Journal of Neurolinguistics 53 (2020) 100876

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling

Research paper

Male fashionistas and female football fans: Gender stereotypes affect neurophysiological correlates of semantic processing during speech comprehension

Angela Grant[a,*], Sarah Grey[a,b], Janet G. van Hell[a]

[a] Department of Psychology and Center for Language Science, 140 Moore Building, The Pennsylvania State University, University Park, PA, 16801, USA

[b] Department of Modern Languages and Literatures, Fordham University, Bronx, NY, 10458, USA

ARTICLE INFO

Keywords:
Semantics
Gender
Stereotypes
ERPs
N400
Social knowledge

ABSTRACT

Recent studies have shown that pre-existing contextual information, such as gender stereotypes, is incorporated online during comprehension (e.g., Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). Stereotypes, however, are not static entities, and social role theory suggests that they may be influenced by the behavior of members of the group (Eagly, 1987). Consequently, our study examines how gender stereotypes affect the semantic processing of statements from both a male and a female speaker, as well as investigating how the influence of stereotypes may change as listeners gain experience with individual speakers. Participants listened to male and female speakers produce sentences about stereotypically feminine (fashion) and stereotypically masculine (sports) topics. Half of the participants heard a stereotype congruent pattern of sentences (e.g., for the male speaker, semantic errors about fashion but no semantic errors on sports sentences) and the other half heard a stereotype incongruent pattern. We found that the N400 effect of semantic correctness is larger in stereotype incongruent conditions. Furthermore, in stereotype congruent conditions, only stimuli presented in the male voice show an N400 effect in the expected direction (larger N400s to semantic violations). Additionally, when we examined ERP changes over the course of the experiment, we found that the degree of change in amplitude was predicted by individual differences in ambivalent sexism. These results suggest that not only are speaker characteristics incorporated during online language processing, but also that social knowledge influences language processing in a manner congruent with social role theory.

1. Introduction

In the specialization of language research into individual disciplines (e.g., phonology, semantics, pragmatics), and under the umbrella of two-stage theories that suggest pragmatic language processing follows semantic and syntactic processing (e.g., Grice, 1975), pragmatic language processing has often been investigated separately and independently from other aspects of language (see McNally, 2013). In recent years, however, the view of pragmatics as a separable aspect of language processing has begun to change, and multiple studies show that pragmatic context affects the neurophysiological correlates of semantic processing (Baetens, der Cruyssen, Achtziger, Vandekerckhove, & Van Overwalle, 2011; Creel & Tumlin, 2011; Delaney-Busch & Kuperberg, 2013; Hagoort,

* Corresponding author. Cognition Laboratory, Department of Psychology, Missouri Western State University, 4525 Downs Drive, St. Joseph, MO, 64507, USA.

E-mail address: agrant5@missouriwestern.edu (A. Grant).

https://doi.org/10.1016/j.jneuroling.2019.100876
Received 2 May 2019; Received in revised form 30 August 2019; Accepted 10 October 2019

Hald, Bastiaansen, & Petersson, 2004; Lau, Holcomb, & Kuperberg, 2013; Nieuwland & Van Berkum, 2006; Rommers, Dijkstra, & Bastiaansen, 2013; Van Berkum, 2008; Van Berkum et al., 2008) and grammatical processing (Grey & Van Hell, 2017; Hanulíková, Alphen, Van Goch & Weber, 2012; Molinaro, Su, & Carreiras, 2016).

The majority of the previous studies, however, have considered pragmatic context as a pre-existing, static construct. While listeners may come in with pre-existing pragmatic expectations, work on speaker identity formation suggests that these expectations are sensitive to incoming information about individual speakers (e.g. Regel, Coulson, & Gunter, 2010). Yet, very few studies have measured the brain's ability to *adjust* its online response to pragmatic contexts, such as social stereotypes and knowledge about individual speakers. Our study addressed this by using event-related potentials (ERPs) to examine the effects of a particular pragmatic cue, gender stereotype congruity, on listeners' online formation of speaker-specific knowledge during semantic processing.

1.1. Stereotypes, language, and the brain

Of the studies that have used ERPs to investigate the effects of pragmatic context on semantic processing, an illustrative example comes from Van Berkum et al. (2008), who examined the effects of social stereotypes - specifically age, gender, and socioeconomic status - associated with speaker identity. Their study utilized a paradigm where listeners heard well-formed sentences, such as “I cannot sleep without my *teddy bear*” uttered by either a child, who is a stereotypically congruent speaker, or an adult man, who represents a stereotypically incongruent speaker (Van Berkum et al., 2008, p. 582; italics in the original). In addition to these pragmatic incongruities, they tested responses to semantic errors, such as “Dutch trains are *sour* and blue” (Van Berkum et al., 2008, p. 583; italics in the original) relative to semantically correct sentences (“Dutch trains are *yellow* and blue”). The authors found that both pragmatic incongruities and semantic errors elicited an N400 effect (an ERP signature of semantic processing that peaks around 400 ms post-stimulus onset; see Kutas & Federmeier, 2011) and that the effect was larger for semantic errors than pragmatic incongruities. Furthermore, there was a late positivity for sentences violating speaker gender stereotypes, but not for those violating age or socio-economic status. This study showed not only that listeners interpret speech in the context of the speaker, but also that this process happens extremely quickly, providing support for constraint-based (Tanenhaus & Trueswell, 1995), rather than two-step (Grice, 1975), models of speech comprehension.

Subsequent work has built on Van Berkum et al.’s (2008) findings by considering how pragmatic context varies within individuals. Van den Brink et al. (2012) used the same materials as Van Berkum et al. (2008) and found that the difference in effect magnitude between semantic and pragmatic incongruities was due partly to characteristics of the listener. Specifically, analyses of male and female listeners revealed that the N400 effect of pragmatic incongruities was driven by female listeners' responses in the first half of the experiment. In addition to sex, Canal, Garnham, and Oakhill (2015) demonstrated that individual differences in gender biases can also modulate the neurophysiological correlates of language processing. Canal et al. examined pronoun processing following biologically (e.g., *mother*) and stereotypically (e.g., *nurse*) gendered nouns during sentence reading. In response to stereotypically gendered nouns, they found a biphasic ERP response to gender-mismatching pronouns. To better understand the individual variability in the processing of stereotype violations, Canal et al. correlated the ERP responses with scores from the hostile subscale of the Ambivalent Sexism Inventory (ASI; Glick & Fiske, 1996). Those with higher scores (i.e., were more sexist) tended to show a fronto-central positivity, while those with lower scores (i.e., were less sexist) tended to show the biphasic anterior negativity and posterior positivity response that had been observed in the larger group ERP averages. Canal et al. interpreted this pattern to suggest that although the less sexist participants processed the stereotype violation as a categorical agreement violation (P600) they also searched for the less (stereotypically) likely antecedent (Nref). Overall, their findings highlight the importance of accounting for individual differences, including non-linguistic factors such as sexism, when examining gender effects in language comprehension.

While Van den Brink et al.‘s study provides important information about the effects of inter-individual differences on linguistic processing, their analysis, like Van Berkum et al.’s (2008), examined stereotypes across multiple categories (e.g., age, socioeconomic status, sex). Given that Van Berkum et al. recorded a different response pattern to gender stereotype violations compared to other stereotype violations, and that Van den Brink et al. (2012) found sex-specific listener differences in the ERP response to the same stimuli, further investigation of the neurophysiological response to gender stereotypes is needed to better understand the source of these differences. Later work, such as Creel and Tumlin’s (2011) study, considered these questions in the grammatical domain, but our study is the first to simultaneously manipulate pragmatic and semantic information in the investigation of the neurophysiological effects of gender stereotypes.

1.2. Social role theory and stereotype asymmetry

In addition to addressing questions about the mechanics of speech comprehension, previous research investigating the neurophysiological correlates of stereotypical gender processing has had important implications for theories of stereotype processing. For example, social role theory assumes that the behavior of group members shapes their stereotype (Eagly, 1987). Applied to gender, this theory predicts asymmetric effects of stereotypes, as women have begun to take on the roles traditionally held by men, but the change in men's perceived roles has been less dramatic (Diekman & Eagly, 2000; Vandello, Hettinger, Bosson, & Siddiqi, 2013). Siyanova-Chanturia, Pesciarelli, and Cacciari (2012) provided support for this prediction in a study that investigated the effects of stereotypical gender (e.g., most firefighters are men) on pronoun resolution in Italian and German speakers. Using a priming paradigm, they found that masculine pronouns preceded by stereotypically feminine professions (whose gender, critically, was not explicitly marked on the noun; e.g., insegnante (teacher) - lui (he)) elicited an N400, but no N400 was observed for the reverse condition (stereotypically masculine professions preceding feminine pronouns; e.g., conducente (driver) - lei (she)). These results

suggest that female pronouns were more easily integrated into stereotypically masculine contexts than the reverse (i.e., participants were more accepting of female drivers than male teachers), an asymmetric effect of gender stereotypes that is predicted by social role theory (Eagly, 1987).

Although Siyanova-Chanturia et al.'s (2012) results appear to indicate that participants' processing of stereotype incongruent females was facilitated compared to stereotype incongruent males, this pattern is not always consistent in the literature. For example, White, Crites, Taylor, and Corral (2009) investigated the association between gender and stereotypical traits using a priming paradigm. White et al. (2009) used the words "Women" and "Men" as primes and measured ERP responses to traits that were either stereotype consistent or inconsistent (e.g., Women-nurturing, Men-aggressive). White and colleagues found larger N400 amplitudes and slower response times following the stereotype inconsistent traits. Behavioral measures showed faster responses for congruent traits when the prime was Women, whereas participants were quicker to respond to incongruent traits when the prime was Men. Siyanova-Chanturia, Warren, Pesciarelli, and Cacciari (2015) observed a similar pattern of results in a study where participants judged if stereotypically gendered professions, such as headmaster or social worker, could apply to definitionally gendered kinship terms, such as brother or mother. They found that responses were overall faster for congruent pairs (e.g., social worker-mother) as well as gender asymmetries such that male targets received more and faster "yes" responses. Siyanova-Chanturia et al. (2015) suggest that these results may be due to "androcentrism" or the use of the male category as the normative standard (Hegarty & Pratto, 2001). It remains to be determined whether the differing asymmetries across these studies are due to methodological differences (i.e., testing pronouns vs. kinship terms) or because these studies are tapping into competing underlying effects: androcentrism vs. social role.

Despite these remaining questions, the pattern that emerges from the studies described above demonstrates that a) gender stereotypes are automatically activated during comprehension (White et al., 2009), and b) this activation has consequences for semantic (e.g., Van Berkum et al., 2008) and grammatical (e.g., Siyanova-Chanturia et al., 2012) processing. Moreover, the work by Van den Brink et al. (2012) and Canal et al. (2015) indicates that individual differences impact how these stereotypes affect linguistic processing.

1.3. Speaker identity formation and stereotype adaptation

Individual differences are born out of experience (see Koenig & Eagly, 2014), but there is still little research on how experience may change cognitive processing as measured by ERPs. Behavioral studies of social role theory, however, suggest that diverse experiences promote stereotype malleability (e.g., Blair, 2002; Koenig & Eagly, 2014). To our knowledge, no study has examined the neurophysiological correlates of such malleability. While the aforementioned ERP studies manipulated existing stereotypical norms, they did not consider the possibility that listeners adapt to these manipulations by forming new knowledge about the speaker's identity. In fact, very few previous studies have examined the neurophysiological correlates of such speaker identity adaptation over time (Baetens et al., 2011; Regel et al., 2010).

Regel et al. (2010) analyzed ERPs to sentences produced by two 'people' at two different sessions. Although the stimuli were presented visually, they established an ironic and a sincere (literal) person during the first session via the probability of ironic or sincere remarks to previously shown passages. Specifically, the ironic person made ironic remarks 70% of the time at this session, while the sincere person only made ironic remarks 30% of the time. At the second session, participants were again presented with passages from these two people, but the proportion of ironic sentences was equalized for the two people. The authors observed an N400 at the second session when the sincere person produced ironic sentences, but the N400 was smaller when ironic sentences were produced by the ironic person. The consequences of these findings are that a) it appears possible for participants to form a speaker identity based solely on the speaker's output, and within a short period of time, and b) participants utilize this knowledge to adjust their expectations, as would be predicted by social role theory.

Work by Baetens et al. (2011) offers further evidence for the speed of speaker identity formation. Baetens and colleagues had participants read scenarios about unknown actors that implied positive traits (e.g., friendliness), and these scenarios were followed by either a consistent or inconsistent statement (e.g., "Diplaq gave his mother a smack", p. 90). Their electrophysiological data showed that character-inconsistent statements as compared to (baseline) consistent statements elicited an N400 of higher magnitude, followed by a late positive potential (LPP). The authors differentiate these two responses as being due to trait inconsistencies (N400) and evaluative incongruence (LPP). The findings from this study show that participants are capable of forming a speaker identity based not only on the speaker's direct output (similar to Regel et al., 2010) but additionally by indirect information, such as a character description, and that listeners use this knowledge to make predictions about the speaker.

1.4. The current study

Our study connects the work on identity formation by Regel et al. (2010) and Baetens et al. (2011) with the ERP literature on stereotype processing (e.g., Van Berkum et al., 2008) by examining speaker identity formation—that is, dynamic changes in listeners' ERP patterns with increased speaker exposure—in response to stereotype congruent and incongruent identities. Furthermore, our study is the first to simultaneously manipulate pragmatic and semantic information in the investigation of the neurophysiological effects of gender stereotypes specifically. This specificity is motivated by the previous post-hoc findings that gender stereotypes elicit different ERP responses compared to other stereotypes (Van Berkum et al., 2008), and because it allows for the formation of stable speaker-specific identities.

We tested the following questions. First, how do gender stereotypes affect the neurophysiological correlates of semantic processing? We predicted that stereotypical knowledge about speaker gender would affect processing of both semantically correct and

incorrect sentences, as measured by increased N400s to stereotype incongruent compared with stereotype congruent sentences. That is, regardless of whether the sentence is semantically correct (e.g., a male voice saying "The jewelry designer took over fashion week with the ring he had created") or incorrect (a male voice saying "The field goal resulting from the double was contested by the coach") both sentences violate stereotypical expectations of males. More specifically, because men stereotypically care more about sports than fashion, it is incongruent to hear about a double (a baseball term) in an American football context, as well as semantically correct sentences about fashion week. The stereotype congruent condition included the same sentences, but with the critical word adjusted such that both the incorrect (e.g., a male voice saying "The jewelry designer took over fashion week with the shirt he had created") and correct (the male voice saying "The field goal resulting from the kickoff was contested by the coach") sentences fit stereotypical expectations of males. In addition to the N400, our primary component of interest, we were also interested in the LPP, which was observed in some previous studies on stereotype and speaker identity processing (e.g., Baetens et al., 2011; Van Berkum et al., 2008) but not others (e.g., White et al., 2009).

Second, we asked whether speaker gender would influence the effect of stereotype congruence. Based on social role theory and previous work establishing that modern women are viewed as having more masculine characteristics (e.g., Diekman & Eagly, 2000) we hypothesized that effects of stereotype congruence would be stronger to stimuli presented in the male voice.

Third, we asked how increased exposure to stereotype incongruent information would change the observed ERP patterns. We predicted that listeners would adjust their knowledge of the speakers to override the standard stereotype, as indexed by a reduction in N400s to the speakers' stereotype incongruent, but not congruent, utterances in the second half of the experiment relative to the first half. Our prediction that incongruent, but not congruent utterances would show a reduction in N400 magnitude over the course of the experiment is supported by experimental work examining stereotype reduction techniques (e.g., Finnegan, Oakhill & Garnham, 2015). Their research shows that repeated exposure to counter-stereotypical, but not stereotypical, information weakens stereotype biases, which are the presumed generators of N400 effects to stereotype incongruent sentences.

2. Methods

2.1. Participants

All participants were native English speakers with minimal exposure to other languages as assessed by a language background questionnaire. Participants were recruited from the Psychology department subject pool at The Pennsylvania State University and compensated with course credit. All participants provided informed consent before the experiment. Thirty-four participants completed the study; six participants' data were excluded due to excessive EEG artifact (4) or failure to meet study criteria (2). The remaining 28 participants had no reported history of neurological or hearing disabilities and were matched for gender: 14 female and 14 male, with an average age of 19.4 years (SD = 2.14). All participants were right-handed (cf. Grey, Tanner, & Van Hell, 2017), and they reported normal or corrected-to-normal vision. None of the participants reported a history of brain trauma or neuropsychological, learning, or hearing disabilities.

3. Materials

We created 160 sentence pairs that were semantically correct or incorrect by changing one word; half the sentence pairs (80) related to fashion and half (80) to football. See Table 1 for examples and Supplementary Materials for the full set of 160 critical sentence pairs. Correct and incorrect critical words were matched on frequency using the Subtlexus database (Brysbaert & New, 2009), t (318) = 1.2871, $p > .05$, as well as length, t (318) = 1.2587, $p > .05$.

All stimulus sentences were rated by a separate sample of English monolingual undergraduates from the Psychology subject pool (N = 79, 39 female). Norming study participants were given the following instructions:

"Your task is to determine whether each sentence matches more with stereotypes about men or women. The scale ranges from 1 to 5. A score of **1** indicates that the preceding sentence strongly matches stereotypes about women. A score of **2** indicates that the preceding sentence slightly matches stereotypes about women. A score of **3** indicates that the preceding sentence doesn't match stereotypes about either gender. A score of **4** indicates that the preceding sentence slightly matches stereotypes about men. A score of **5** indicates that the preceding sentence strongly matches stereotypes about men."

A repeated measures ANOVA consisting of the within-subjects factors of Topic (Football, Fashion), Correctness (Correct, Incorrect), and a between-subjects factor of Gender revealed main effects of Topic (F (1,77) = 491.161, $p < .001$, $\eta_p^2 = 0.864$) and Gender (F (1,77) = 23653.889, $p < .001$, $\eta_p^2 = 0.997$), as well as a marginal main effect of Correctness (F (1,77) = 3.862, $p < .001$, $\eta_p^2 = 0.048$). These main effects were qualified, however, by interactions between Correctness x Gender (F (1,77) = 49.263, $p < .001$, $\eta_p^2 = 0.390$), Topic x Correctness (F (1,77) = 31.804, $p < .001$, $\eta_p^2 = 0.292$), and Topic x Gender (F (1,77) = 16.778, $p < .001$, $\eta_p^2 = 0.179$). The 3-way interaction was not significant.

We conducted follow-up pairwise analyses of each interaction. For the Correctness × Gender interaction, we observed a difference between how women and men rated incorrect, but not correct, sentences. Specifically, female participants were more likely to rate incorrect sentences as characteristic of men, while male participants were more likely to rate incorrect sentences as more characteristic of women. For the Topic × Gender interaction, we observed that the effect of Topic (i.e., fashion being associated with women and football being associated with men) was stronger in female than male participants. The critical interaction for our study, however, was the Topic × Correctness interaction. We observed that, as expected, incorrect sentences in each topic were more likely

Table 1
Example stimuli in the comprehension task.

Speaker Voice	Male		Female	
Stereotype	Congruent	Incongruent	Congruent	Incongruent
Semantic Incorrect	The jewelry designer took over fashion week with the shirt he had created.	After the punter missed the baseball, the entire team felt defeated.	The kickoff returner ran down the track to catch the ball.	Vogue's shoe specialist wrote an article about a suit in the latest edition.
Semantic Correct	After the punter missed the football, the entire team felt defeated.	Vogue's shoe specialist wrote an article about a boot in the latest edition.	The jewelry designer took over fashion week with the ring he had created.	The kickoff returner ran down the stadium to catch the ball.

Note. ERPs were time-locked to the critical words (underlined).

to be rated as characteristic of its non-stereotypically associated gender. That is, incorrect fashion sentences were rated as more characteristic of men, and incorrect football sentences were rated as more characteristic of women.

Stimulus sentences were produced by a male and a female native speaker of American English. Stimuli were digitally recorded in a sound-attenuated chamber using a SM58 Shure microphone feeding into a PMD670 Marantz recorder. During recording, both speakers produced the stimuli in a randomized order, with each condition (Stereotype Congruent and Incongruent) in both versions (Correct and Incorrect) intermixed. They read the sentences at a natural speech rate. Exported. wav files were then imported into Praat (Boersma & Weenink, 2017), and sentence and target onsets were marked by a native speaker of English. An in-house Praat script segmented the audio into individual stimulus files and equalized each file to 70 db sound intensity. The final versions of the incorrect sentences were composed of the incorrect critical word inserted into the correct sentence using Praat (Boersma & Weenink, 2014), to prevent any anticipatory prosodic effects of the critical word. Stimuli were presented to the participants using E-Prime.

Four experimental lists were created from the stimuli: two Stereotype Congruent lists and two Stereotype Incongruent. Within condition, the two lists had different randomized orders of the stimuli. Each of the four lists comprised 160 critical sentences presented in the male voice, and an additional 80 sentences presented in the female voice. Each participant heard either Stereotype Congruent or Stereotype Incongruent sentences in order to encourage the formation of speaker-specific identities. Examples of the stimulus sentences are provided in Table 1; the full set of sentences is listed in Supplementary Materials. In the Stereotype Congruent condition, 14 participants (7 females, 7 males) heard 80 sports correct and 80 fashion incorrect sentences produced by the male speaker, and 40 sports incorrect and 40 fashion correct sentences produced by the female speaker. In the Stereotype Incongruent condition, 14 participants (7 females, 7 males) heard 80 sports incorrect and 80 fashion correct sentences presented in the male voice, and 40 sports correct and 40 fashion incorrect sentences presented in the female voice. Participants also heard 160 filler sentences (split evenly across speakers), 80 of which were grammatically correct and 80 of which were grammatically incorrect, in order to mask the experimental targets of the study.

3.1. Procedure

All participants were tested in a single 2.5 h-long session. After obtaining consent, participants first completed language history and handedness (Snyder & Harris, 1993) questionnaires, and then began the EEG sentence comprehension task. All sentences were presented bi-aurally using in-ear earphones (Etymotic earphones, model ER4S; Elk Grove Village, IL). Each trial began with a self-timed "Ready?" screen, and the sentence would begin 500 ms following the participants' response. Participants saw a crosshair while they listened to the sentence. Immediately following the end of each sentence, either the "Ready" screen (75% of trials) or a comprehension question unrelated to the critical word (25% of trials) would be presented. Responses to the comprehension questions were collected using an E-Prime serial response box (Schneider, Eschman, & Zuccolotto, 2012). Participants were instructed to stay as still as possible throughout and to blink during the self-timed "Ready?" screen. Participants completed the sentence listening task while sitting in a comfortable chair in a sound-attenuated darkened chamber. After the listening task, participants completed individual difference measures, including the Flanker (Eriksen & Eriksen, 1974) and automated operation-span (Unsworth, Heitz, Schrock, & Engle, 2005) tasks, as well as the Ambivalent Sexism Inventory (ASI; Glick & Fiske, 1996). A subset of 18 participants also completed a post-experiment questionnaire to gather self-ratings of sports and fashion knowledge.

3.2. Data acquisition and analysis

Continuous EEG was recorded from 30 Ag/AgCl active electrodes attached to an elastic cap (Brain Products ActiCap, Germany) in accordance with the extended 10–20 system (Jasper, 1958: Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10). Additional electrodes were placed on each mastoid. Eye movements were monitored with bipolar montages consisting of electrodes placed at the outer canthus of each eye and above and below the left eye. Scalp electrodes were referenced during recording to an electrode placed on the scalp vertex; during offline data processing all scalp electrodes were re-referenced to the algebraic mean of activity over the left and right mastoids. Impedances at all sites were held under 10 kΩ.

EEG was amplified using a Neuroscan SynampsRT system with a 0.05–100 Hz bandpass filter, and digitized with a 500 Hz sampling rate. Following re-referencing, an offline 30 Hz half-amplitude low-pass filter (24 dB/octave roll-off) was applied to the continuous EEG data. ERPs, time-locked to the onset of the critical word, were averaged off-line for each participant at each electrode site in each condition, relative to a 200 ms pre-stimulus baseline using EEGLAB/ERPLAB (Delorme & Makeig, 2004; Lopez-Calderon & Luck, 2014) plug-ins for MATLAB. All artifact-free trials were included in the averages. Trials characterized by eye blinks, excessive muscle artifact, or drift, were not included in the averages. An average of 11.8% of trials was excluded, and this number did not differ reliably across conditions.

ERP components of interest were quantified using mean amplitude measures in a 300–700 ms time window to identify N400 effects. This window was chosen because N400s in response to auditory stimuli may have later onsets or longer durations than N400s to visual stimuli (see e.g., Creel & Tumlin, 2011; Van Berkum et al., 2008). We did not observe evidence of an LPP. For our analysis including the effect of Voice Gender, we conducted two linear mixed-effects models (which account for unbalanced samples, as in our male/female voice conditions), one for midline electrodes and one for lateral regions, that each included random intercepts for subjects using the lme4 package (version 1.1–19) of R (version 3.5.1). Our experimental factors were entered as contrast-coded fixed effects in a 2x2x2x3 (4) factorial design: Voice Gender (female = −0.5, male = 0.5), Stereotype Congruence (congruent = −0.5, incongruent = 0.5) and Semantic Correctness (correct = −0.5, incorrect = 0.5). Electrode/Region was entered as a treatment-coded fixed effect: for Electrode (baseline = Fz; treatment levels = Cz, Pz), and for Region (baseline = Left Anterior; treatment levels = Left Posterior, Right Anterior, Right Posterior). Regions were estimated as an average of the following electrodes: left anterior electrodes (F7, F3, FC1, and FC5), right anterior electrodes (F8, F4, FC2, and FC6), left posterior electrodes (CP5, CP1, P7, P3) and right posterior electrodes (CP6, CP2, P8, P4).

Additional continuous fixed effects were estimated for individual difference variables. Participants' scores on the ASI and its benevolent subscale were allowed to interact with the Stereotype Congruence and Voice Gender experimental factors, and additional fixed effects of Flanker and OSpan performance were added but not allowed to interact with our experimental effects. Random effects were limited to random intercepts per participant given that a) we estimated condition level averages as our dependent variable and b) the majority of our experimental factors have only two levels, which is not optimal for random slope estimation (Bolker, 2012). Models were fit using a restricted maximum likelihood estimation technique. A fixed effect was considered significant if the absolute value of the *t*-statistic was greater than or equal to 2.0 (Linck & Cunnings, 2015) and any *p*-values reported were estimated using sjPlot's tab_model function (version 2.6.1).

For our analysis including the effect of experimental half, we conducted two linear mixed-effects models, using the same procedures as the voice gender analysis. The differences were as follows. First, this analysis was conducted using the male voice data only, due to the reduced number of trials in the female voice condition. Second, the contrast-coded fixed effects included Experimental Half (first = −0.5, second = 0.5) as well as Stereotype Congruence (congruent = −0.5, incongruent = 0.5), and Semantic Correctness (correct = −0.5, incorrect = 0.5). Finally, participants' scores on the ASI and its benevolent subscale were allowed to interact with the Stereotype Congruence and Experimental Half factors.

4. Results

4.1. Behavioral

Descriptive statistics from the behavioral data are provided in Table 2. For further information on the comprehension questions, post-experiment questionnaire, and sexism inventory, see below.

4.1.1. Comprehension questions and post-experiment questionnaire

Participants completed comprehension questions for 25% of the trials in the comprehension task. Average accuracy on the comprehension questions was 88% (*SD* = 0.33), indicating that participants were paying attention to the task. The post-experiment questionnaire asked participants to identify the errors on a subset of 30 sentences that they had heard during the task. Participants were also highly accurate on this task (*M* = 88%, *SD* = 0.09), indicating that they were sensitive to the experimental manipulations. In addition, the post-experiment questionnaire asked a subset of participants (*n* = 18) to report how knowledgeable they were about sports and fashion. This subset reported moderate knowledge of both topics on average, and there was no significant difference between female and male listeners, although females rated themselves descriptively higher in terms of fashion knowledge (female *M* = 3.55, *SD* = .72; male *M* = 2.94, *SD* = 1.24) and males tended to rate themselves higher for sports knowledge (male *M* = 3.63, *SD* = 0.95; female *M* = 3.25, *SD* = 0.89).

4.1.2. Ambivalent Sexism Inventory (ASI)

An independent samples *t*-test on the scores of the male and female participants revealed that the difference between the two groups was significant. Males reported significantly more overall sexism (male *M* = 2.54, *SD* = 0.52, female *M* = 2.10, *SD* = 0.55; *t* (26) = 2.37, *p* = .03, *d* = 0.882) and benevolent sexism (male *M* = 2.43, *SD* = 0.72, female *M* = 1.85, *SD* = 0.70; *t* (26) = 2.35, *p* = .04, *d* = 0.852). There were no significant group differences on the hostile sexism subscale (male *M* = 2.65, *SD* = 0.63, female *M* = 2.35, *SD* = 0.52; *t* (26) = 1.546, *p* = .13, *d* = 0.606).

4.2. ERP results

Fig. 1 presents the grand mean ERP waveforms for correct and incorrect semantic conditions in the Stereotype Congruent group and Fig. 2 presents the ERPs for these conditions in the Stereotype Incongruent group. Visual inspection of the waveforms in the 300–700 ms time window indicated an extended negativity beginning at approximately 300 ms in response to semantically incorrect sentences, which was larger in the Stereotype Incongruent than the Stereotype Congruent group. We did not observe evidence of an LPP.

4.2.1. Analysis of voice gender

In the linear mixed effect model of the midline electrodes (summarized in Supplementary Table 2) we observed a treatment effect of Electrode, such that responses at Pz were significantly less negative than those at Fz. We also observed a main effect of Semantic Correctness, a Stereotype Congruence x Semantic Correctness interaction, and a Stereotype Congruence x Voice Gender interaction that were all superseded by a 3-way interaction between Stereotype Congruence, Semantic Correctness, and Voice Gender (Estimated Beta Coefficient = 2.37, *CI* = 0.06 to 4.69, $p = .045$). As observed in Fig. 3, when stimuli are congruent with gender stereotypes, we observe the expected effect of correctness (larger negativities to incorrect than correct stimuli) for stimuli presented in the male voice. For stimuli presented in the female voice, however, we observe the opposite pattern: larger negativities to correct than incorrect trials. For incongruent stimuli, the pattern is different, such that we see the expected effect of correctness for stimuli presented in both the male and female voice, although the N400 effect is larger for stimuli presented in the male than the female voice.

In the model of the lateral regions (summarized in Supplementary Table 3) we observe treatment effects of Region (posterior sites bilaterally are less negative than left anterior sites) as well as a main effect of Semantic Correctness. That main effect, however, is superseded by two-way interactions with Stereotype Congruence and Voice Gender. The first interaction, plotted in Fig. 4, shows that the effect of Semantic Correctness is limited to Stereotype Incongruent contexts (Estimated Beta Coefficient = −1.11, *CI* = −2.01 to −0.20, $p = .016$). The second interaction with Voice Gender, plotted in Fig. 5, shows that the effect of Semantic Correctness is limited to stimuli presented in the male voice (Estimated Beta Coefficient = −1.11, *CI* = −2.01 to −0.21, $p = .016$).

There were no significant main effects or interactions with our individual difference factors. However, there was a marginal interaction of Voice Gender and overall ASI score in the lateral analysis (Estimated Beta Coefficient = 0.82, *CI* = −0.01 to 1.65, $p = .052$). As plotted in Fig. 6, ASI scores appear to predict the N400 response to stimuli presented in the female, but not the male, voice.

4.2.2. Analysis of experimental half

In our linear mixed effect model of the midline electrodes (summarized in Supplementary Table 4) we observed a treatment effect of Electrode, such that responses at Pz were significantly less negative than those at Fz. We also observed a main effect of Semantic Correctness, such that Incorrect stimuli elicited larger negativities than Correct stimuli. Additionally, we observed interactions between Experimental Half and both the overall ASI and benevolent ASI scores, as well as a three-way interaction between Stereotype Congruence, Experimental Half, and ASI scores (Estimated Beta Coefficient = 3.89, *CI* = 1.06 to 6.72, $p = .007$). The three-way interaction, plotted in Fig. 7, shows that the N400 response of participants who scored higher on the ASI did not change over the course of the experiment. Participants with lower scores, however, showed significant differences in their response between the two halves, with a general weakening of the N400 response at Half 2 in the Stereotype Congruent condition and strengthening of the N400 response at Half 2 in the Stereotype Incongruent condition. The same pattern of results was present over lateral sites, as summarized in Supplementary Table 5.

5. Discussion

This study used ERPs to examine the effects of stereotype congruity on listeners' online formation of speaker identities during semantic processing. We predicted that knowledge of gender stereotypes would affect processing of semantic errors, as measured by increased N400s to stereotype incongruent compared to stereotype congruent sentences. We also predicted that this difference would be larger in response to the male than female voice. Both of these hypotheses were born out in the data (see e.g., Fig. 3). Finally, we predicted that listeners, particularly in the Stereotype Incongruent group, would be able to adjust their stereotype-based expectations

Table 2
Summary of behavioral data.

Measure	Accuracy *M (SD)*	Reaction Time *M (SD)*
Comprehension[a]	0.88 (.33)	3080 ms (2038)
Error Identification[b]	0.88 (.09)	-
OSpan	0.78 (.09)	2269 ms (281)
Flanker Effect	-	54 ms (24)

[a] During the Listening task.
[b] During the Post-Experiment Questionnaire.

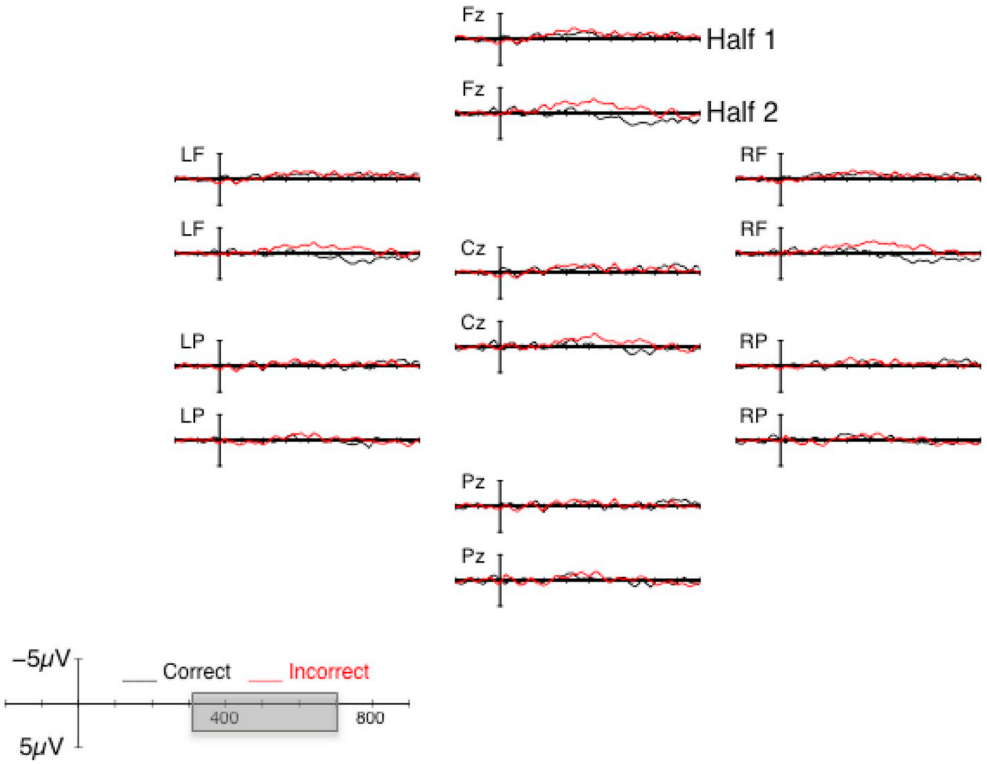

Fig. 1. Effect of Semantic Correctness to the male voice in the Congruent Group. Grand mean ERP waveforms for the Congruent condition for correct semantics (black line) and incorrect semantics (red line). Each tick mark represents 100 ms and the 300–700 ms analysis window is highlighted in gray; negative voltage is plotted up. Calibration scale is ± 5 μV. Grand mean ERP waveforms for each experimental half are plotted separately: the top plot for each electrode shows responses during Half 1, while the bottom plot shows responses during Half 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

of the speakers as indexed by a reduction in N400s to errors in the second half of the experiment. Our findings regarding this point were more nuanced. We observed an adjustment in the N400 response over the course of the experiment only in participants who scored lower on the ASI. Among those participants, we observed a reduction in N400 amplitude only in the Stereotype Congruent group, and in fact observed an increase in N400 amplitude among participants who were exposed to Stereotype Incongruent identities.

5.1. Stereotype congruence affects semantic processing

Returning to our first question, the N400 effect we observed to social stereotype incongruities replicates previous work (e.g., Van Berkum et al., 2008; White et al., 2009). Our observation of a monophasic N400 rather than a biphasic N400-LPP is similar to the results of White et al. (2009). Their study, like ours, focused exclusively on gender stereotype violations. This similarity in findings is notable given that their design used visual presentation of single word primes (e.g., “Women: Nurturing”), as compared to our use of more naturalistic auditory sentence presentation. Other studies that have used a similar auditory paradigm to ours but have manipulated multiple stereotype categories, such as Van Berkum et al. (2008) and Van den Brink et al. (2012), also observed an N400, but the N400 effect in their studies was occasionally followed by a late positivity, the LPP. Specifically, LPPs were observed either in response to gender stereotype violations (Van Berkum et al., 2008) or in the second half of the experiment (Van den Brink et al., 2012). It is possible that we did not observe this positivity because we focused exclusively on gender stereotypes and also presented them from consistent speakers. This may have enabled the participants to form stable speaker identities that elicited less evaluative incongruence and required less of the reflective processing typically associated with the LPP (Baetens et al., 2011; Van den Brink et al., 2012).

It should be noted that although we have characterized our findings as an N400, the distribution of the negativities we observed is more anterior than centro-parietal. Such a distribution is consistent with recent work by Molinaro et al. (2016) and Proverbio, Orlandi, and Bianchi (2017). Both of these studies observed long-lasting anterior negativities in response to stereotype violations. Specifically, Molinaro et al. (2016) suggested that, based on the right-lateralized, anterior distribution and long-lasting nature of their observed effects, social information like stereotypes may be processed differently than semantic knowledge. Proverbio et al. (2017) also observed extended negativities in response to stereotype violations, although their data showed effects over both hemispheres, as do ours. Our results thus add to an emerging literature suggesting that the negativity in response to stereotype violations may be a functionally different component than the N400.

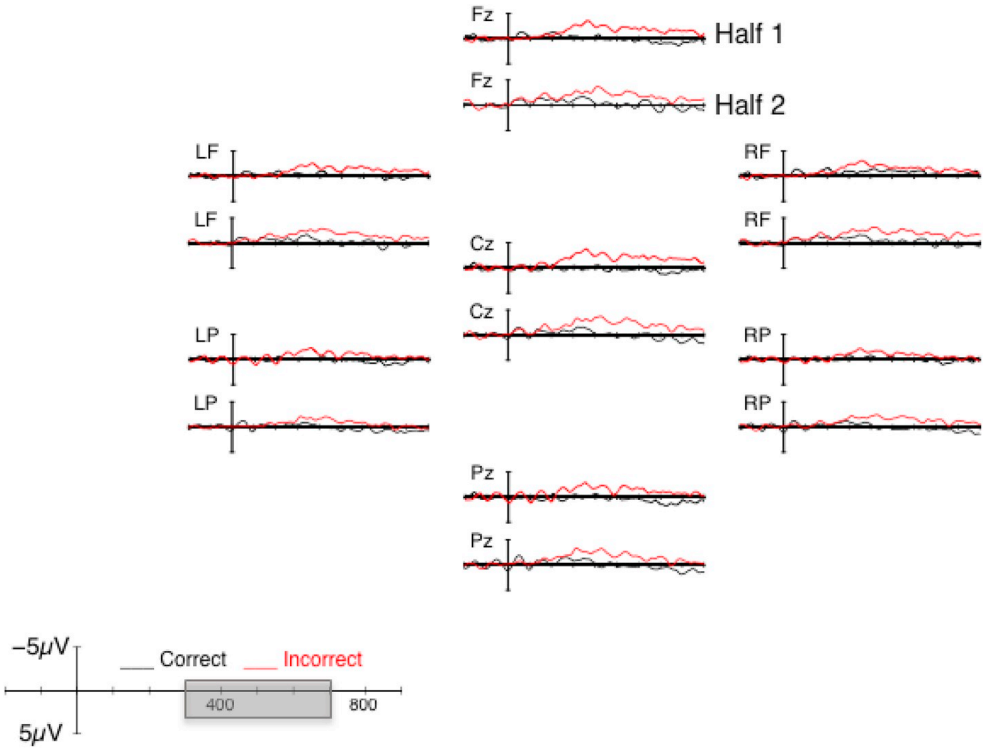

Fig. 2. Effect of Semantic Correctness to the male voice in the Incongruent Group. Grand mean ERP waveforms for the Incongruent condition for correct semantics (black line) and incorrect semantics (red line). Each tick mark represents 100 ms and the 300–700 ms analysis window is highlighted in gray; negative voltage is plotted up. Calibration scale is ± 5 μV. Grand mean ERP waveforms for each experimental half are plotted separately: the top plot for each electrode shows responses during Half 1, while the bottom plot shows responses during Half 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

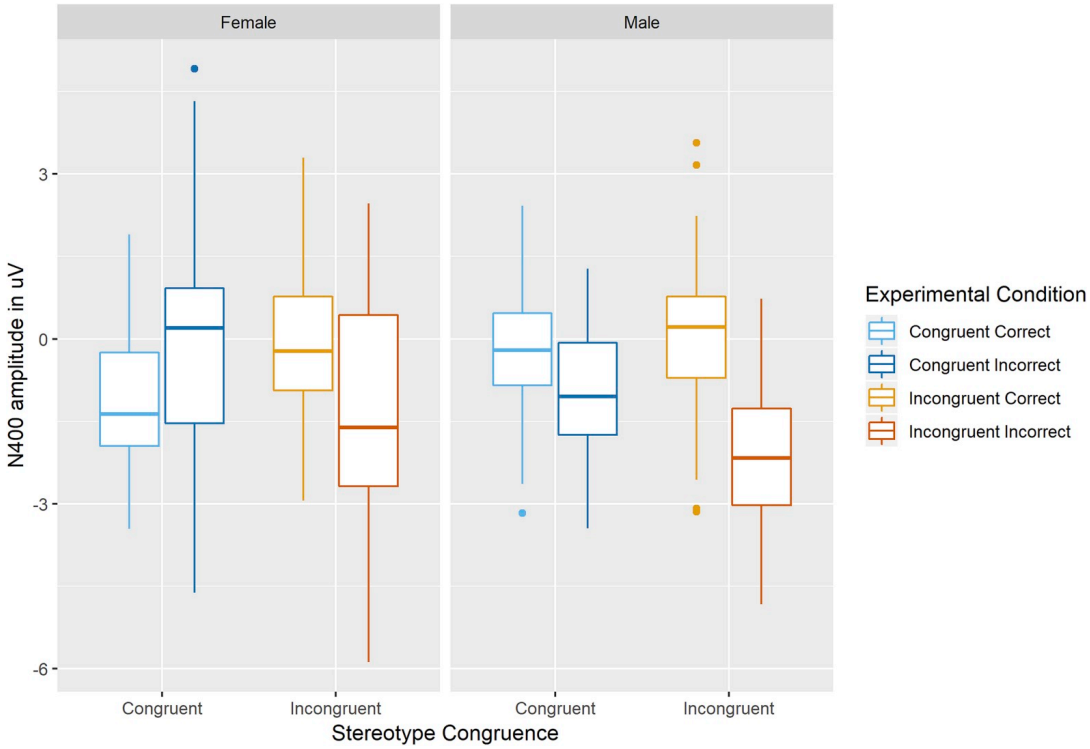

Fig. 3. Three-way interaction of Voice Gender, Stereotype Congruence, and Semantic Correctness present in the midline analysis collapsed across both halves of the experiment. The left panel displays data from the female voice, and the right panel displays data from the male voice. Negative voltage is plotted down.

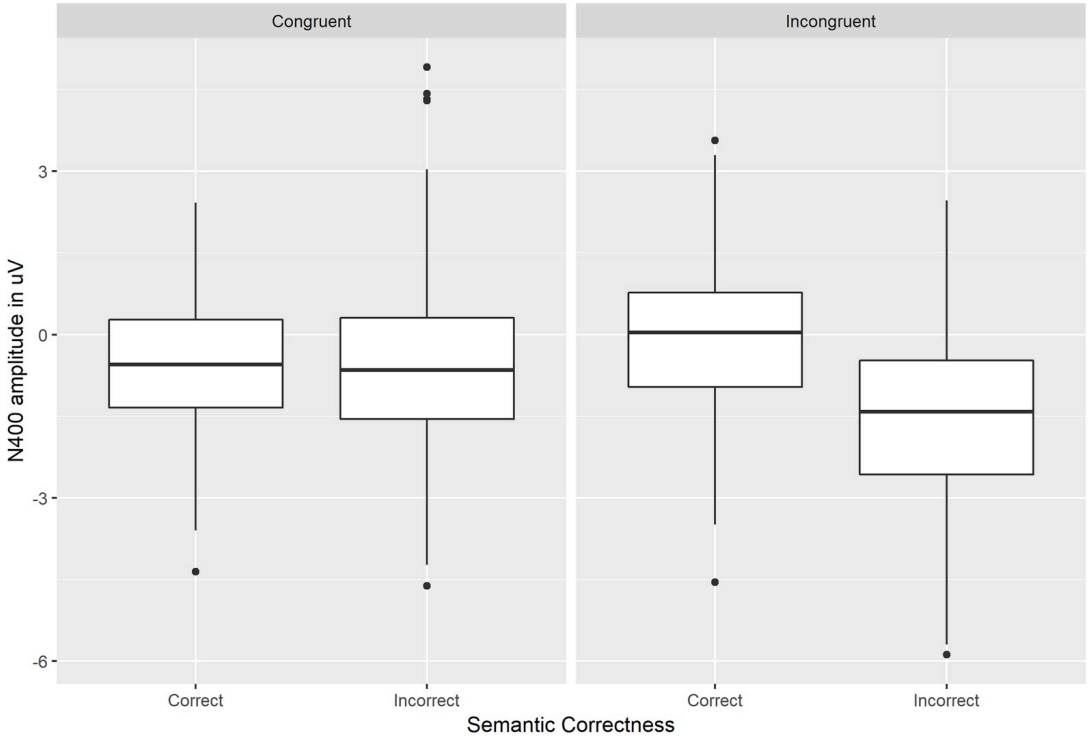

Fig. 4. Two-way interaction of Stereotype Congruence and Semantic Correctness in the lateral analysis collapsed across both halves of the experiment. The left panel displays data from the Stereotype Congruent condition, and the right panel displays data from the Stereotype Incongruent condition. Negative voltage is plotted down.

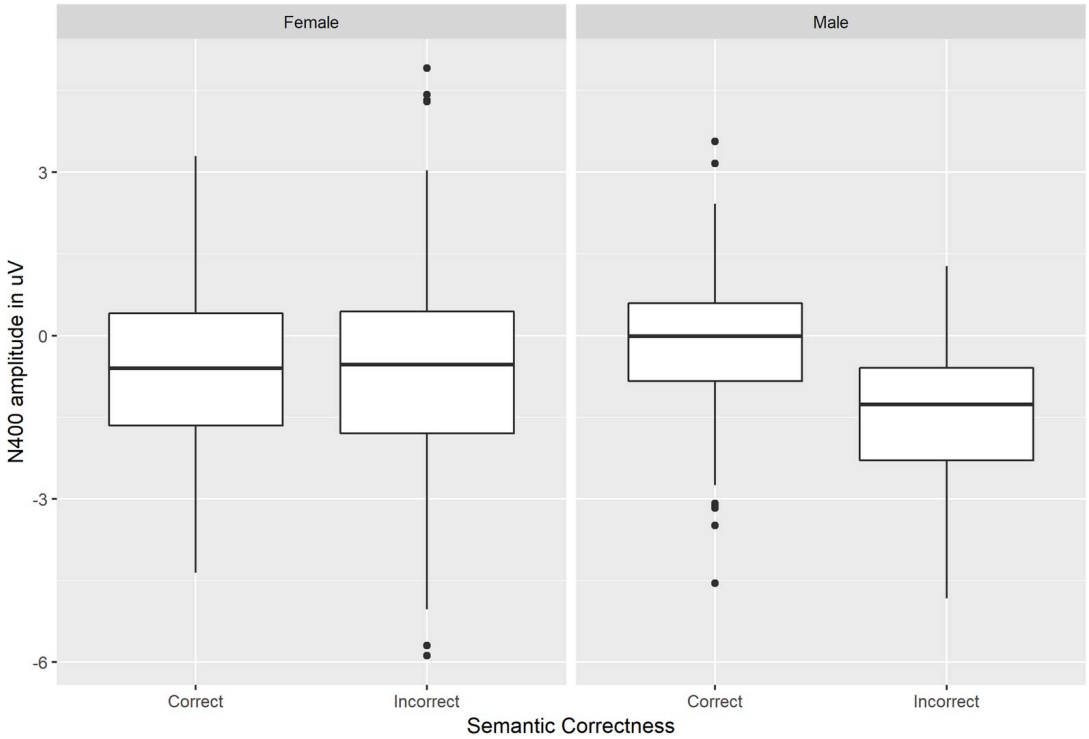

Fig. 5. Two-way interaction of Voice Gender and Semantic Correctness in the lateral analysis collapsed across both halves of the experiment. The left panel displays data from the female voice, and the right panel displays data from the male voice. Negative voltage is plotted down.

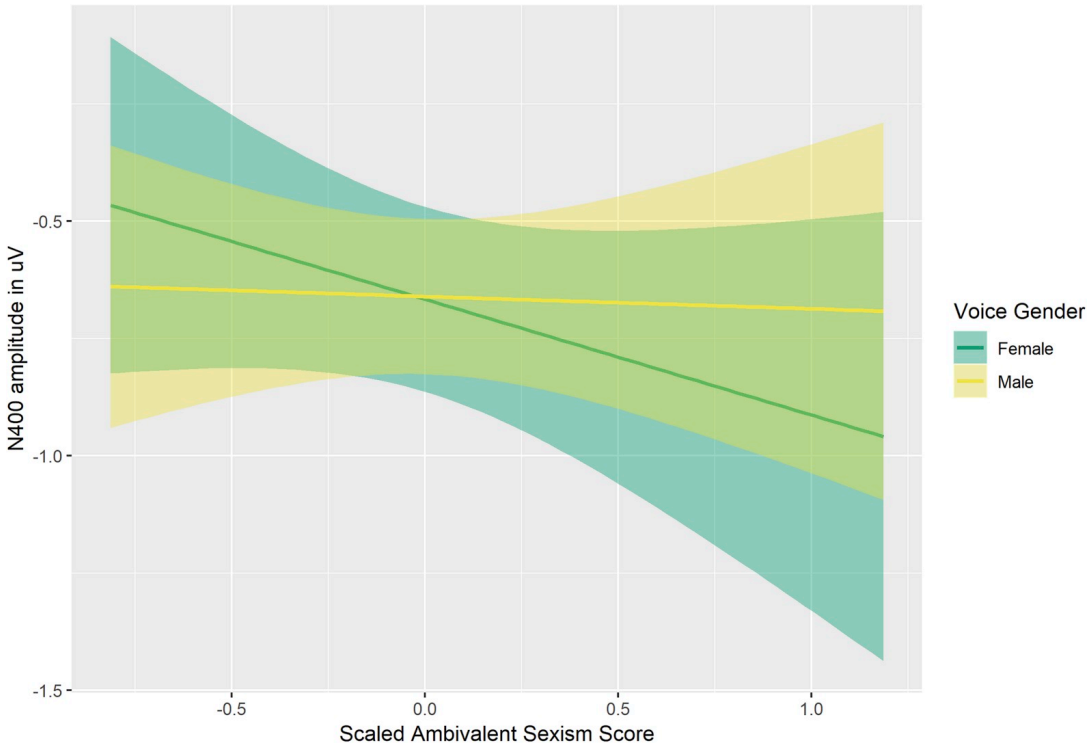

Fig. 6. Smoothed conditional means of the marginal interaction between Voice Gender and ASI score in the lateral analysis collapsed across both halves of the experiment. Shaded areas represent 95% CIs. Higher ASI scores were associated with larger negativities in response to the female voice but not the male voice.

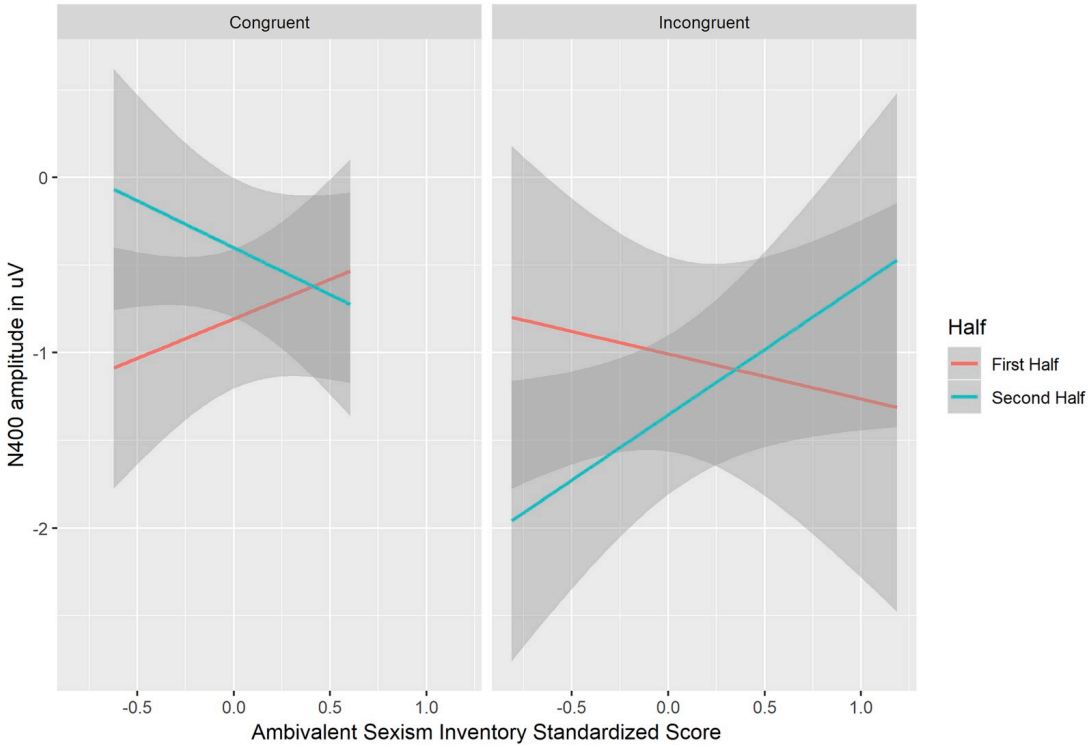

Fig. 7. Smoothed conditional means of the interaction between Stereotype Congruence, Experimental Half and ASI score in the midline analysis. Shaded areas represent 95% CIs. The left panel displays data from the Stereotype Congruent condition, and the right panel displays data from the Stereotype Incongruent condition. Negative voltage is plotted down.

5.2. Changing gender roles affect the influence of stereotypes

In addition to evaluating the influence of stereotype congruence, our second hypothesis questioned how voice gender would affect that influence, given that gender stereotypes have been observed to be more predictive of behavior in males than females (e.g., Vandello et al., 2013). Based on the social role theory proposal that the behavior of group members shapes their stereotypes, we expected the effect of stereotype congruence to be larger in the male voice, and that was indeed what we found. Interestingly, however, we observed not only a weaker effect of stereotype congruence in the female voice, but a reversal of the N400 effect in "Stereotype Congruent" conditions. This means that sentences like "The jewelry designer took over fashion week with the ring he had created." elicited a larger negativity at "ring" than the negativity elicited by the word "track" in the following sentence: "The kickoff returner ran down the track to catch the ball." This reversal could be indicative of changing expectations of women, as investigated by Diekman and Eagly (2000). Participants in their study not only assessed gender-stereotypic characteristics of men and women in the present day, but also for the average man and woman in the past (1950, 1975) and future (2025, 2050). They found that participants generally thought of women in the present day as having more masculine traits than women in the past, and that women in the future would be even more masculine.

We complemented our analysis of voice gender with an examination of change in the N400 over the course of the experiment. We observed that this change was dependent on individual differences in expressed sexism, as well as the stereotype congruence of the stimuli. Participants who scored higher on the ASI were less likely to show a change in their N400 amplitude over the course of the experiment compared to lower-scoring participants. This suggests that when stereotypes are more strongly held, they are more resistant to change. Among participants with lower scores on the ASI, we observed an increase in N400 amplitude over the course of the experiment to Stereotype Incongruent stimuli, and a decrease in amplitude to Stereotype Congruent stimuli. The direction of these effects contrasts with our prediction that increased exposure to the speaker would result in a reduction in N400 amplitude to Stereotype Incongruent stimuli. One explanation for these results is that we were only able to evaluate change over the course of the experiment to stimuli presented in the male voice. Given that stereotypes about males appear to be undergoing less change than stereotypes about females (see Vandello et al., 2013) it is possible that more than approximately 1 h of exposure is needed to change expectations from more masculine to more feminine for a particular speaker. Previous studies that showed speaker adaptation within a single experiment divided the two halves over two days (e.g., Regel et al., 2010) and this 24-h interval between sessions may have allowed for consolidation of the speaker's identity that was not possible in the current study.

In addition to male stereotypes being less in flux, androcentrism may also have affected the change in the N400 to stimuli presented in the male voice. That is, the observed direction of effects wherein continued stereotype congruent information facilitated processing (smaller N400s over time) and stereotype incongruent information hindered processing (larger N400s over time) coincides with the use of males as the "normative standard" (Hegarty & Pratto, 2001; Siyanova-Chanturia et al., 2015). Additional evidence for the androcentric perspective comes from the marginal interaction we observed between ASI-score and Voice Gender, indicating that participants with larger ASI scores showed larger N400s to stimuli presented in the female, but not the male, voice. This pattern of results supports an androcentric perspective because larger N400s are associated with greater difficulty in semantic retrieval/integration, suggesting that participants who endorse more sexist statements show more difficulty in processing stimuli presented in a female voice.

In sum, our data not only replicate the finding that stereotype violations elicit an N400, but also clarify that influence through our focus on gender stereotypes, and extend it by examining the role of voice gender and individual differences in sexism. We interpret our finding that stereotype congruence is a better predictor of responses to male speakers than female speakers as evidence that larger-scale social changes in behavior, such as the expansion of women's roles in the workplace, can moderate gender stereotypes, as predicted by social role theory. Furthermore, we find that individual differences in sexism moderated whether participants' neurophysiological responses to both stereotype congruent and incongruent information would change over the course of a single experimental session.

5.3. Limitations and future directions

The current study opens up several potential avenues for future work. Future replications should include more stimuli presented in a female voice in order to assess change within an experimental session for both genders. Such an analysis would test our hypothesis that it is the relative inflexibility of male gender stereotypes that caused us to observe either no change over the course of the experiment or increased N400s to stereotype incongruent stimuli. Another promising possibility is to examine the role of participant gender, as well as voice gender. Given that our male participants scored significantly higher on the Ambivalent Sexism Inventory overall and its benevolent subscale, and that the Ambivalent Sexism Inventory moderated the effect of stereotype congruence, future work is needed to tease apart the role of participant gender from that of individual differences in sexism. Additionally, an analysis that incorporates participant gender could try to distinguish the relative roles of gender and empathy in determining sensitivity to stereotypes. Both our norming data and Van den Brink et al. (2012) found that female participants were generally more sensitive to stereotype violations, but Van den Brink et al. found that the gender effect was due to individual differences in empathy.

6. Conclusions

The ERP data reported here demonstrate several important findings. First, social stereotypes influence language processing. Listeners' automatic, neurophysiological response to semantic violations (i.e., the N400 response) was larger when those violations

also violated a stereotype. Second, the expectations generated by stereotypes are malleable, and subject to the information listeners have acquired about the current speaker. We observed significant changes in the amplitude of the N400 response over the course of the experiment, and these changes were moderated both by the stereotype congruence of the information presented and by the participants' pre-existing level of endorsement for those stereotypes. Our study is the first to investigate the neurophysiological effects of gender stereotypes specifically in an ecologically valid auditory sentence paradigm, and the results indicate that listeners use both pre-existing stereotypical information and information they acquire online about the speaker during semantic processing.

Declaration of competing interest

None.

Acknowledgements

This work was supported in part by grants from the United States National Science Foundation SMA-1514276/1659920 to Sarah Grey and Janet van Hell, and BCS-1349110, OISE-0968369, and OISE-1545900 to Janet van Hell. We would like to thank Garrett Swan, Leah Pappas, Jennifer Kline, Jasmyn Butryn and Tim Poepsel for their help with stimulus creation, as well as Grace Kim, Ingemarie Donker and Emmanuel Akande for their help testing, and Kaitlyn Litcofsky for her advice throughout the project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jneuroling.2019.100876.

References

Baetens, K., der Cruyssen, L. Van, Achtziger, A., Vandekerckhove, M., & Van Overwalle, F. (2011). N400 and LPP in spontaneous trait inferences. *Brain Research, 1418*, 83–92. https://doi.org/10.1016/j.brainres.2011.08.067.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242–261. https://doi.org/10.1207/S15327957PSPR0603.

Bolker, B. (2012). Mixed model simulations. Retrieved from https://rpubs.com/bbolker/4187.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Canal, P., Garnham, A., & Oakhill, J. (2015). Beyond gender stereotypes in language comprehension: Self sex-role descriptions affect the brain's potentials associated with agreement processing. *Frontiers in Psychology, 6*, 1–17. https://doi.org/10.3389/fpsyg.2015.01953.

Creel, S. C., & Tumlin, M. A. (2011). Online acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*(3), 264–285. https://doi.org/10.1016/j.jml.2011.06.005.

Delaney-Busch, N., & Kuperberg, G. (2013). Friendly drug-dealers and terrifying puppies: Affective primacy can attenuate the N400 effect in emotional discourse contexts. *Cognitive, Affective, & Behavioral Neuroscience, 13*(3), 473–490. https://doi.org/10.3758/s13415-013-0159-5.

Delorme, A., & Makeig, S. (2004). Eeglab: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 134*, 9–21.

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.

Eagly, A. H. (1987). *Sex differences in social behavior: A social role interpretation.* Hillsdale, NJ: Lawrence Erlbaum.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149.

Glick, P., & Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*(3), 491–512.

Grey, S. E., Tanner, D., & Van Hell, J. G. (2017). How right is left? Handedness modulates neural responses during morphosyntactic processing. *Brain Research, 1669*, 27–43.

Grey, S., & Van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics, 42*, 93–108.

Grice, H. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.). *Syntax and semantics* (pp. 41–58). NewYork: Academic Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*, 438–441. https://doi.org/10.1126/science.1095455.

Hanulíková, A., Alphen, P. M. Van, Goch, M. M. Van, & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience, 24*, 878–887.

Hegarty, P., & Pratto, F. (2001). The effects of social category norms and stereotypes on explanations for intergroup differences. *Journal of Personality and Social Psychology, 80*, 723–735. https://doi.org/10.1037/0022-3514.80.5.723.

Jasper, H. H. (1958). The ten–twenty system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371–375.

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology, 107*(3), 371–392. https://doi.org/10.1037/a0037215.

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.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502. https://doi.org/10.1162/jocn_a_00328.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207. https://doi.org/10.1111/lang.12117.

Lopez-Calderon, J., & Luck, S. J. (2014). Erplab: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213. https://doi.org/10.3389/fnhum.2014.00213.

McNally, L. (2013). Semantics and pragmatics. *WIREs Cognitive Science, 4*, 285–297. https://doi.org/10.1002/wcs.1227.

Molinaro, N., Su, J. J., & Carreiras, M. (2016). Stereotypes override grammar: Social knowledge in sentence comprehension. *Brain and Language, 155–156*, 36–43. https://doi.org/10.1016/j.bandl.2016.03.002.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098.

Proverbio, A. M., Orlandi, A., & Bianchi, E. (2017). Electrophysiological markers of prejudice related to sexual gender. *Neuroscience, 358*http://dx.doi.org/10.1016/j.neuroscience.2017.06.028.

Regel, S., Coulson, S., & Gunter, T. C. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research, 1311*, 121–135. https://doi.org/10.1016/j.brainres.2009.10.077.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience, 25*, 762–776. https://doi.org/10.1162/jocn.

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.

Siyanova-Chanturia, A., Warren, P., Pesciarelli, F., & Cacciari, C. (2015). Gender stereotypes across the ages: On-line processing in school-age children, young and older adults. *Frontiers in Psychology, 6*, 1388. https://doi.org/10.3389/fpsyg.2015.01388.

Snyder, P., & Harris, L. J. (1993). Handedness, sex, familial sinistrality effects on spatial tasks. *Cortex, 29*, 115–134.

Tanenhaus, M. K., & Trueswell, C. (1995). Sentence comprehension. In J. L. Miller, & P. D. Eimas (Eds.). *Speech, language, and communication* (pp. 217–262). San Diego, CA: Academic Press.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505. https://doi.org/10.3758/BF03192720.

Van Berkum, J. J. A. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science, 17*, 376–380.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054.

Van den Brink, D., Van Berkum, J. J. A., Bastiaansen, M. C. M., Tesink, C. M. J. Y., Kos, M., Buitelaar, J. K., et al. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive and Affective Neuroscience, 7*(2), 173–183. https://doi.org/10.1093/scan/nsq094.

Vandello, J. A., Hettinger, V. E., Bosson, J. K., & Siddiqi, J. (2013). When equal isn't really equal: The masculine dilemma of seeking work flexibility. *Journal of Social Issues, 69*(2), 303–321. https://doi.org/10.1111/josi.12016.

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.