then shown with a duration of 1 second. The prime words were shown in a random order but were of the categories; masculine, feminine, or neutral. Another fixation cross was then shown for a duration between 0.4 and 0.6 seconds. Following the second fixation cross was the target word. The target word was either a masculine, feminine or a non-word. Participants had to press ‘y’ if the word was a real word and ‘n’ for a non-word before the experiment would continue to the next trial sequence. Priming words were displayed in uppercase letters while the target words were shown in lowercase for easier differentiation between categories The trial sequence is also displayed in Figure 1.

In total the participant was exposed to 37 congruent trials, 35 incongruent trials, 44 neutral trials and 232 non-words. It was ensured that half of the trials were non-words in order to prevent any bias when clicking either 'y' or 'n'. Each trial lasted around 4 minutes.

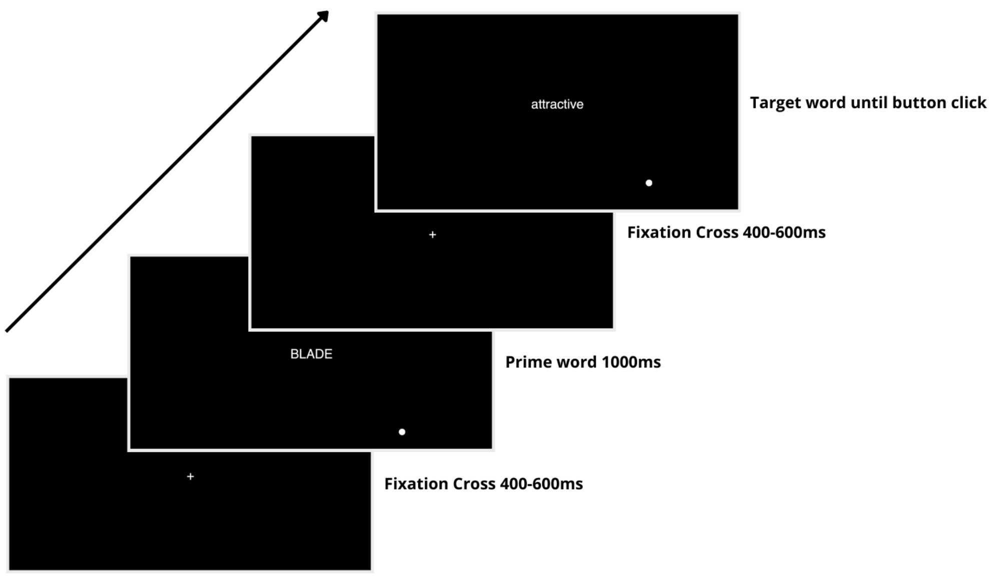


Figure 1: A trial sequence

EEG data was recorded throughout the experiment and specific triggers were recorded at the onset of the fixation cross, prime words, and target words. This enabled us to mark the specific trials as events in the EEG recording, allowing for analysis of these events at a later stage. While running the experiment, logging of the responses, reaction times, correctness, and condition was also recorded and saved to a CSV file.

Procedure (MVT)

The experiment followed a within-subject design, where each participant was exposed to a set of trials. This design allows for control of individual differences since each participant serves as their own control (Charness et al., 2012). The participant was exposed to three conditions: congruent, incongruent, neutral, and non-word. These are presented in Table 1.

Condition Table
Prime and Target Conditions

Condition	Prime Type	Target Type
Congruent	Feminine	Feminine
Congruent	Masculine	Masculine
Incongruent	Feminine	Masculine
Incongruent	Masculine	Feminine
Neutral	Neutral	Masculine
Neutral	Neutral	Feminine
Non-word	Feminine;Masculine	Non-word

Table 1

In the introduction of the experiment, participants were greeted with an introductory message explaining the experiment and the procedure and they were then presented with the cognitive task explained above. Participants were instructed to pay extra attention to the priming and target words.

Analysis

Behavioural Analysis (SM)

The behavioural data was preprocessed and analysed in Rstudio version 2024.4.0.735 (Posit Team, 2024). During the preprocessing of the data, outliers were removed by only including reaction time values smaller than 5 s. This was done to exclude trials where the participant may have been unfocused. Furthermore, the conditions were balanced to include the same number of trials in each condition to do a paired samples t-test as the analysis. Each condition occurred 27 times after they had been balanced.

Electrophysical Analysis

Preprocessing (MVT)

Prior to doing any analysis, the electrophysical data needed preprocessing. The trials were preprocessed separately as they were conducted on separate days and hence there were differences in how the electrodes were placed on the scalp.

The MNE package was used in Python to preprocess the data (Larson et al., 2024). One of the first steps was to do common average referencing. In this process, the reference signal is changed from being the activity from the central electrode (Cz) to the average electrical activity across all electrodes. This is done to ensure that activity from other electrodes is not heavily impacted in case the central electrode displays a lot of noise.

Once common average referencing had been done, the next step involved looking into the stimuli triggers. A total of 29 spurious triggers were identified. These were rejected by removing all stimuli triggers that had an onset less than 2 ms after a previous trigger. Once the spurious triggers had been removed, there was a total amount of 180 stimuli triggers from the first day, and 270 on the second day leaving a total of 450 stimuli triggers including both prime and target words in all conditions. In EEG experiments, it is common to correct for ocular artefacts by doing an ICA analysis. This was however not shown to be necessary as the frontal electrodes were not used and potential artefacts from eye blinks did not affect the relevant channels.

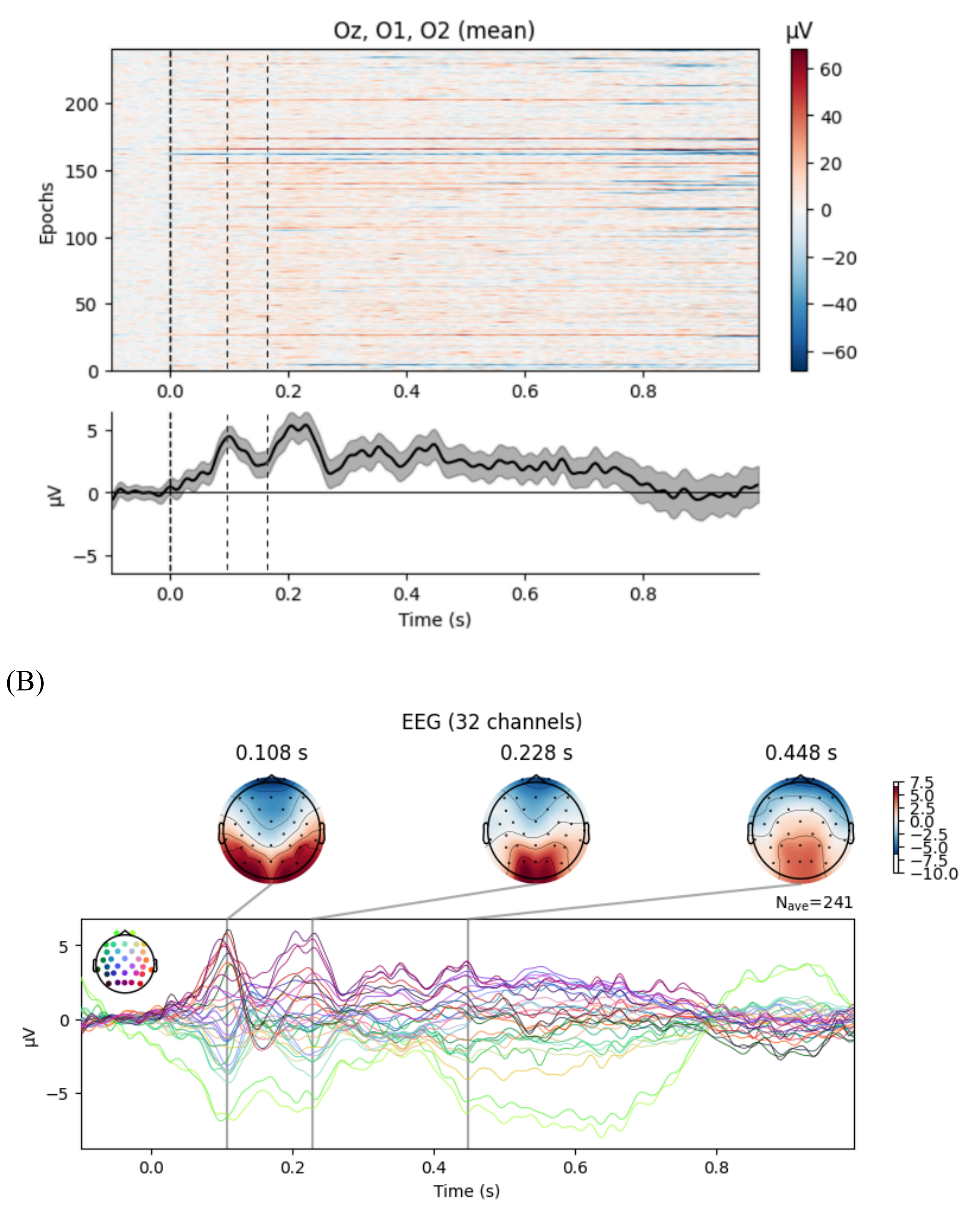
The electrical activity was high pass filtered at 0.1 Hz and low pass filtered at 40 Hz following the standard EEG practices (Bigdely-Shamlo et al., 2015). A high pass filter at 0.1 Hz allows for filtering out slow drifts in the data and a low pass filter of 40 Hz is used to filter out muscle-related artefacts or other types of noise as they are typically around 50Hz or more. Once triggers were identified and the data had been filtered, a diode check was done to ensure triggers were displayed at the time given in the data. It was found that the trigger stimuli were delayed by 33.33 ms which is equivalent to two refresh frame rates of 60Hz lasting 16.67 ms each. Additionally, the prime stimuli were delayed by one frame rate i.e. 16.67 ms. The timestamp for each trigger was then adjusted accordingly. By plotting the channels and visually inspecting the data, a few channels were identified as too noisy and were removed. In the sample from the first day, P3, P010, and P09 were removed whereas no channels were removed in the second sample.

After the raw data had been preprocessed, the next step involved creating epochs. Epochs were sampled from -100 ms to 1000 ms as we were interested in the time window from 100 ms prior to the onset of the stimulus to 1s after the stimulus had been shown. A rejection criterion of 160 microvolts was used as any activity either above or below 150 microvolts was deemed as artefacts. A baseline correction was used to correct for temporal drifts that are unrelated to the experiment and it was set from 100 ms prior to stimulus onset until stimulus onset. Epochs were then downsampled to 250 Hz to allow for less computational processing when analysing the epochs. This resulted in 14 congruent epochs, 18 incongruent epochs, and 15 neutral epochs in the first sample. In the second sample, there were 23 congruent epochs, 17 incongruent epochs, and 29 neutral epochs.

Sanity Checks (SM)

Before doing any analysis, a few sanity checks were done to ensure the quality and validity of the data. When doing the sanity checks, we were interested in seeing an effect at P100 and N170 as a result of visual stimuli (Bentin et al., 1999; Woodman, 2010). Plot A in Figure 2 shows a clear P100 effect indicating that the measured data is valid as the participant has registered the visual stimuli. Additionally, plot B supports this as it is clear that the occipital channels detect a signal around 100 ms post-stimulus onset and a negative deflection at approximately 170 ms indicating that the word is being processed at this ERP.

(A)



(B)

Figure 2

By inspecting Figure 3 (GFP) it is evident that the same general trend is seen across conditions. The Global Field Power (GFP) is a way to visualise the spatial activity in the brain across time and hence the plot displays that both positive and negative deflections are present at approximately the same time in both conditions. This shows that no condition includes wrongful data and that both conditions look sensible.

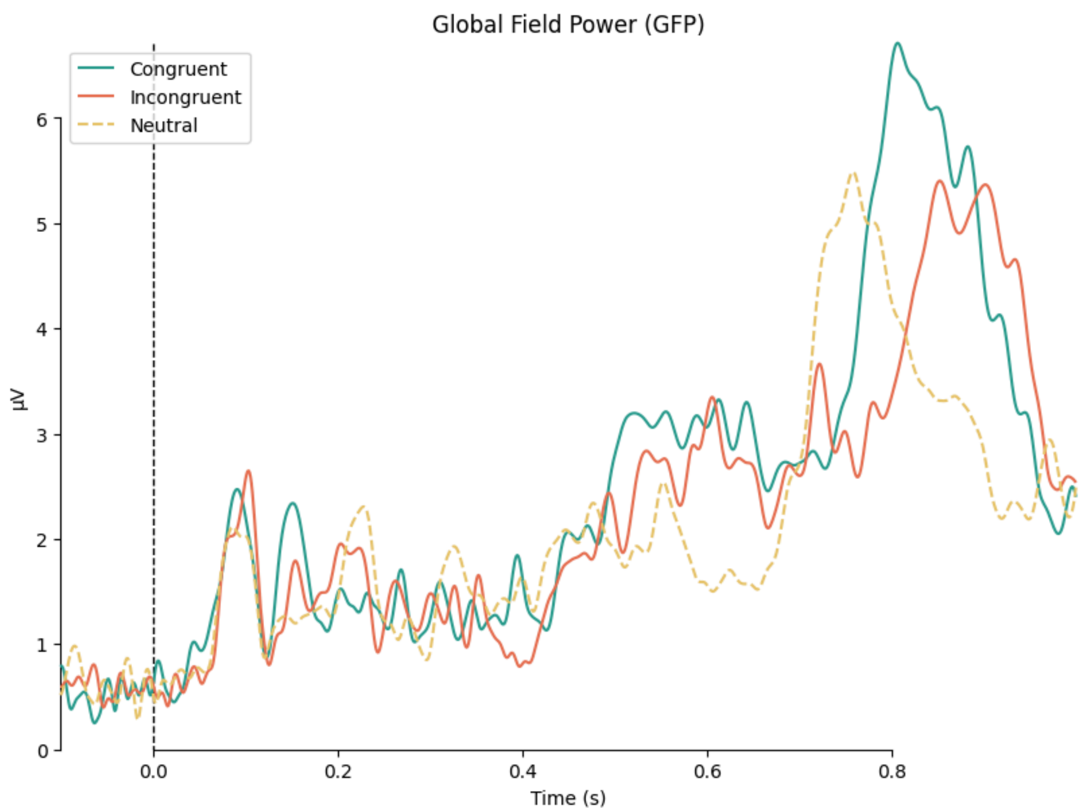


Figure 3

Analysis Methods (SM)

Based on prior research, specific channels were chosen to detect the relevant components The channels for N400 were F3, F4, Fz, C3, Cz, and Pz (Kutas & Federmeier, 2011; Wang et al., 2017; White et al., 2009), and the channels for P600 were F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4 (Kaan et al., 2000; Osterhout et al., 1997; Wang et al., 2017). A windowed mean subject level t-test was used to determine whether there was a difference in amplitude between the incongruent, congruent, and neutral conditions. In order to conduct the windowed mean t-test, data was averaged across trials as well as the relevant channels, resulting in one data point per participant. As we only had one participant it was sufficient to conduct a subject-level t-test, otherwise a group-level t-test would have been necessary. A t-test was deemed sufficient as we only investigated the differences between the two conditions. If the experiment had included multiple predictors, a general linear model would have been more suitable; however, as there was only one predictor (condition) of brain activity (voltage) we did not consider it necessary. The t-test for N400 was conducted at 300 ms-500 ms and the time window for P600 was at 500 ms - 800 ms as identified by previous research (Kutas & Federmeier, 2011; Wang et al., 2017; White et al., 2009).

Results

Behavioural Results (MVT)

As the assumption of normality by the Shapiro-Wilk test was met, a paired t-test was conducted on the reaction times from the behavioural experiment. A total of three t-tests were conducted between conditions and none of them showed significant results. Conducting a t-test between the congruent (M= 2.55, SD=0.132) and incongruent (M= 2.58, SD=0.131) conditions yielded no significant results t(26)=0.656, p=.518, d=.019). The results between the neutral and congruent condition, as well as the neutral and incongruent condition, did not reveal significant results either. The results from the t-test can be seen in Appendix 1.

The conditions are visualised in Figure 4 showcasing means and standard deviations.

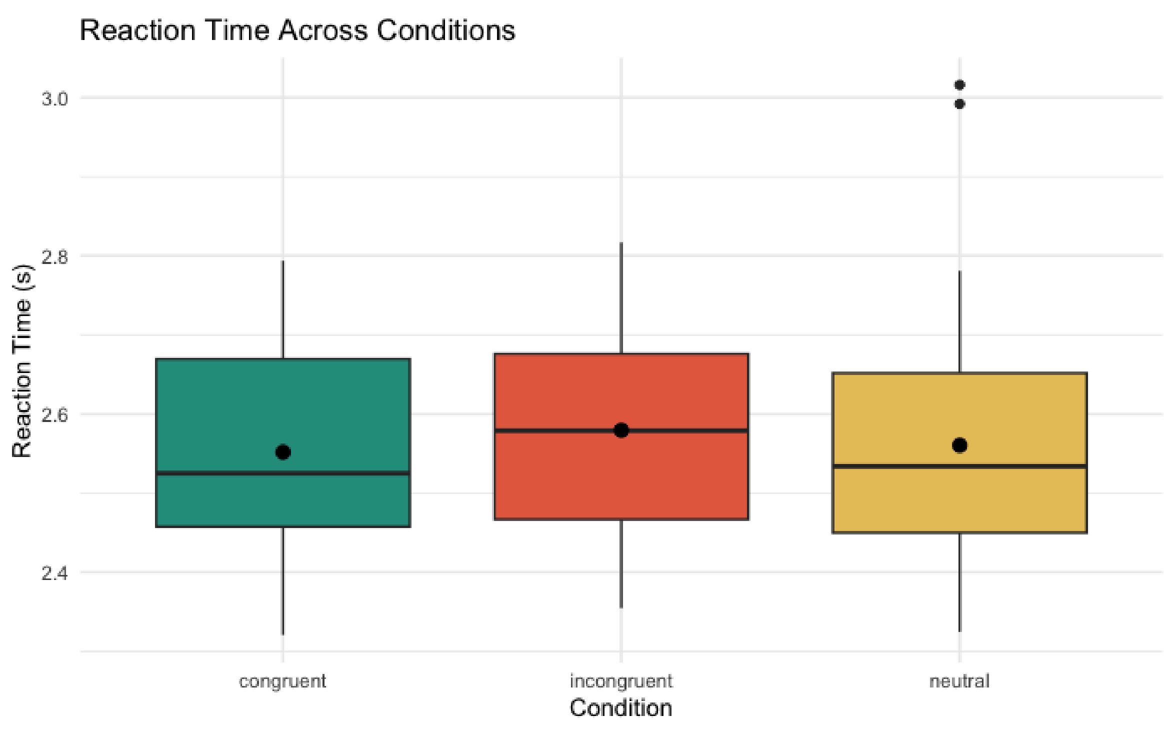


Figure 4

Electrophysical Results (SM)

Three separate t-tests were conducted for each of the two components investigated. Each condition was tested against the neutral condition for baseline comparison. The hypothesis is tested through the t-tests between the congruent and incongruent conditions. The t-test for N400 revealed no significant results $t(38) = 0.253$, $p = .802$, and neither did the t-test for P600 $t(44) = 0.676$, $p = .502$. The results between the conditions can be seen in Table 2. A visualisation of the results from both components can be seen in Figure 5.

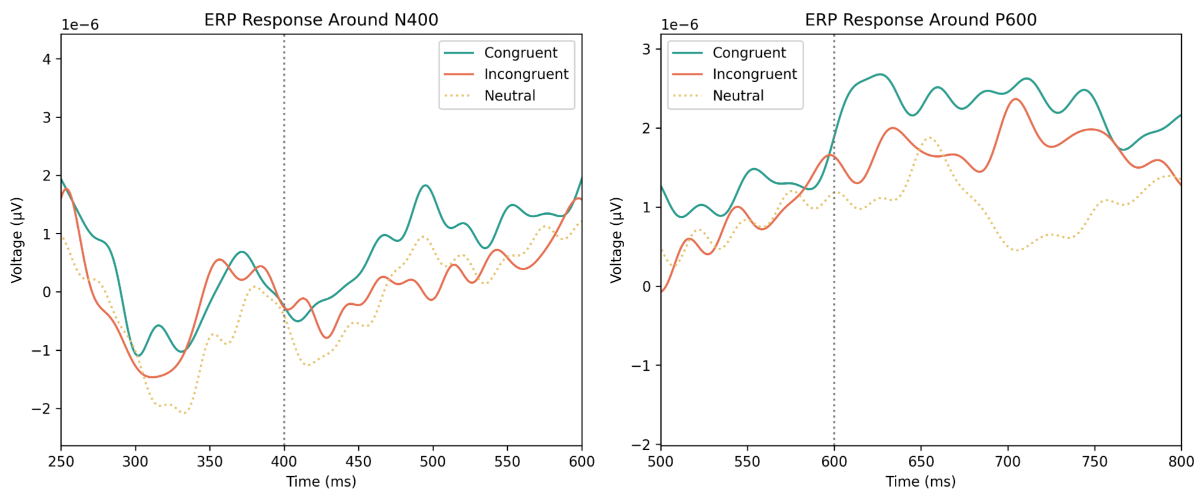


Figure 5

P600 and N400 Comparisons

Effect Sizes, Degrees of Freedom, and P-values

Comparison	Effect Size	Degrees of Freedom	P-value
P600: Neutral vs Congruent	0.113	38.000	0.910
P600: Neutral vs Incongruent	0.821	50.000	0.415
P600: Congruent vs Incongruent	0.676	44.000	0.502
N400: Incongruent vs Congruent	0.253	38.000	0.802
N400: Neutral vs Congruent	0.484	50.000	0.630
N400: Neutral vs Incongruent	−0.060	44.000	0.953

Table 2: P600 and N400 t-test results table.

Discussion

Key Insights and Takeaways

Findings (MVT)

From the behavioural results, none of the conditions had any significant results as shown by the t-tests. By looking at the visualisation from Figure 5 it might seem that there is a minuscule difference and that reaction times for the incongruent condition were a bit slower but none of the conditions significantly affected reaction time. Therefore, the hypothesis of the behavioural design cannot be supported. It still might be that people hold gendered stereotypes for specific words, especially since they can be categorised on a scale as shown in the study by Roberts & Utych (2020).

The N400 component did not reveal any statistical differences between conditions. For the N400 component, small variations are present in Figure 6 (from 400 ms-550 ms), which are consistent with the hypothesis. However, these variations were not deemed large enough by the t-test to show any statistically significant differences. These results indicate that if any semantic incongruencies were present, the effect of these was not large enough to elicit a prominent N400 response.

The P600 component also did not reveal any statistical differences between conditions. None of the results from the P600 component was consistent with our hypothesis and Figure 6 even shows the opposite effect (600ms-700ms) from the one, which was expected that the congruent task seemed to elicit a larger P600 response. However, these effects are very slight and insignificant. These results indicate that the participant did not experience any syntactic anomalies deemed large enough to elicit a significant P600 effect.
The lack of significant results in both the behavioural data and EEG may indicate flaws in the experimental design or other issues, which will be discussed later on.

Comparison with Previous Studies (SM)

Previous research has demonstrated that the brain seems to be orchestrated in such a way, that different words hold stereotypes that can elicit an EEG response with a connected ERP (Macrae et al., 1994; Osterhout et al., 1997; Siyanova-Chanturia et al., 2012). In the study by Wang et al. (2017), they demonstrated the presence of stereotype activation in neural responses in both the N400 and P600 components and the responses were larger in amplitude compared to the LSA task. The research by White et al., (2009) also found significant effects of gender stereotypes in the N400 component.

The evidence of gender stereotypes in the brain seems to be strong and one might then ask: why did this research not find any significant results in any conditions if the effects seem to be so substantial in previous research? One key difference from this study compared to the other studies on the subject lies in the experimental design.

The other studies employed a more traditional way of priming using a gender category. This study used a different research approach to the effects of gendered stereotypes as the priming words consisted of both adjectives and nouns. Unlike the other studies, this study aimed to uncover any subtle and inherent gender biases in everyday language which was not primed by a gender category. While the study did not find any significant effects, it still provides valuable insights into how stereotypes may be neurologically encoded, but also equally important, how they may not be.

The study by Contreras et al. (2011) also suggested that social categories of stereotypes are processed differently from the non-social categories. These implications could contribute to the explanation of why stereotype activation might not have been detected as prominently in our study as compared to previous research. Specifically, the differential processing of social versus non-social categories could imply that the experimental design failed to adequately isolate and capture the neural mechanisms specific to social stereotype activation. This also suggests that perhaps stereotype activation should not be compared equivalently to semantic processing. Prototype theory proposes a different way of storing information in the brain (Rosch, 1975). This theory may imply we have more specific categories for social and non-social information and new concepts are compared to these categories, meaning social and non-social information will not be processed equally. If this is the case, one may hypothesise that semantic and social information will be related to different components in EEG.

Methodological Limitations (MVT)

A primary limitation of the present study was the small sample size of one participant. People differ in their brain activity patterns which means one participant is not sufficient to be able to generalise results to a bigger population. This decreases the reliability of the results and limits the possibility of drawing any conclusions. The participant took part in the experiment five times. As a result, he might subconsciously have recognised the words presented meaning the effect of the words would decrease over trials as he got used to the different conditions. The last trials are therefore not representative.

Furthermore, the participant identified as male which may impact how the stereotypically gendered words are perceived. When investigating gender stereotypes in males and females, it is important to have both genders represented as previous research has shown that men and women have different perceptions of gender stereotypes (Diekman & Eagly, 2000). This means that an individual identifying as male will relate certain attributes to females; however, females may have a different perception and are more likely to relate other attributes to their gender. Theoretically, this means the links between the conceptual networks are more strongly tied to some attributes compared to others depending on one's gender (Collins & Loftus, 1975). Additionally, the participant was proficient in English but it was not his primary language. This might have influenced his perception of gendered words as he may not have had as strong associations with the English words compared to words in his mother tongue.

The experimental design of the current study poses several weaknesses. One of the flaws of the present study is that there may be instances where the target word and prime word are semantically related. In this case, it is not possible to determine whether a potential effect would be due to stereotype activation and the presence of stereotypes, or rather just a semantic connection. An instance was noticed where “brutal” was the prime and “rebel” was the target word. In this case, one may have a semantic category for crime in which both of these are included. Previous research has shown that semantic congruency also can be related to the N400 and P600 components (Dien & O’Hare, 2008; Wang et al., 2017). Hence, word pairs that have both semantic and stereotypic congruency may elicit both effects which cannot be differentiated in the analysis. In order to avoid this, it is important to ensure that the words presented are semantically unrelated.

Another aspect of the present study that may pose a limitation is the uncertainty in ERPs. The experimental setup is based on a decision task (judging whether the target word is a real word or non-word) and a portion of cognitive resources will be devoted to the decision making process. The distinction between whether a potential ERP arises from the decision making process or stereotype activation further complicates the analysis. Throughout the paper, previous research has been presented supporting the involvement of the N400 and P600 components in stereotype activation. However, to our knowledge, none of these studies have considered the potential impact of the decision-making process on parietal-frontal channels used for analysis, and consequently, on the responses attributed to stereotype activation. The decision making process may induce more activity in the frontal regions of the brain and perhaps contribute to the negative deflection at 400 ms and positive deflection at 600 ms (Yates & De Oliveira, 2016).

Additionally, the present experiment was about 4 minutes long, limiting the number of trials. To be able to extrapolate conclusions from a short experiment like the current one, either a great number of participants is needed, or the experiment needs to be longer to get a greater number of data points for each participant. In this case, the duration of the task was limited as the words from the database were limited (Roberts & Utych, 2020).

Using EEG as a method to detect brain activity also has its limitations. As EEG measures electrical activity in the brain, other sources emitting electromagnetic activity can create a lot of noise in the data. When doing the experiment on the first day, there were some monitors in the experimental room which created a lot of noise in the data. This makes it harder to analyse the data and prevents the results from being as clear compared to if there was no noise. Additionally, EEG signals are in general impacted by fine muscle movements such as clenching one's jaw, eye blinks, and movement of the head. These muscle movements create artefacts in the data and can often reduce the number of trials in an ERP study.

Theoretical Considerations and Limitations (SM)

The evidence of spreading activation, stereotypes in the brain, and lexical decision tasks are well-documented documented theoretical foundations on which this study has been built (Collins & Loftus, 1975; Fischler, 1977; Foster et al., 2016; Mills et al., 2012; Osterhout et al., 1997). It is also for this reason that the validity of the theories is not discussed in this section but will contrarily focus on the variations that can influence these theoretical frameworks.

Theory states that stereotype activation is affected by the associative strengths between the links in a network of stereotypes (Kunda & Spencer, 2003). In other words, a stereotype is more accessible in one's mind depending on the strength of the belief. The strength of the belief may depend on the frequency and occurrence in people's lives. This means that some stereotypes will be more prominent than others as individuals become more familiar with them when they occur more frequently. As a result, the magnitude of the N400 component may be impacted by the associative strength of the links. If someone, for example, has a strong belief that men are very brave, perhaps because they have encountered this co-occurrence frequently, then that specific association will be stronger and the N400 magnitude will be impacted by this.

Various factors can impact the N400 latency i.e. at what time after stimulus presentation the N400 effect occurs. Wang et al. (2017) found that N400 starts earlier in stereotype activation compared to spreading activation in a semantic network. They attributed this to be due to the fact that social information may be more readily available in the brain as humans are more used to processing social information compared to nonsocial (semantic) information (Wang et al., 2017). Furthermore, it has previously been found that language proficiency impacts the N400 latency. If you are more proficient in the language presented, then N400 will occur reliably earlier compared to someone who is less proficient (Kutas & Federmeier, 2011). Based on the findings, it can be suggested that the time point at which N400 occurs may differ depending on various factors and a time span around 400 ms post-presentation of the stimuli should therefore be investigated.

One way this study could differ compared to other studies is the way culture plays a role in the experimental design. Since stereotypes vary across cultures there is not a “one-size-fits-all” experimental design (e.g. using equivalent nouns and adjectives across languages) that can be employed across cultures and the effects may also be either larger or smaller across cultures. For instance, the study by Wang et al. (2017) consisted of only Chinese participants and the results may have been different if it was a Chinese-speaking participant from a different country with a different culture, who had completed their experiment. For some cultures, it may be that gender stereotypes are more ingrained and the effect seen would have been greater. It is also for this instance that these cross-cultural EEG studies can be used to compare and assess general stereotypical views of genders in societies (Ellemers, 2018; Macrae et al., 1994).

The language spoken itself can also impact stereotypes, especially for languages that contain grammatical genders. In a study done by Flaherty (2001), she examined the impact of a gendered language system on perception by comparing Spanish which has robust grammatical genders, with English which does not have robust grammatical genders. Her results showcased that Spanish had a noticeable difference in assignment tasks. The Spanish-speaking participants were more likely to assign gendered attributes to objects based on the grammatical gender of the object. This research highlights the importance of considering linguistic structures when examining gender stereotypes and their manifestations in culture. This difference in language on perception also highlights the importance of why there is not a “one-size-fits-all” experimental design for investigating gender stereotypes.

Key Improvements for a Full-Scale Study (MVT)

In order to scale this to a full-scale study, several aspects would have to be changed or improved. As mentioned previously, the primary flaws in the present study are the limited participant pool and inadequate experimental design. A full-scale study would need to include significantly more participants and both genders should be represented. Additionally, stereotype activation could be investigated in other modalities as well such as auditory or visual (in terms of pictures). Other types of stereotypes, such as race or ethnicity should also be investigated to draw any conclusions about stereotypes in general.

Prior to the EEG experiment, it would have been beneficial to test whether the participant holds gender stereotypic beliefs. If an individual does not have any gender stereotypic beliefs, then according to spreading activation theory, the connection between the nodes in a stereotype network will not be very strong and hence little to no effect should be seen. A number of tests have been developed to test whether an individual holds gender stereotypes. These include the BEM Sex Role Inventory (Bem, 2011) and IDR-GST (Mills et al., 2012). Using one of these would help determine the degree to which individuals hold stereotypic beliefs and perhaps could be a predictor of how great the N400 and P600 effects are. An interesting investigation could be whether the level of stereotypic beliefs can predict how great the effect is at either N400 or P600. For further investigation one could perform a multilevel linear model including both the micro voltage of brain activity at 400 ms and 600 ms after the presentation of the stimuli, as well as participants' scores on a gender stereotype test. In this way, one could investigate whether the ERP components are dependent on the strength of stereotypic beliefs.

Instead of presenting the participants with either adjectives or nouns in a random order to find the effects of spreading activation, the experimental design could have been changed by categorising the prime words as nouns and the target words as adjectives (e.g. woman, emotional). This could potentially result in a greater degree of spreading activation related to stereotypic masculine and feminine categories.

Furthermore, the experiment duration would have to be increased to get a greater number of trials and data points for each participant. The database used (Roberts & Utych, 2020) was limited in the number of words and in order to scale the experiment one would have to either build upon this database or search the field for other databases that perhaps could be combined with the current. To further develop the database, researchers would have to ask people to score words based on their level of femininity and masculinity.

Conclusion (SM)

The present study investigates whether gender stereotypes can be detected in the brain through a lexical decision task. Based on a theoretical foundation, stereotype activation was thought to impact both the N400 and P600 components when investigating EEG responses. The experimental setup followed a typical LDT where there were three conditions being; congruent, incongruent, and neutral. Results from t-tests showed no significant results between conditions and therefore no conclusions can be generalised to a population. The study was confined due to the small sample size of one participant resulting in limited statistical power. Furthermore, flaws in the experimental setup were identified and in order to scale it to a full-scale study various aspects would have to be modified. Limitations to the theory were also identified as suggestions for why the study did not show any significant results.

References

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2016). *Neuroscience: Exploring the brain* (4th ed., international ed). Wolters Kluwer.

Bem, S. L. (2011). *Bem Sex Role Inventory* [dataset]. https://doi.org/10.1037/t00748-000

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260.

Bermúdez-Margaretto, B., Beltrán, D., Cuetos, F., & Domínguez, A. (2019). Novel Word Learning: Event-Related Brain Potentials Reflect Pure Lexical and Task-Related Effects. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00347

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00016

Binder, J. R., & Desai, R. H. (2011). The Neurobiology of Semantic Memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*, *7*(7), 764–770. https://doi.org/10.1093/scan/nsr053

Dickter, C. L., & Bartholow, B. D. (2007). Racial ingroup and outgroup attention biases revealed by event-related brain potentials. *Social Cognitive and Affective Neuroscience*, *2*(3), 189–198. https://doi.org/10.1093/scan/nsm012

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*(10), 1171–1188. https://doi.org/10.1177/0146167200262001

Dien, J., & O'Hare, A. J. (2008). Evidence for automatic sentence priming in the fusiform semantic area: Convergent ERP and fMRI findings. *Brain Research*, *1243*, 134–145. https://doi.org/10.1016/j.brainres.2008.09.045

Elbert, T., Lutzenberger, W., Rockstroh, B., & Birbaumer, N. (1985). Removal of ocular artifacts from the EEG--a biophysical approach to the EOG. *Electroencephalography and Clinical Neurophysiology*, *60*(5), 455–463. https://doi.org/10.1016/0013-4694(85)91020-x

Ellemers, N. (2018). Gender Stereotypes. *Annual Review of Psychology*, *69*(Volume 69, 2018), 275–298. https://doi.org/10.1146/annurev-psych-122216-011719

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, *5*(3), 335–339. https://doi.org/10.3758/BF03197580

Fiske, S. T. (2017). Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) versus Variable Stereotypes (Race, Ethnicity, Religion). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(5), 791–799. https://doi.org/10.1177/1745691617708204

Flaherty, M. (2001). How a Language Gender System Creeps into Perception. *Journal of Cross-Cultural Psychology*, *32*(1), 18–31. https://doi.org/10.1177/0022022101032001005

Foster, P. S., Hubbard, T., Campbell, R. W., Poole, J., Pridmore, M., Bell, C., & Harrison, D. W. (2016). Spreading activation in emotional memory networks and the cumulative effects of somatic markers. *Brain Informatics*, *4*(2), 85–93. https://doi.org/10.1007/s40708-016-0054-2

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, *85*(4), 616–626. https://doi.org/10.1037/0022-3514.85.4.616

Jia, L., Dickter, C. L., Luo, J., Xiao, X., Yang, Q., Lei, M., Qiu, J., & Zhang, Q. (2012). Different brain mechanisms between stereotype activation and application: Evidence from an ERP study. *International Journal of Psychology: Journal International De Psychologie*, *47*(1), 58–66. https://doi.org/10.1080/00207594.2011.580348

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and Clinical Neurophysiology. Supplement*, *52*, 3–6.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*(4), 522–544. https://doi.org/10.1037/0033-2909.129.4.522

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Larson, E., Gramfort, A., Engemann, D. A., Leppakangas, J., Brodbeck, C., Jas, M., Brooks, T., Sassenhagen, J., McCloy, D., Luessi, M., King, J.-R., Höchenberger, R., Goj, R., Favelier, G., Brunner, C., van Vliet, M., Wronkiewicz, M., Rockhill, A., Holdgraf, C., … luzpaz. (2024). *MNE-Python* (v1.7.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10999175

Macrae, C., Milne, A., & Bodenhausen, G. (1994). Stereotypes as Energy-Saving Devices: A Peek Inside the Cognitive Toolbox. *Journal of Personality and Social Psychology*, *66*, 37–47. https://doi.org/10.1037/0022-3514.66.1.37

McKinney, W. & others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Mills, M. J., Culbertson, S. S., Huffman, A. H., & Connell, A. R. (2012). Assessing gender biases: Development and initial validation of the gender role stereotypes scale. *Gender in Management*, *27*(8), 520–540. https://doi.org/10.1108/17542411211279715

Neely, J. H., Keefe, D. E., & Ross, K. L. (1989). Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1003–1019. https://doi.org/10.1037/0278-7393.15.6.1003

OpenAI. (2024). *ChatGPT* (4.0) [Computer software]. https://chat.openai.com/chat

Osterhout, L., Bersick, M., & Mclaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285. https://doi.org/10.3758/BF03211283

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Posit Team. (2024). *RStudio: Integrated Development Environment for R* [R]. Posit Software, PBC. http://www.posit.co/

Proverbio, A. M., Alberio, A., & De Benedetto, F. (2018). Neural correlates of automatic beliefs about gender stereotypes: Males are more prejudicial. *Brain and Language*, *186*, 8–16. https://doi.org/10.1016/j.bandl.2018.08.006

Roberts, D. C., & Utych, S. M. (2020). Linking Gender, Language, and Partisanship: Developing a Database of Masculine and Feminine Words. *Political Research Quarterly*, *73*(1), 40–50. https://doi.org/10.1177/1065912919874883

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*. https://www.academia.edu/24474401/Cognitive_representations_of_semantic_categories

Siyanova-Chanturia, A., Pesciarelli, F., & Cacciari, C. (2012). The Electrophysiological Underpinnings of Processing Gender Stereotypes in Language. *PLoS ONE*, *7*(12), e48712. https://doi.org/10.1371/journal.pone.0048712

Tsamadi, D., Falbén, J. K., Persson, L. M., Golubickis, M., Caughey, S., Sahin, B., & Macrae, C. N. (2020). Stereotype-based priming without stereotype activation: A tale of two priming tasks. *Quarterly Journal of Experimental Psychology*, *73*(11), 1939–1948. https://doi.org/10.1177/1747021820925396

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wang, P., Yang, Y.-P., Tan, C.-H., Chen, Q.-W., & Cantfort, T. (2017). Gender Stereotype Activation versus Lexical Semantic Activation: An ERP Study. *The Journal of General Psychology*, *144*(4), 283–308. https://doi.org/10.1080/00221309.2017.1310714

Wang, P., Yang, Y.-P., Tan, C.-H., Zhao, X.-X., Liu, Y.-H., & Lin, C.-D. (2016). Stereotype activation is unintentional: Behavioural and event-related potenials evidence. *International Journal of Psychology*, *51*(2), 156–162. https://doi.org/10.1002/ijop.12135

White, K. R., Crites, S. L., Taylor, J. H., & Corral, G. (2009). Wait, what? Assessing stereotype incongruities using the N400 ERP component. *Social Cognitive and Affective Neuroscience*, *4*(2), 191–198. https://doi.org/10.1093/scan/nsp004

Woodman, G. F. (2010). A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention. *Attention, Perception & Psychophysics*, *72*(8), 10.3758/APP.72.8.2031. https://doi.org/10.3758/APP.72.8.2031

Yates, J. F., & De Oliveira, S. (2016). Culture and decision making. *Organizational Behavior and Human Decision Processes*, *136*, 106–118. https://doi.org/10.1016/j.obhdp.2016.05.003

Appendices

Appendix 1

Comparisons of Reaction Times
Incongruent, Neutral, and Congruent Conditions

Comparison	t-value	Degrees of Freedom (df)	p-value
Incongruent vs Congruent	−0.656	26.000	0.518
Neutral vs Congruent	−0.770	35.000	0.446
Neutral vs Incongruent	0.297	28.000	0.769