GLOSSA psycholinguistics

Gardner, B., & Brown-Schmidt, S. (2024). Biased inferences about gender from names. *Glossa Psycholinguistics*, 3(1): 2, pp. 1–40. DOI: https://doi.org/10.5070/G6011185

eScholarship University of California

# Biased inferences about gender from names

**Bethany Gardner,** Department of Psychology and Human Development, Vanderbilt University, US, bethany.gardner@vanderbilt.edu

**Sarah Brown-Schmidt,** Department of Psychology and Human Development, Vanderbilt University, US, sarah.brown-schmidt@vanderbilt.edu

How do alternative forms of reference to individuals – first, last, and full names – guide inferences about the gender of the referent? Given distributional correspondences between English first names and gender, first names provide probabilistic information about an individual's gender. While English last names do not vary with gender, men are more likely to be referred to by last name alone. Across four experiments, we demonstrate that inferences about gender are shaped by a persistent bias to infer that people are male, along with probabilistic information carried by the first name. When an individual was introduced by last name alone, participants overwhelmingly used *he* to subsequently refer to the person, suggesting that participants inferred that the person was male. This bias was still present when the individual was introduced using a first or full name, with participants less likely to use *she* than the distributional characteristics of the first names would predict. When explicitly asked to recall the gender of an individual who was introduced by last name alone, participants preferentially responded that the person was male. This bias persisted even when the person was introduced using a first or full name. Repeated reference attenuated, but did not eliminate, this bias. We discuss implications for models of how world knowledge is linked to language use.

## 1. Introduction

When we talk about people, the way we talk about them can shape or support beliefs and inferences about the person. For example, if I state that "Jane ordered pizza," this asserts some new information about Jane (that Jane ordered pizza), but also supports some inferences about Jane (that Jane is the sort of person that likes pizza). In addition, given that the name *Jane* is probabilistically associated with individuals who are female, the use of this name may lead to the inference that Jane is female. How might this inference about gender change if we introduced Jane using a full name, *Jane Smith*, or just a last name, *Smith*?

Making inferences about gender is rapid and automatic whenever a new person is introduced into a conversation or a story (Duffy & Keir, 2004; Garnham et al., 2002; Kennison & Trofe, 2003; Reynolds et al., 2006), even when gender is not relevant to the current context (Carreiras et al., 1996). Specifically, when a character is introduced using a gender-stereotyped role noun, then referred to later with a pronoun, readers are slower when the pronoun is incongruent with the gender stereotype (e.g., *engineer…she*, *nurse…he*) than when it is congruent (e.g., *engineer…he*, nurse…*she*). This suggests that readers infer the character's gender based on gender cues associated with the role noun, then have to revise this inference based on new information from the pronoun, which incurs a processing cost (Oakhill et al., 2005; Osterhout et al., 1997; Pyykkönen et al., 2010; Sturt, 2003).

When the gender of a referent is unspecified, the person is often assumed to be male (Gastil, 1990; Hamilton, 1991; Moulton et al., 1978; Silveira, 1980). When reading a story about a character with a gender-neutral name who was never referred to with third-person pronouns or other gendered terms, about 75% of participants labelled the character as male (Davis Merritt & Kok, 1995; Davis Merritt & Wells Harrison, 2006). Despite knowing that around half of people are women, it has been argued that comprehenders tend to have a *people = male* bias, where the default person is male, and men belong to the unmarked category (Silveira, 1980).

Similar biases are evident in language production. When participants read short stories that used role nouns (e.g., *After the shop on High Street closed for the night, a baker stayed to tidy up. Before the baker took out the trash…*), then wrote continuations of the story, they were less likely to use *she* to refer to the character than the distributional statistics about the role nouns would predict (Boyce et al., 2019). While participants in a separate norming study estimated that 49% of bakers are women, participants in the sentence completion study used *she* to refer to the baker only around 30% of the time. A third experiment that probed memory for these characters found that participants were less likely to recall characters as female than would be expected given the norming study's estimates of the gender distributions.

In a related study during the 2016 US presidential election cycle (von der Malsburg et al., 2020), one group of participants estimated the likelihoods of Clinton, Trump, and Sanders winning;

a second group completed a sentence about the next US president (*The next US president will be sworn into office in January 2017. After moving into the Oval Office, one of the first things that…*); and a third group completed a self-paced reading task where *she*, *he*, or *they* referred to the next president. Data were collected at multiple intervals during the election cycle. While participants throughout the election cycle estimated a 50–60% chance Clinton would win, participants in the sentence completion task used *she* to refer to the next president only around 10% of the time and *they* around 50% of the time. Participants also showed significant delays when reading sentences that used *she* as compared to *he* and *they* when referring to the next US president. An auxiliary experiment found no general reading time penalty for *she* vs. *he*, indicating that these results were driven by difficulty in interpreting *she* when it co-referred with *the president*, as opposed to *she* being intrinsically slower to process. Similarly, when participants were asked to write about a generic person (e.g., *Before a pedestrian crosses the street…*) and then describe the person they imagined, participants imagined men two times as often as women, but were two and a half times as likely to use masculine names to refer to the characters (Hamilton, 1988). In all three studies, participants used masculine language forms (*he*, strongly masculine names) at higher rates than they inferred the referent to be male. These findings point to a bias in favor of producing masculine language forms, above and beyond a bias to infer gender-unspecified people as male.

Separate evidence suggests that people's estimates of how gender is distributed within different contexts generally reflects real world distributions. Participant estimates of the gender ratios in different occupations were strongly correlated with UK government data (Garnham et al., 2015; Misersky et al., 2014). In the cases where the estimated and actual gender ratios diverged the most, it was in the direction of overestimating the proportion of men. This suggests that it is unlikely that the observed differences between beliefs about and language for role nouns (Boyce et al., 2019; von der Malsburg et al., 2020) are primarily driven by beliefs about role nouns that overestimate the prevalence of women, instead of by a bias towards masculine language forms.

The studies discussed so far have investigated gender inferences from role nouns, which carry probabilistic information about gender through corresponding knowledge of gender distributions in the world (e.g., what proportion of presidents are women). Personal names can also carry probabilistic information about a person's gender. In English, for example, most first names have strong gender associations. Androgynous first names are relatively infrequent in the US, and specific names rarely maintain an androgynous gender association over time (Lieberson et al., 2000). While English last names do not mark gender per se, men are more likely than women to be referred to by last name, particularly in professional contexts (Atir & Ferguson, 2018; Files et al., 2017; Rubin, 1981; Stewart et al., 2003; Takiff et al., 2001; Uscinski & Goren, 2011). Speakers have a range of choices when referring to a person, including pronouns, first names, last names, role nouns, titles, and honorifics. These referential alternatives provide different types

of probabilistic cues to the person's gender that may guide the inferences about gender that a comprehender makes.

In addition to shaping inferences about gender, these referential choices also impact evaluations of the person. For example, when scientists were referred to by last names as opposed to gender-neutral full names, they were subsequently judged as more eminent, famous, and deserving of awards (Atir & Ferguson, 2018). When students evaluated a transcript of a class introduction, professors who were referred to by title were afforded higher status than those referred to by first name (Stewart et al., 2003; Takiff et al., 2001). However, when female professors were referred to by title, they were perceived as less accessible; this double bind between respect and accessibility was not found for male professors (Takiff et al., 2001). One explanation for why referring to people by last name or title affords them higher status is that it makes them, overall, seem more masculine.

The aim of the present research is to examine how alternative ways of referring to people affect inferences about the person's gender. Focusing on the use of a person's name in English, we leverage the fact that first names (e.g., *Jordan*, *Mary*, *Brian*) are probabilistically associated with different genders. The *use* of last names is also probabilistically associated with gender, but indirectly, by virtue of the fact that men are more likely to be referred to by last name than women, particularly in professional settings (Atir & Ferguson, 2018; Files et al., 2017; Rubin, 1981; Stewart et al., 2003; Takiff et al., 2001; Uscinski & Goren, 2011). More generally, there is a people = male bias, where people are assumed to be male by default (Gastil, 1990; Hamilton, 1991; Moulton et al., 1978; Silveira, 1980). Here, we contrast two hypotheses about how these three different sources of information shape inferences about gender.

One hypothesis is that the people = male bias (Gastil, 1990; Hamilton, 1991; Moulton et al., 1978; Silveira, 1980) is present only when the referential form itself does not provide direct, probabilistic information about gender. If so, a person introduced by last name would be much more likely to be assumed to be male than female, due to both the people = male bias and the fact that that people referred to by last name tend to be male. In contrast, when a person is introduced with a first name, probabilistic information carried by the gender distribution of that name would guide gender inferences, instead of the people = male assumption. While this pattern would differ from findings for role nouns (Boyce et al., 2019; von der Malsburg et al., 2020), such a pattern of findings may be expected, given that gender associations for English first names cluster at the endpoints (Lieberson et al., 2000) more than gender associations for job-related role nouns (Garnham et al., 2015; Misersky et al., 2014). If, on average, first names carry stronger gender cues than role nouns, people may form inferences based primarily on names, without defaulting to the "people = male" assumption.

Alternatively, if the people = male bias persists in the face of probabilistic information about gender, inferences about gender should result from a combination of these cues. On this hypothesis,

we would not expect introducing people with a first name to eliminate the *people=male* bias. Instead, a person would be less likely to be inferred to be female than predicted by the gender distribution of the first names. A series of four experiments tests these competing hypotheses in a paradigm where participants were introduced to characters using first, last, or full names. We consider two measures of gender inferences about the characters: use of gendered third-person pronouns to refer to the characters (Experiments 1 and 3) and explicit questions about the gender of the characters (Experiments 2 and 4).

## 2. Experiment 1

The aim of Experiment 1 was to examine the relationship between how a character in a sentence is introduced – by their first, last, or full name – and inferences about that character's gender. Participants read sentences that introduced a character by name (e.g., Jordan, Smith, or Jordan Smith) and continued with fragments that invited continuation with a pronoun referring to the character. Participants wrote completions for the fragments, and we used the gender information that was (or was not) carried on the pronoun as a measure of the participants' inferences about that character's gender.

### 2.1 Methods

#### 2.1.1 Participants

457 participants were included in the dataset, with each participant assigned to 1 of 3 between-participants conditions (First = 152, Full = 153, Last = 152). The sample size was selected a priori, based on Boyce et al. (2019). Participants were recruited on Amazon Mechanical Turk and required to be over the age of 18, located in the US, and have started learning English before the age of 5; they were paid $1.50 for a task that took approximately 10 minutes. A total of 570 participant responses were collected, and exclusion rationales and participant demographics are reported in Supplement §2.1.[1] Critically, participants who guessed that the study was about gender bias were excluded.

#### 2.1.2 Norming study

In order to select a set of first names that range from feminine to androgynous to masculine, we first conducted a norming study on a set of 90 names. 30 masculine and 30 feminine names were selected from lists of the most common names for assigned male at birth (AMAB) and assigned female at birth (AFAB)[2] babies in the US (USSSA, 2019). An additional 30 androgynous

[1] https://github.com/bethanyhgardner/gender-bias-names/blob/main/supplement.pdf.

[2] We use assigned male at birth (AMAB) and assigned female at birth (AFAB) to indicate that these datasets only have information about what sex children were assigned at birth, not their gender identities later. For more information about current best practices for talking about gender, see GLAAD (2020).

names were selected from a list of names that were given at least one third of the time to AFAB children in the US and also at least one third of the time to AMAB children (Flowers, 2015). 50 participants on Amazon Mechanical Turk, following the same inclusion criteria as Experiment 1, were asked to rate the 90 names on a scale of 1–7, with 1 being "definitely masculine" and 7 being "definitely feminine." From these results, we selected 21 names to represent a range of ratings from masculine to feminine, with different levels of androgyny in between. The 21 names were not perfectly centered ($M = 4.19$), partially due to the fact that androgynous names that lean masculine are much more frequent than androgynous names that lean feminine (Lieberson et al., 2000). The norming data were compared to US census data from 1930–2015 (USSSA, 2020). The proportion given to AFAB children in the census data and gender rating from the norming data showed a very strong positive correlation, $r(19) = .92$, $p < .001$, and differences between the measures did not consistently over- or underestimate the femininity of the names (Supplement §1).

### 2.1.3 Materials and procedure

We created 21 prompts that introduced a human character by name and ended with a fragment that was easiest to continue with a pronoun referring to the character. The prompts did not include gendered pronouns, other names, or additional human characters. 3 between-participant conditions manipulated the form of name used to refer to the character: *First Name* (e.g., *Jordan woke up early to walk the dog. After making coffee*), *Last Name* (e.g., *Smith woke up…*), and *Full Name* (e.g., *Jordan Smith woke up…*). The participants' task was to read a sentence using one of these names and then write a completion to the continuing fragment. We measured which pronouns, if any, were used to refer to the character as a measure of the participant's inference about the character's gender.

The combinations of names and prompts were counterbalanced by creating 3 lists for each condition; each participant was randomly assigned to 1 of the resulting 9 lists. Each list for the First Name condition included the 21 first names selected from the norming study and all 21 prompts, and the 3 lists counterbalanced which names went with which prompts. Each list for the Last Name condition included 21 last names and all 21 prompts, and again the 3 lists counterbalanced which names went with which prompts. The last names were selected from a list of the most common surnames in the US (US Census Bureau, 2016), avoiding last names that are also commonly used as first names (Supplement §1). Each of the 3 lists for the Full Name condition included 21 full names and all 21 prompts, and the 3 lists counterbalanced which names went with which prompts, as well as the combinations of first and last names. Each Full Name list had a different combination of first and last names; however, due to experimenter error, there was 1 combination that appeared in 2 lists (this item was included in the analysis) and 1 first name missing from 1 list in the Full Name condition. This resulted in 104 name

combinations. In addition to the critical stimuli, each participant saw 8 filler items using the names of 26 US presidential 2020 candidates in May 2019. These fillers (8–9 per list) served two purposes: first, they were used as a distraction from the focus of the study. Second, they were used to pilot items for an unrelated study about forms of reference in political language. After completing the production task, participants were asked for basic demographic information: gender, age, race/ethnicity, and education level. The participant gender question was written as an open-ended response, following best practices for trans-inclusive study design (Cameron & Stinson, 2019; NASEM, 2022; Vincent, 2018).

## 2.2 Predictions

If bias in gender inferences is limited to cases when no other direct information about gender is provided, we would expect people to use probabilistic gender information to infer gender when this information is available (First and Full Name conditions), resulting in rates of *she* responses that match the gender distributions of the first names. A bias to infer characters as male would only appear when a direct cue to gender is not provided, resulting in a bias towards *he* responses only in the Last Name condition.

Alternatively, if the bias to assume people are male occurs even when probabilistic cues to gender are available, we would expect characters to be more likely to be inferred as male than female in all three conditions, with combined effects of the first name's gender associations and a bias in gender inferences in the First and Full Name conditions. If this is the case, while *she* responses will increase as the first names become more feminine, the rate of *she* responses will be lower than predicted by the gender associations of the first names.

A secondary question was whether introducing a character with the full name (e.g., Jordan Smith) rather than the first name only (e.g., Jordan) would attenuate the influence of the gender information carried by the first name. This question was motivated by the observation that, in English, it is more common to refer to men than women by their last names (Files et al., 2017; Rubin, 1981; Stewart et al., 2003; Takiff et al., 2001; Uscinski & Goren, 2011), and thus the full name may act as a cue to masculinity.

## 2.3 Results

Responses that used *he/him/his* pronouns to refer to the named character were categorized together and will be referred to as *he* responses. Likewise, responses that used *she/her/hers* pronouns to refer to the named character are categorized as *she* responses. Responses that did not use a gendered pronoun were coded as *other*; these responses most commonly repeated the character's name (e.g., *After making coffee...Jordan sat down*), but also included responses with no grammatical subject (e.g., *...sat down*) and responses not referring to the

character (e.g., *...it started raining*). Uses of singular *they* were infrequent (Supplement §2.3). For the First Name and Full Name conditions, the rates of *he* and *she* responses were roughly equal, following the balanced distribution of first names in our stimuli (**Table 1**). In the Last Name condition, responses overwhelmingly biased towards *he*. Notably, in the Last Name condition, *she* responses were slightly less common than other responses that did not gender the character.

**Table 1:** Experiment 1: Number of *she*, *he*, and *other* responses and the ratio of *she* responses to *he* and *other* responses for each condition.

| **Experiment 1: Number of Responses by Condition** | | | | |
|---|---|---|---|---|
| | ***She*** | ***He*** | ***Other*** | **Ratio of *She* \| *He* + *Other*** |
| **First** | 1395 | 1572 | 225 | 0.776 |
| **Full** | 1535 | 1514 | 131 | 0.933 |
| **Last** | 251 | 2616 | 325 | 0.085 |

Responses were analyzed using logistic mixed-effect regression models with *lme4* in R (Bates et al., 2015), predicting the log odds of *she* responses (coded as 1) as opposed to *he* and *other* responses (coded as 0). *Other* responses were coded as 0 because they were not frequent enough to be placed in a third category. Participant and item were included as random intercepts, with items defined as the unique first, last, and first + last name combinations. Treating the names as the random items meant that the condition manipulations were fully between-participant and between-item, so fitting a random slope model was not possible. The fixed effect of Condition was coded with orthogonal Helmert contrasts, with the first contrast comparing the Last Name condition to the First and Full Name conditions, and the second contrast comparing the First Name condition to the Full Name condition. All models are reported with Bonferroni corrections for multiple comparisons. Overall, participants were less likely to respond *she* than *he* and *other* ($\beta = -1.43$, $z = -4.65$, $p < .001$). Participants in the First and Full Name conditions were more likely to produce *she* than participants in the Last Name condition ($\beta = 2.82$, $z = 4.03$, $p < .001$). The comparison between First and Full Name conditions was not significant (**Table 2**).

The second model included each first name's normed Gender Rating as a covariate (**Table 3**). This analysis included the First and Full Name conditions only, as the Last Name condition did not contain first names. Condition was coded with mean-centered contrasts, comparing the First and Full Name conditions. The Gender Rating for each first name was mean centered, with positive numbers more feminine and negative numbers more masculine.

**Table 2:** Experiment 1: Model results for the effect of Condition on the likelihood of *she* responses (= 1) as opposed to *he* and *other* responses (= 0).

| **Experiment 1: Condition** | | | | |
|---|---|---|---|---|
| | ***Refer to using* she** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –1.428 | 0.308 | –4.644 | **<0.001** |
| **Condition: Last** (–.66) **vs. First** (+.33) + **Full** (+.33) | 2.824 | 0.702 | 4.026 | **<0.001** |
| Condition: First (–.5) vs. Full (+.5) | 0.620 | 0.700 | 0.886 | 0.376 |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 1.029 | | | |
| $\tau_{00\ Item}$ | 7.234 | | | |
| $N_{Participant}$ | 457 | | | |
| $N_{Item}$ | 104 | | | |
| Observations | 9564 | | | |

†Bonferroni corrected $\alpha = .0167$.

**Figure 1** shows the proportions of *he*, *she*, and *other* responses for the First and Full Name conditions by the Gender Rating of the first name. As the rating of the name became more feminine, *she* responses increased ($\beta = 1.59$, $z = 21.97$, $p < .001$). However, inspection of the data shows that the increase in *she* responses was not symmetrical to the increase in *he* responses. In the mostly-feminine range of first names (1 to 2 on the X-axis), *he* responses outnumbered *other* responses. In the mostly-masculine range (–3 to –2), *she* responses occurred at similar rates as *other* responses. Particularly in the First Name condition, *she* responses did not surpass *he* responses until the first name in the prompt was biased somewhat feminine, rather than at the midpoint on the scale. With mean-centered fixed effects, the significant intercept term ($\beta = –0.51$, $z = –4.28$, $p < .001$) reflects overall fewer *she* than *he* and *other* responses, and the effect of Condition ($\beta = 0.53$, $z = 2.22$, $p < .05$) reflects more *she* responses in the Full Name than the First Name condition. The interaction between Condition and Gender Rating was not significant, indicating that the effect of Gender Rating was of a similar magnitude in the First and Full Name conditions.[3]

[3] A reviewer pointed out that the mean for the 21 first names ($M = 4.21$) is higher than the center of the scale (= 4). An alternative way of conducting this analysis would be to center Gender Rating at 4, the mean of the response scale (dashed line in Figure 1), instead of at the item mean (0 in in Figure 1). This alternative analysis yielded the same pattern of results in this and subsequent experiments, though the absolute value of the intercept was consistently larger (see details in Supplement §2.2).

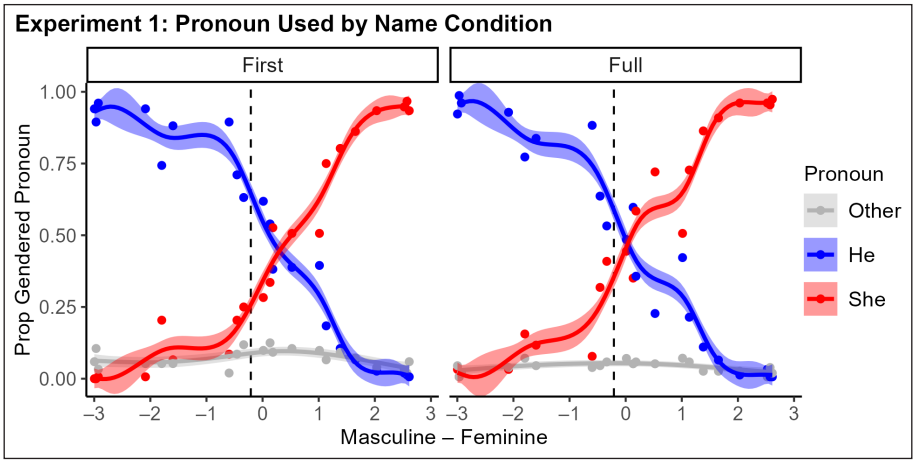


**Figure 1:** Experiment 1: Proportions of *he* (blue), *she* (red), and *other* (gray) responses in the First and Full Name conditions by the mean-centered gender rating of the first name. Points indicate means for each of the 21 first names, and solid lines indicate a smooth function on the raw data. On the x-axis, 0 is the mean of the 21 names, and the dashed line indicates the center of the original response scale in the norming study.

**Table 3:** Experiment 1: Model results for the effects of Condition and Gender Rating on the likelihood of *she* responses (= 1) as opposed to *he* and *other* responses (= 0) in the First and Full Name conditions.

| **Experiment 1: Condition and Gender Rating** | | | | |
|---|---|---|---|---|
| | ***Refer to using* she** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –0.513 | 0.120 | –4.282 | **<0.001** |
| Condition (First = –.5, Full = +.5) | 0.532 | 0.240 | 2.218 | 0.027† |
| **Gender Rating** (Centered, Masc –, Fem +) | 1.593 | 0.073 | 21.967 | **<0.001** |
| Condition × Gender Rating | –0.175 | 0.139 | –1.257 | 0.209 |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 0.889 | | | |
| $\tau_{00\ Item}$ | 0.501 | | | |
| $N_{Participant}$ | 305 | | | |
| $N_{Item}$ | 83 | | | |
| Observations | 6372 | | | |

†Bonferroni corrected α = .0125.

An additional analysis examined if including *other* responses with *he* responses impacted our findings. The results revealed the same patterns as above, with the following exceptions: In the model testing the effects of Condition and Gender Rating, the intercept ($\beta$ = –0.22, $z$ = –1.75, p = .08) and the difference between the First and Full Name conditions ($\beta$ = –0.25, $z$ = –1.57, p = .12) were not significant. A second exploratory analysis added a quadratic effect of Gender Rating to evaluate whether the increase in *she* responses as a function of Gender Rating was nonlinear. For example, we might expect to see an effect of Gender Rating that is weaker at the endpoints (strongly gendered names) than at the midpoint (androgynous names), with a larger *he* response bias for androgynous names than for strongly gendered names. The quadratic effect of Gender Rating was not significant, nor did it significantly interact with Condition, inconsistent with this possibility. A final exploratory analysis included participant gender as a covariate, testing if male participants showed a larger *he* response bias; this analysis revealed no significant effects after Bonferroni corrections for multiple comparisons. These three analyses are discussed in more detail in Supplement §2.3–2.5.

## 2.4 Discussion

We investigated whether the form of reference—first name, last name, or full name—affected people's inferences about a character's gender, measured through the pronouns they used to complete a sentence referring to the character. When participants were not given explicit cues to gender (Last Name condition), participants overwhelmingly used *he* to refer to the character. Moreover, in the Last Name condition, participants were approximately equally likely to not use a gendered pronoun at all (*other* responses) as they were to use *she*. Although probabilistic cues to the referent's gender did shape inferences, with more *she* responses when a first name was given (First and Full Name conditions), inspection of the data indicates that the bias towards *he* responses persisted. A character's name needed to be more strongly feminine for participants to preferentially refer to them with *she*. In addition, participants showed a pattern of asymmetry for mostly-masculine and mostly-feminine names. In the First and Full Name conditions, androgynous names that leaned feminine (e.g., *Jackie*) still elicited *he* responses, but androgynous names that leaned masculine (e.g., *Chris*) elicited *other* responses, rather than *she* responses. As the first names became more feminine, the rate of *she* responses remained flat and parallel to the rate of *other* responses, then increased more sharply, whereas the rate of *he* responses decreased more gradually. We also hypothesized that introducing a person with a first and last name would attenuate the gender cue from the first name, such that the preference for *he* responses would be greater in the Full Name condition as compared to the First Name condition. The data were not consistent with this prediction; instead, the preference for *he* responses was numerically larger in the First Name condition.

The fact that the gender ratings of the first names predicted *she* responses indicates that participants were willing to produce feminine language forms in this task. The observed bias towards masculine language forms instead points to biased inferences about gender. However, one potential concern with this interpretation is that the pronouns produced in reference to the characters may not entirely match participants' underlying inferences about the characters' genders. Instead, it is possible that some *he* responses in the Last Name condition come from generic masculine usage, with participants producing *he* in an ostensibly gender-unspecified manner. In the 19th century, the generic masculine was prescribed as correct, explicitly replacing alternatives like singular *they* and *he or she* that had been in use. This was contested by feminists in the 1970s and 1980s, who argued that this language was not inclusive and perpetuated biases of masculine as the default (Bodine, 1975). While this guidance has been replaced in formal language policies by *he or she* and occasionally singular *they* constructions (APA, 2019; APA Publication Manual Task Force, 1997; Robertson, 2021), some speakers retain the generic masculine usage. If so, some instances of *he* responses in the data – particularly those in the Last Name condition, where no direct information about the character's gender is included – may reflect this generic use. To provide a more direct test of the influence of referential form on gender inferences per se, in Experiment 2 we ask participants to make explicit inferences about the referent's gender.

# 3. Experiment 2

The aim of Experiment 2 was to examine the relationship between how a character in a story is referenced (e.g., by their first, last, or full name) and later explicit judgments about that character's gender. Participants read a series of seven short stories that introduced a human character with a name and described them completing an everyday action. After a brief delay task, participants were prompted with each character's action and asked to indicate the character's gender in a free-response box. Note that participants were only asked explicit questions about gender after having read all seven stories first, in contrast to Experiment 1, where participants generated a sentence completion after reading each story preamble. This design choice was used to avoid participants reading the stories with the expectation that they would be later asked about gender.

## 3.1 Methods

### 3.1.1 Participants

1351 participants were included in the dataset, with each participant assigned to 1 of 3 conditions (First = 451, Last = 448, Full = 452). The sample size was determined a priori, based on Boyce et al. (2019). Participants were recruited on Amazon Mechanical Turk using the same inclusion criteria and payment as in Experiment 1. A total of 1534 responses were collected; exclusion

rationales and participant demographics are reported in Supplement §3.1. Unlike in Experiment 1, participants were not excluded for guessing the study was about gender bias, since this task explicitly asked about gender inferences.

### 3.1.2 Materials and procedure

The names were combined into 3 between-participants conditions as in Experiment 1 (*First Name, Last Name, Full Name*). Participants saw two-sentence stories that referred to a character by name twice and did not contain any gendered pronouns (**Figure 2**). The stories described everyday actions selected to avoid strong gender stereotypes (e.g., making coffee, walking a dog). Because the task involved a memory component, participants completed 7 critical trials (as opposed to 21 in Experiment 1). The materials included a total of 7 stories, 21 first names, 21 last names, and 63 first + last name pairs. Within each of the 3 conditions, 9 lists counterbalanced which names were included (3 sets) and the combinations of names and stories (also 3 sets); participants were randomly assigned to 1 of these 27 lists. In the First Name condition, each list included 7 out of the 21 first names, distributed evenly across the gender ratings from masculine to feminine. In the Last Name condition, each list included 7 out of the 21 last names, distributed randomly. In the Full Name condition, we used the same 3 combinations of first and last names as in Experiment 1, with the exception that we corrected an error in the Experiment 1 lists where a duplicate name appeared. As in Experiment 1, the names of 26 US presidential 2020 candidates acted as filler items to pilot a separate study, with each participant seeing 1 of these items.

After reading each story, participants typed the name of the character as an attention check. Participants then completed 16 simple math questions as a distraction task. Next, participants were given a summary of the main action in each story and asked to type the gender of the character into a free response box (**Figure 2**). The free response box allowed participants to express uncertainty (e.g., *gender wasn't specified* or *I can't remember*). Critically, the memory prompt referenced the action and not the name.

## 3.2 Predictions

The results of Experiment 1 indicate that the bias to infer characters as male was not eliminated when probabilistic information about the gender of a character – given by their first name – was provided. Instead, the findings support a model of gender inference where probabilistic cues about gender are combined with a people = male bias. However, pronouns produced in reference to the character may not necessarily reflect underlying inferences about the character's gender. In particular, some *he* responses in the Last Name condition may have been driven by a generic masculine usage, where participants produced *he* but did not necessarily infer the character as male.

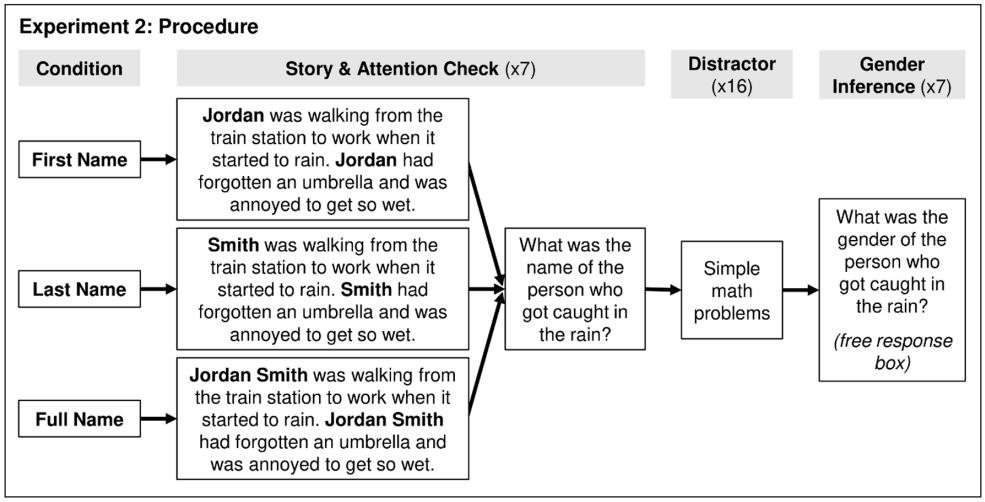


**Figure 2:** Experiment 2: Procedure and example stimuli.

If the results of Experiment 1, where participants were less likely to produce *she* than predicted by the gender association of the first names, do reflect a bias to infer characters as male, the pattern of results should be the same when participants are asked directly about the characters' genders. Characters will be more likely to be recalled as female as the rating of the first name becomes more feminine (First and Full Name conditions). However, a bias to infer characters as male will be present in all three conditions and strongest when direct probabilistic information about gender is not provided (Last Name condition).

Alternatively, the bias to use the pronoun *he* when describing the characters in Experiment 1, particularly in the Last Name condition, may not have reflected a bias in underlying gender inferences about the character, and instead reflected the use of the generic masculine (Bodine, 1975). If so, in Experiment 2, we would expect to observe no bias to recall characters as male in the First and Full Name conditions, where probabilistic cues to gender are available. In the Last Name condition, we would expect characters to be more likely to be recalled as male than as female, due to the tendency to refer to men using last names more often than women.

## 3.3 Results

Responses were coded as *male* (e.g., "m," "man," "male"), *female* (e.g., "f," "woman," "female"), or *other* (e.g., "It wasn't specified," "I don't remember"). As in Experiment 1, the rates of *male* and *female* responses were roughly equal in the First and Full Name conditions, following the balanced distribution of the first names, but participants overwhelmingly responded *male* in the Last Name condition (**Table 4**). Overall, participants were less likely to respond *female* than *male* or *other* ($\beta$ = –0.86, $z$ = –5.71, p < .001). Participants in the First and Full Name conditions were more likely to respond *female* than participants in the Last Name condition ($\beta$ = 2.00, $z$ = 5.83, p < .001). There was no difference between the First and Full Name conditions.

**Table 4:** Experiment 2: Number of *female*, *male*, and *other* responses and the ratio of *female* responses to *male* and *other* responses for each condition.

| **Experiment 2: Number of Responses by Condition** | | | | |
|---|---|---|---|---|
| | ***Female*** | ***Male*** | ***Other*** | **Ratio of *Female* \| *Male* + *Other*** |
| **First** | 1579 | 1543 | 35 | 1.001 |
| **Full** | 1446 | 1633 | 85 | 0.842 |
| **Last** | 406 | 2498 | 232 | 0.149 |

**Table 5:** Experiment 2: Model results for the effect of Condition on the likelihood of *female* responses (= 1), as opposed to *male* and *other* responses (= 0).

| **Experiment 2: Condition** | | | | |
|---|---|---|---|---|
| | ***Recall as female*** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –0.861 | 0.151 | –5.710 | **<0.001** |
| **Condition: Last** (–.66) **vs. First** (+.33) + **Full** (+.33) | 2.000 | 0.343 | 5.843 | **<0.001** |
| Condition: First (–.5) vs. Full (+.5) | –0.231 | 0.345 | –0.669 | 0.50 |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 0.196 | | | |
| $\tau_{00\ Item}$ | 1.782 | | | |
| $N_{Participant}$ | 1351 | | | |
| $N_{Item}$ | 105 | | | |
| Observations | 9457 | | | |

†Bonferroni corrected $\alpha = .0167$.

Next, the effect of Gender Rating was analyzed for the First and Full Name conditions (**Table 6**), again following the same model specifications as in Experiment 1. The intercept term was significant ($\beta = –0.18$, $z = –2.99$, $p < .01$), indicating that participants were less likely to respond *female* in the First and Full Name conditions overall. *Female* responses became more likely as the names became more feminine ($\beta = 0.78$, $z = 22.34$, $p < .001$), but did not surpass *male* responses until the first name in the prompt was biased somewhat feminine, rather than at the mean (**Figure 3**). The interaction between Gender Rating and Condition was not significant, indicating that the linear effect of Gender Rating was similar in the First and Full Name conditions.

**Table 6:** Experiment 2: Model results for the effects of Condition and Gender Rating on the likelihood of *female* responses (= 1) as opposed to *male* and *other* responses (= 0) in the First and Full Name conditions.

| **Experiment 2: Condition and Gender Rating** | | | | |
|---|---|---|---|---|
| | ***Recall as female*** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –0.176 | 0.059 | –2.999 | **0.003** |
| Condition (First = –.5, Full = +.5) | –0.223 | 0.117 | –1.907 | 0.057 |
| **Gender Rating** (Centered, Masc –, Fem +) | 0.783 | 0.035 | 22.338 | **<0.001** |
| Condition × Gender Rating | –0.066 | 0.069 | –0.961 | 0.336 |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 0.114 | | | |
| $\tau_{00\ Item}$ | 0.141 | | | |
| $N_{Participant}$ | 903 | | | |
| $N_{Item}$ | 83 | | | |
| Observations | 6321 | | | |

†Bonferroni corrected α = .0125.

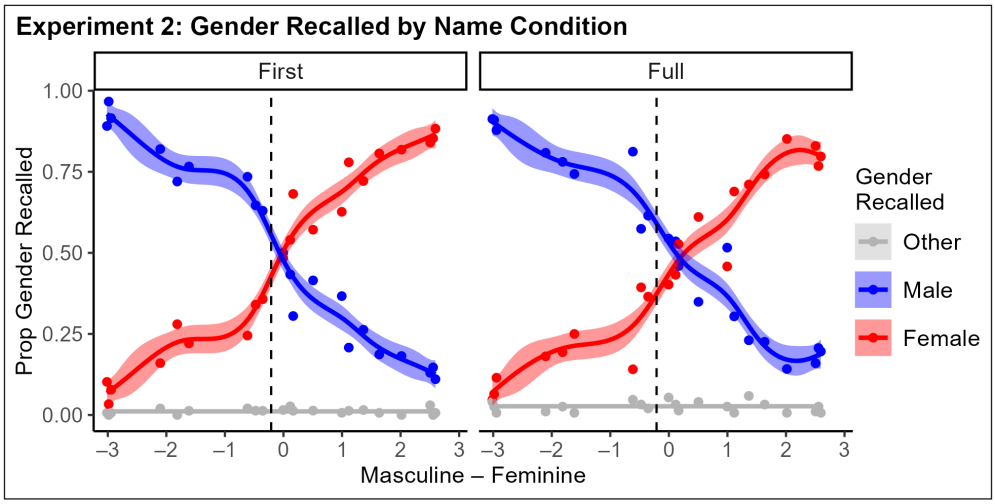


**Figure 3:** Experiment 2: Proportions of *male* (blue), *female* (red), and *other* (gray) responses in the First Name and Full Name conditions by the mean-centered gender rating of the first name. Points indicate means for each of the 21 first names, and lines indicate a smooth function on the raw data. Here, 0 is the mean of the 21 names, and the dashed line indicates the center of the original response scale in the norming study.

Next, we conducted the same three exploratory analyses as in Experiment 1: exclusion of the *other* responses, a quadratic effect of Gender Rating, and participant gender effects (Supplement §3.3–3.5). The rate of *other* responses (3.72%) was lower than in Experiment 1, and excluding *other* responses did not affect the substantive pattern of results. When adding a quadratic effect of Gender Rating to the Condition and Gender Rating model, neither the quadratic effect nor its interaction with Condition were significant. This finding is inconsistent with the hypothesis that the Condition effect would be larger at the midpoint of Gender Rating compared to the endpoints. Adding participant gender as a covariate revealed a significant interaction between Participant Gender and the Last vs. First + Full contrast ($\beta = -.42$, $z = -2.93$, $p < .01$), such that male participants were less likely than non-male participants to respond *female* in the First and Full Name conditions ($\beta = -.26$, $z = -3.75$, $p < .001$), whereas there was no effect of participant gender in the Last Name condition ($\beta = .15$, $z = 1.23$, $p = .22$). Across conditions, the effect of Gender Rating was smaller for male participants than for non-male participants ($\beta = -0.16$, $z = -2.64$, $p < .01$).

## 3.4 Discussion

Experiment 2 was designed to address the possibility that the preference for *he* responses in Experiment 1, especially in the Last Name condition, was due to participants using generic masculine forms, rather than a bias in the underlying gender inferences. Our findings were inconsistent with this as the primary explanation of the data. As in Experiment 1, participants' judgments about the genders of characters introduced in short narratives exhibited a bias to infer characters as male. This bias was strongest in the Last Name condition, compared to the conditions where the character was introduced including a first name (First and Full Name conditions). Participants did not recall the character as female 50% of the time at the midpoint on the name gender rating continuum, but instead when the names were somewhat feminine.

The bias to infer characters as male was present in Experiment 2, but smaller than in Experiment 1 (see comparison between experiments in **Figure 8**). This difference is primarily due to an attenuated difference between the Last Name and First + Full Name conditions, where participants were 16.84 times more likely to produce a *she* response in the First + Full Name conditions in Experiment 1, but 7.39 times more likely to respond *female* in the First + Full Name conditions in Experiment 2. The mismatch between knowledge about the gender associations of first names and inferences about the genders of characters with those names – where characters only began being preferentially inferred as female when first names were somewhat feminine, not at the midpoint – was consistent across the two experiments. In the First and Full Name conditions, the odds ratios were 0.70 for a *she* response compared to a *he* or *other* response and 0.77 for a *female* response compared to a *male* or *other* response (**Figure 8**).

One reason that the bias to infer characters as male was smaller in Experiment 2, aside from residual uses of generic *he* in the Last Name condition, is that people may be less likely to assume characters are male by default when the task requires them to think more directly about gender. This would be consistent with prior results, where after writing about a generic person, participants were two and a half times more likely to use masculine names to describe the character, as compared to two times more likely to explicitly label the character as male (Hamilton, 1988). Before considering this further, we first explore whether the people = male bias can be attenuated when people are provided with more information about, and repeated reference to, a character.

## 4. Experiment 3

Experiments 1 and 2 examined gender inferences after brief introductions to characters, but in many settings, we receive significantly more individuating information about a person before needing to refer to them or reflect on their gender. If so, having more information about a person as an individual may reduce reliance on the people = male bias in typical settings. To address this question, Experiment 3 investigates whether the bias to infer characters as male persists after repeated reference to the character in a narrative that highlights an aspect of their life. In addition to providing individuating information, the use of repeated reference in the narrative provides multiple opportunities and more time to process an inference about gender. We also explore whether the form of reference shapes perceptions of character traits beyond gender.

Participants in Experiment 3 read a paragraph-length story about a character, written as a short news story highlighting an accomplishment. Characters were always introduced with a full name, then referred to 3 more times, which varied by the same conditions as in prior experiments (*First Name*, *Last Name*, *Full Name*). Participants continued the story by completing a sentence fragment, and we measured which pronouns, if any, were used to refer to the character. After the sentence completion task, participants rated the character in terms of likeability, accomplishment, and importance. This process was repeated for 7 stories and characters. Prior research demonstrates that professionals who are referred to by last name, a convention associated with masculinity in English (Files et al., 2017; Uscinski & Goren, 2011), were judged as more accomplished and deserving of awards (Atir & Ferguson, 2018). Given these findings, we hypothesized that characters who are rated as more accomplished and important may be more likely to be referred to with *he*. Judgments of status and likeability frequently trade off in women (Stewart et al., 2003; Takiff et al., 2001), and so characters who are rated as more likeable may be more likely to be referred to with *she*.

## 4.1 Methods

### 4.1.1 Participants

Participants were recruited on Amazon Mechanical Turk, following the same criteria and procedures as in Experiments 1 and 2. The sample size (1350 planned) was chosen to generate the same number of data points as in Experiment 1 (150 participants per condition, completing 21 trials) and Experiment 2 (450 participants per condition, completing 7 trials). Because trials in Experiment 3 were longer than in Experiment 1, each participant completed only 7 trials. The final sample (N = 1272) included 405 in the First Name condition, 510 in the Last Name condition, and 357 in the Full Name condition, with conditions unbalanced due to variable rejection rates on MTurk. Participant exclusions and demographics are reported in Supplement §4.1.

### 4.1.2 Materials and procedure

As in Experiments 1 and 2, participants were randomly assigned to 1 of 3 between-participants conditions: *First Name, Last Name,* and *Full Name*. Participants saw paragraph-length stories that included the character's name 4 times, but did not use any gendered pronouns. The first reference to the character in the story always used a full name. The following 3 references to the character used either their first, last, or full name, according to the condition (**Figure 4**). The materials included a total of 7 stories, each written as a short news article highlighting the character's accomplishment: publishing a study, running a successful campaign event, having a bestseller, releasing a new album, breaking a running record, founding an animal rescue, and donating holiday meals. In addition to the 7 stories, the materials included 3 combinations of the 21 first names and 21 last names. Similar to Experiment 2, there were 9 lists within each condition, counterbalancing the combinations of names and stories and the combinations of first and last names; each participant was randomly assigned to 1 of the resulting 27 lists. Each list had first names evenly distributed across the gender ratings from masculine to feminine. After reading each of the 7 stories, participants were given a sentence fragment, which contained a 5$^{th}$ instance of the character's name, varying by condition. Participants were asked to complete the sentence. Next, they were asked to rate the character on a 1–7 scale as Likeable, Accomplished, and Important. The names in these prompts again varied according to condition.

## 4.2 Predictions

The results of Experiments 1 and 2 support a model where gender is inferred based on a combination of probabilistic cues to gender – here in the form of a person's first name – and an overall people = male bias. However, the characters were only mentioned once in Experiment 1 and twice in Experiment 2, and the Last Name condition contained no cues about the characters'

genders through the names themselves, only more indirect cues from the fact that men are more likely to be referred to by last name. The bias to infer people as male may only be present in the initial inferences established during these brief introductions. If so, additional information about the referent and time to process gender inferences may attenuate or eliminate this tendency. We would then expect to find no bias towards *he* responses when the probabilistic information about gender carried by the first name is repeatedly presented (First and Full Name conditions).

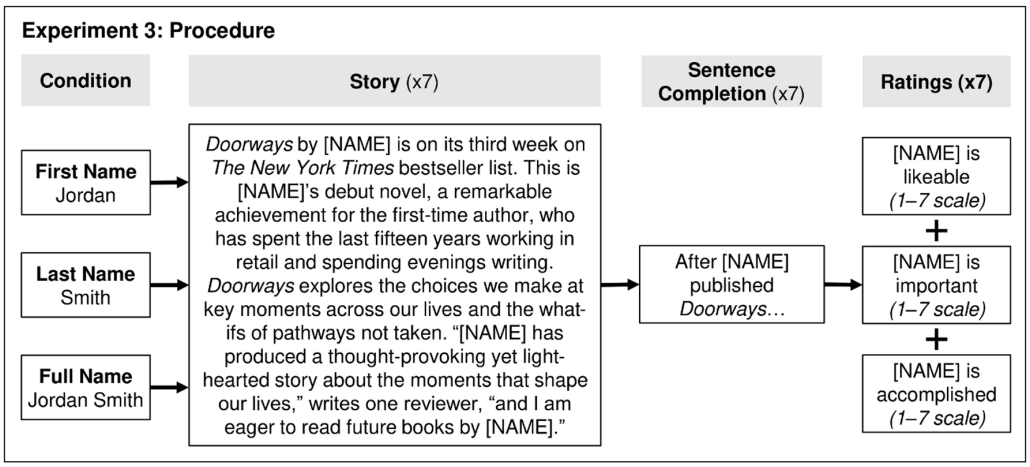


**Figure 4:** Experiment 3: Procedure and example stimuli.

Alternatively, if biased gender inferences persist after repeatedly encountering cues to a character's gender, we would expect to observe a bias to infer characters as male in all conditions, with *he* responses occurring more frequently than predicted by the gender distributions of the first names. One reason to expect the bias to persist comes from work showing that revising initial inferences about a character's gender incurs processing costs while reading (Carreiras et al., 1996; Garnham et al., 2002; Kennison & Trofe, 2003; Sturt, 2003).

## 4.3 Results

Responses were categorized as *he*, *she*, and *other*. The sentence completion prompts were less constrained than Experiment 1, with only 53% of responses beginning with a pronoun (compared to 93% in Experiment 1). As a result, we analyzed pronouns used to refer to the character at any position in the response (69% of responses). **Table 7** shows the proportions of responses across conditions, with *other* responses occurring in about a third of trials, an increase compared to Experiment 1. Responses were analyzed using logistic mixed-effect regression models, as before. Unlike Experiments 1 and 2, the contrasts for Condition were weighted to account for uneven numbers of participants in each condition. Recall that in all conditions, the first of 4 repetitions of the name was always a full name. As a result, we now analyze Gender Rating in all 3 conditions (**Table 8**).

**Table 7:** Experiment 3: Number of *she*, *he*, and *other* responses, the ratio of *she* responses to *he* and *other* responses, and the ratio of *she* to *he* responses for each condition.

| Experiment 3: Number of Responses by Condition | | | | | |
|---|---|---|---|---|---|
| | ***She*** | ***He*** | ***Other*** | **Ratio of *She* \| *He* + *Other*** | **Ratio of *She* \| *He*** |
| **First** | 941 | 992 | 902 | 0.497 | 0.949 |
| **Full** | 848 | 899 | 752 | 0.514 | 0.943 |
| **Last** | 1079 | 1378 | 1113 | 0.433 | 0.783 |

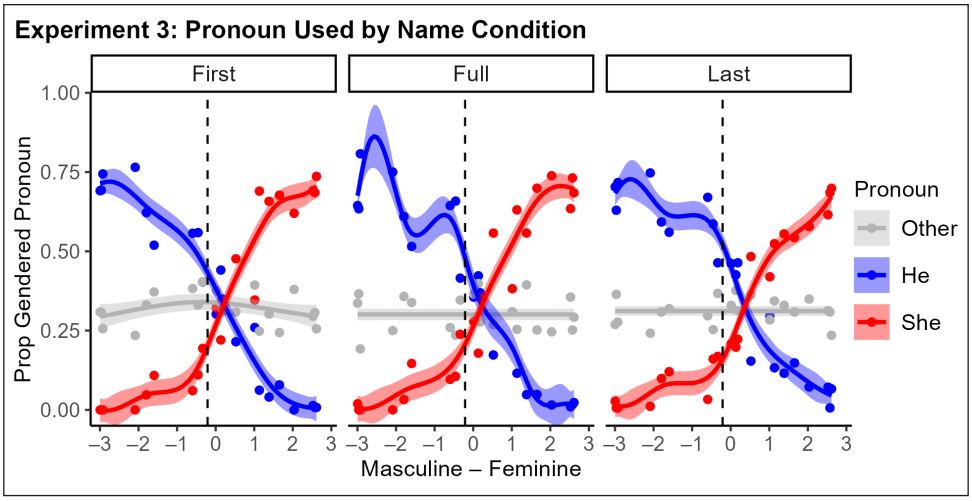


**Figure 5:** Experiment 3: Proportions of *he* (blue), *she* (red), and *other* (gray) responses in the First, Last, and Full Name conditions by the mean-centered gender rating of the first name. Points indicate means for each of the 21 first names, and lines indicate a smooth function on the raw data. Here, 0 is the mean of the 21 names, and the dashed line indicates the center of the original response scale in the norming study.

Overall, participants were less likely to produce *she* responses than *he* and *other* responses ($\beta = -1.53$, $z = -15.09$, $p < .001$). While *she* responses increased as the first name became more feminine ($\beta = 1.15$, $z = 19.02$, $p < .001$), *she* responses only surpassed *he* responses when names were somewhat feminine, not at the mean (**Figure 5**). Neither main effect of Condition was significant. The interaction between Gender Rating and the Last vs. First + Full condition contrast was significant ($\beta = 0.12$, $z = 2.15$, $p < .05$), such that the effect of Gender Rating was larger in the First + Full Name conditions compared to the Last Name condition. This interaction is likely due to the fact that the First + Full conditions had 4 repetitions of the gendered first name, whereas the Last Name condition only had 1 use of the first name. The interaction between Gender Rating and the First vs. Full Name condition contrast was not significant.

**Table 8:** Experiment 3: Model results for the effects of Condition and Gender Rating on the likelihood of *she* responses (= 1) as opposed to *he* and *other* responses (= 0).

| **Experiment 3: Condition and Gender Rating** | | | | |
|---|---|---|---|---|
| | ***Refer to using* she** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –1.524 | 0.101 | –15.090 | **<0.001** |
| Condition: Last (–.6) vs. First (+.4) + Full (+.4) | 0.153 | 0.092 | 1.674 | 0.094 |
| Condition: First (–.48) vs. Full (+.52); Last (.02) | 0.091 | 0.116 | 0.786 | 0.432 |
| **Gender Rating** (Centered, Masc –, Fem +) | 1.148 | 0.060 | 19.017 | **<0.001** |
| Condition (Last vs. First + Full) × Gender Rating | 0.105 | 0.049 | 2.153 | 0.031† |
| Condition (First vs. Full) × Gender Rating | –0.056 | 0.063 | –0.894 | 0.371 |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 0.793 | | | |
| $\tau_{00\ Item}$ | 0.421 | | | |
| $N_{Participant}$ | 1272 | | | |
| $N_{Item}$ | 63 | | | |
| Observations | 8904 | | | |

†Bonferroni corrected $\alpha = .0083$.

We conducted the same three exploratory analyses as in Experiments 1 and 2, following the same model specifications and applying Bonferroni corrections for multiple comparisons. Responses coded as *other* were similar to the types in Experiment 1 (e.g., repeating the character's name), but represented a larger proportion of the data (31%). When excluding *other* responses (Supplement §4.3), the Last vs. First + Full contrast was significant ($\beta = 0.26, z = 2.63, p < .01$), such that participants were less likely to produce *she* in the First and Full Name conditions than in the Last Name condition, similar to the results of the first two experiments. The intercept ($\beta = –0.42, z = –3.42, p < .001$) and the interaction between Condition and Gender Rating both remained significant ($\beta = 0.42, z = 5.46, p < .001$) in this subset of the data.

Adding a quadratic effect of Gender Rating to the primary model, which included *other* responses, revealed a significant quadratic effect of Gender Rating ($\beta = –0.11, z = –3.67, p < .001$). Inspection of the data suggests that this effect may be due to stronger effects of name rating towards the center of the gender rating scale (androgynous names) than at the end points (strongly-gendered names). The interaction between the quadratic effect of Gender Rating and the Last vs. First + Full contrast was significant ($\beta = –0.10, z = –3.24, p < .001$). Probing this

interaction indicated that the quadratic effect of Gender Rating was significant in the First and Full Name conditions ($\beta$ = –0.15, $z$ = –4.28, p < .001), but not in the Last Name condition ($\beta$ = –0.06, $z$ = –1.67, p = .09), which likely reflects that the Last Name condition only included the gendered first name in 1 out of the 4 repetitions. This analysis also indicated a significant Condition effect for the Last vs. First + Full contrast ($\beta$ = 0.24, $z$ = 3.00, p < .01), such that participants were more likely to produce *she* in the First and Full Name conditions compared to the Last Name condition (Supplement §4.4).

Adding participant gender as a covariate revealed that male participants were less likely than non-male participants to produce *she* responses as compared to *he* and *other* responses across all three conditions ($\beta$ = –.33, $z$ = –3.53, p < .001). Participant Gender did not significantly interact with Condition or Gender Rating (Supplement §4.5). Finally, we conducted exploratory analyses of the Accomplishment, Likeability, and Importance ratings (Supplement §4.6). Because these ratings were near ceiling at the positive ends of the scales, the results were largely nonsignificant, with the exception that more likeable characters were more likely to be referred to with *she*.

## 4.4 Discussion

Experiment 3 examined whether the bias to produce *he* persists with more information about the character and more time to process an inference about their gender. Participants read paragraph-length news stories that mentioned a character four times and described their noteworthy accomplishment. All characters were introduced by their full name, and the following three references carried varying gender cues. In the First and Full Name conditions, the first name was repeated; in the Last Name condition, the first name was not repeated, but the choice of form of reference is an indirect cue to gender, given that men are more likely to be referred to by last name. Despite the fact that all characters were first introduced with their full name, the bias towards producing *he* persisted. *She* responses increased as the gender rating of the first names became more feminine, but only overtook *he* responses when the names were somewhat feminine, not at the midpoint. The effect of gender rating was stronger in the First and Full Name conditions, where the gendered first name was repeated four times, than in the Last Name condition, where the first name only appeared once.

Across conditions, the odds ratio of a *she* response (vs. a *he* or *other* response) was 0.24 in Experiment 1 and 0.22 in Experiment 3 (**Figure 8**). Note, however, that there was a higher rate of *other* responses in Experiment 3. It is unclear how much this reflects a greater flexibility in the sentence completion prompts, with the stories in Experiment 3 allowing more felicitous continuations not using a third-person pronoun to refer to the character than the sentences in Experiment 1. Excluding *other* responses, the odds ratio of a *she* response was 0.32 in Experiment 1 and 0.65 in Experiment 3. This suggests that the *he* response bias was attenuated in comparison

to Experiment 1, but still present. This difference across studies was most notable in the Last Name condition: In Experiment 1, where the Last Name condition did not contain direct cues to gender, the odds ratio of a *she* response (vs. a *he* response, *other* responses excluded) was 0.04. In Experiment 3, where the first mention was by full name, the odds ratio of a *she* response (vs. a *he* response, *other* excluded) was 0.51. Thus, providing the comprehender with probabilistic information about a character's gender may attenuate, but cannot completely override, the bias to infer characters as male instilled by reference by a last name.

# 5. Experiment 4

Experiment 4 investigated if the bias to infer characters as male after a short delay persists when participants have more information about the characters, see repeated cues about their gender, and are asked directly about their gender inferences later. Participants read a series of paragraph-length stories about a character, written as a short news story highlighting an accomplishment. Characters were introduced with a full name, then referred to three more times following the same conditions as prior experiments (*First Name, Last Name, Full Name*). After reading each story, participants rated the character on Likeability, Accomplishment, and Importance. After reading stories about 7 characters and rating each character, there was a brief delay during which participants completed simple math problems. Next, participants were cued with the activity described in each of the 7 stories, one at a time, and were asked to recall the gender of the character.

Experiments 2 and 4 differ from Experiments 1 and 3 in that these studies directly ask about gender inferences. In addition, a key feature of Experiments 2 and 4 is that participants in these studies read stories about all 7 characters and only then were asked about the gender of the 7 characters. While we can assume that participants inferred the gender of the characters as they were reading (Duffy & Keir, 2004; Garnham et al., 2002; Kennison & Trofe, 2003; Osterhout et al., 1997; Reynolds et al., 2006; Sturt, 2003), the instructions did not guide participants to read the stories with the intention of remembering the characters' genders, and the reading task did not prompt participants to read with the intention of designing a story continuation referring to the character. Thus, in both Experiments 2 and 4, while we expect that participants will make an inference about the gender of each character as they read, they are not aware that they will be later asked about this inference.

## 5.1 Methods

### 5.1.1 Participants

Participants were recruited on Amazon Mechanical Turk, following the same criteria and procedures as in Experiments 1–3. The sample size (1350 planned) was chosen to generate

the same number of data points as Experiments 1–3; here, 450 participants in each of the 3 conditions completed 7 trials each. A total of 1361 responses were recorded. The final sample (N = 1253) included 422 participants in the First Name condition, 415 in the Last Name condition, and 416 in the Full Name condition. Participant exclusions and demographics are shown in Supplement §5.1.

### 5.1.2 Materials and procedure

Participants read the same stories as in Experiment 3, with characters that were introduced with a full name and subsequently referenced 3 more times, varying according to the 3 between-subjects conditions. After reading each story, participants rated the characters on Likeability, Accomplishment, and Importance. After a short delay, participants were asked to recall the gender of each character, as in Experiment 2. They were cued by the action in the story, without using the name or any gendered pronouns, and entered their answers in a free-response box. The 9 lists within each condition, counterbalancing names and prompts, were identical to Experiment 3; again participants were randomly assigned to 1 of 27 lists. **Figure 6** shows the procedure and an example story.

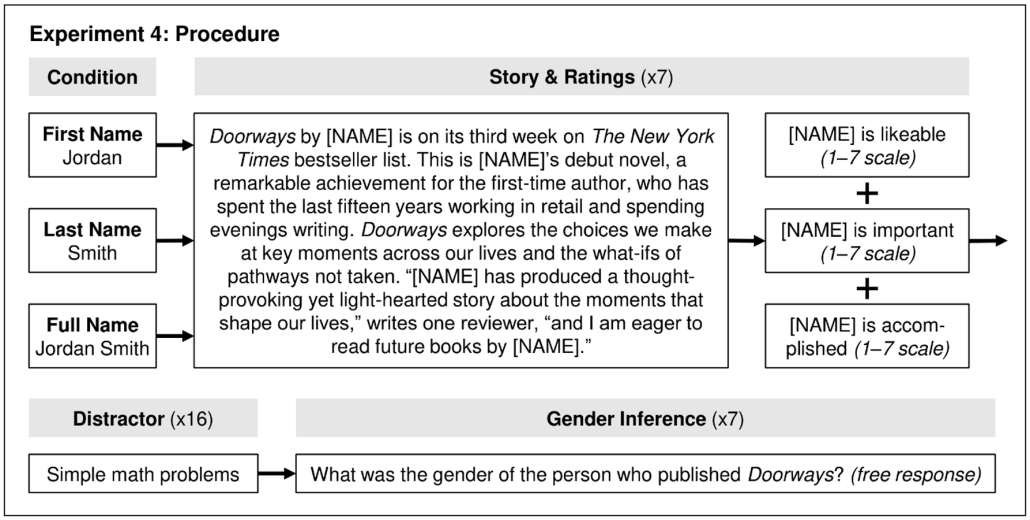


**Figure 6:** Experiment 4: Procedure and example stimuli.

## 5.2 Predictions

Thus far, our findings support a model of gender inference where probabilistic cues to gender are combined with a bias to infer that a character is male. In Experiment 4, we test the predictions of this model when the characters are presented in more detail and where the probabilistic cue to gender (in the form of the first name) is more strongly established, but when the sequence of

the experiment does not prompt participants to be making inferences about gender when first reading the stories. While the results of Experiment 3 demonstrated a persistent bias to produce *he*, here we ask if the bias to infer characters to be male persists when directly asked about the character's gender. If so, the bias towards recalling characters as male will be present even when probabilistic information about gender is repeatedly provided (First and Full Name conditions). Alternatively, if the bias to infer referents as male is attenuated after probabilistic cues about gender are well-established and when gender inferences are asked about directly, we would expect no bias towards recalling characters as male in the First and Full Name conditions.

## 5.3 Results

As in Experiment 2, responses were coded as *male*, *female*, or *other* (**Table 9**) and analyzed using logistic mixed-effect regression models predicting the log odds of a *female* response as opposed to *male* and *other* responses (**Table 10**). Characters were less likely to be recalled as female overall ($\beta = -0.26$, $z = -3.14$, $p < .01$), and somewhat more likely to be recalled as female in the First + Full Name conditions than in the Last Name condition ($\beta = 0.13$, $z = 0.94$, $p < .05$). The comparison between First and Full Name conditions was not significant. Participants responded *female* more frequently as the first names became more feminine ($\beta = 0.76$, $z = 16.65$, $p < .001$), but *female* responses did not overtake *male* responses until the first names were somewhat feminine (**Figure 7**). The interaction between Gender Rating and Condition was significant for the Last vs. First + Full contrast ($\beta = 0.13$, $z = 3.81$, $p < .001$), due to a larger effect of Gender Rating in the First + Full Name conditions, where the first name was repeated four times, as compared to the Last Name condition, where the first name was only presented once. An interaction between Gender Rating and the First vs. Full contrast ($\beta = -0.10$, $z = -2.45$, $p < .05$) was due to a larger effect of Gender Rating in the First Name condition than in the Full Name condition.

**Table 9:** Experiment 4: Number of *female*, *male*, and *other* responses and the ratio of *female* responses to *male* and *other* responses for each condition.

| **Experiment 4: Number of Responses by Condition** | | | | |
|---|---|---|---|---|
| | ***Female*** | ***Male*** | ***Other*** | **Ratio of *Female* \| *Male* + *Other*** |
| **First** | 1381 | 1511 | 62 | 0.878 |
| **Full** | 1380 | 1416 | 116 | 0.901 |
| **Last** | 1292 | 1529 | 84 | 0.801 |

**Table 10:** Experiment 4: Model results for the effects of Condition and Gender Rating on the likelihood of *female* responses (= 1) as opposed to *male* and *other* responses (= 0).

| **Experiment 4: Condition and Gender Rating** | | | | |
|---|---|---|---|---|
| | ***Recall as female*** | | | |
| ***Predictors*** | ***Log-Odds*** | ***SE*** | ***z*** | ***p*** |
| **(Intercept)** | –0.256 | 0.082 | –3.138 | **0.002** |
| Condition: Last (–.67) vs. First (+.33) + Full (+.33) | 0.126 | 0.062 | 2.048 | 0.041† |
| Condition: First (–.49) vs. Full (+.51) | 0.068 | 0.072 | 0.944 | 0.345 |
| **Gender Rating** (Centered, Masc –, Fem +) | 0.764 | 0.046 | 16.648 | **<0.001** |
| **Condition (Last vs. First + Full) × Gender Rating** | 0.131 | 0.035 | 3.809 | **<0.001** |
| Condition (First vs. Full) × Gender Rating | –0.103 | 0.042 | –2.447 | 0.014† |
| *Random Effects* | | | | |
| $\tau_{00\ Participant}$ | 0.201 | | | |
| $\tau_{00\ Item}$ | 0.360 | | | |
| $N_{Participant}$ | 1253 | | | |
| $N_{Item}$ | 63 | | | |
| Observations | 8771 | | | |

†Bonferroni corrected $\alpha = .0083$.

Finally, we conducted the same set of supplementary analyses as in prior experiments (Supplement §5.3–5.5). Excluding *other* responses (2.99% of total responses) revealed a similar pattern of results as the primary analysis. Adding a quadratic effect of Gender Rating revealed no new significant effects after Bonferroni correction for multiple comparisons. Adding Participant Gender as a covariate revealed that male participants were overall less likely than non-male participants to recall the character as female ($\beta = –.20$, $z = –3.27$, $p < .001$). As in Experiment 3, the Accomplishment, Likeability, and Importance ratings were near ceiling at the positive ends of the scales, and more likeable characters were more likely to be recalled as female. Additionally, interactions between each of the three character ratings and Gender Rating indicated that the effects of Likeability, Accomplishment, and Importance on gender inferences were stronger with more feminine names (Supplement §5.6).

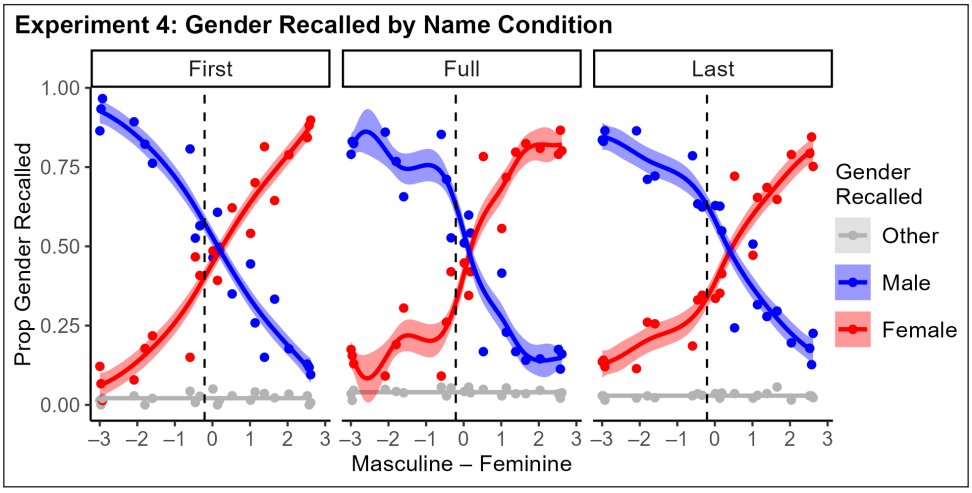


**Figure 7:** Experiment 4: Proportions of *male* (blue), *female* (red), and *other* (gray) responses in the First, Last, and Full Name conditions by the mean-centered gender rating of the first name. Points indicate means for each of the 21 first names, and lines indicate a smooth function on the raw data. Here, 0 is the mean of the 21 names, and the dashed line indicates the center of the original response scale in the norming study.

### 5.4 Discussion

Experiment 4 examined whether the bias to explicitly recall characters as male persists when participants see additional information about, and repeated reference to, the character in a narrative, and when they have more time to develop inferences about the character but are less prompted to do so intentionally by the structure of the experiment. Although all characters were first introduced with their full name, which provided probabilistic information about their gender, participants were overall less likely to infer the character as female than to infer them as male or not indicate a gender inference. Characters needed to have first names that were rated somewhat feminine before they were more likely to be recalled as female, while characters with androgynous first names were more likely to recalled as male. The effect of the first name's gender rating was strongest in the First Name condition, where the first name appeared in all 4 references to the character, and weakest in the Last Name condition, where the first name only appeared once. The bias to infer characters as male was smaller in Experiment 4 as compared to Experiment 2, but still not eliminated: the odds ratio of a *female* (vs. *male* or *other* response) across conditions was 0.42 in Experiment 2, and 0.77 in Experiment 4 (**Figure 8**).

## 6. General discussion

### 6.1 Overview of findings

The present studies investigated how choices in how we refer to a person affect inferences about that person's gender. Specifically, we examined how referring to a character by first name, last

name, or full name impacted two measures of the readers' gender inferences: pronoun use in a sentence completion task and responses to an explicit question about gender. We considered two competing hypotheses about the gender inference process. One hypothesis was that people only show a bias to assume referents are male when few cues about gender are available (e.g., you only know the person's last name). Alternatively, we hypothesized that gender inferences might be shaped by a combination of the people = male bias (Silveira, 1980), along with other probabilistic cues to gender.

Across four experiments, using both short and long character introductions and two measures of gender inference, we observed that inferences about gender were shaped by a persistent people = male bias, along with clear use of probabilistic cues to gender. The results of Experiment 1 showed that characters who were referred to by last name only were overwhelmingly referred to with *he*. While providing more direct cues to gender through a first name attenuated this bias, it did not eliminate it. Instead, a character's first name had to be at least somewhat feminine before *she* responses became more common than *he* responses. The increases in *she* and *he* responses were asymmetric, with androgynous names that leaned feminine still eliciting *he* responses, but androgynous names that leaned masculine eliciting responses that did not use a pronoun instead of *she* responses. An alternative interpretation of these findings, however, is that bias to use *he*, particularly in the Last Name condition, was due to participants using the generic masculine, and not due to biased inferences about gender. To address this possibility, Experiment 2 asked participants about their gender inferences directly. The results of Experiment 2 revealed a persistent, if smaller, overall bias to recall the characters as male. In addition, when cues to gender were provided through the use of the character's first name, the name had to be somewhat feminine before the character was preferentially recalled as female. This mismatch between knowledge about the gender distributions of the first names and inferences about the characters from those names remained similar between Experiments 1 and 2.

In the first two experiments, participants read sentence-length descriptions of the characters. One question, then, is whether the observed bias in gender inferences would persist when a character is referred to multiple times and more is known about them. If the people = male bias attenuates as more information about the person has accrued and more time has been spent thinking about them, it may be attenuated in longer texts. To address this question, Experiments 3 & 4 examined gender inferences in paragraph-length news stories, where characters were introduced by full name and subsequently referred to three more times (manipulated by condition). In comparison to Experiments 1 & 2, these stories provided repeated cues to the character's gender and more time for the reader to process inferences about the character, as well as more closely resembling a way we might read about a new person in everyday life. While the people = male bias was numerically smaller in Experiments 3 & 4 compared to Experiments 1 & 2, characters were still less likely to be referred to with *she* and less likely to be recalled as female

than the distributions of the first names would predict. The strongest biases were observed when no direct information about gender was provided (Last Name condition in Experiments 1 & 2), where about 80% of characters were referred to with *he* or recalled as male. It is worth noting the magnitude of the people = male bias is roughly the same as results from 20–30 years ago (Davis Merritt & Kok, 1995; Hamilton, 1988), despite continuing social advances in gender equality.

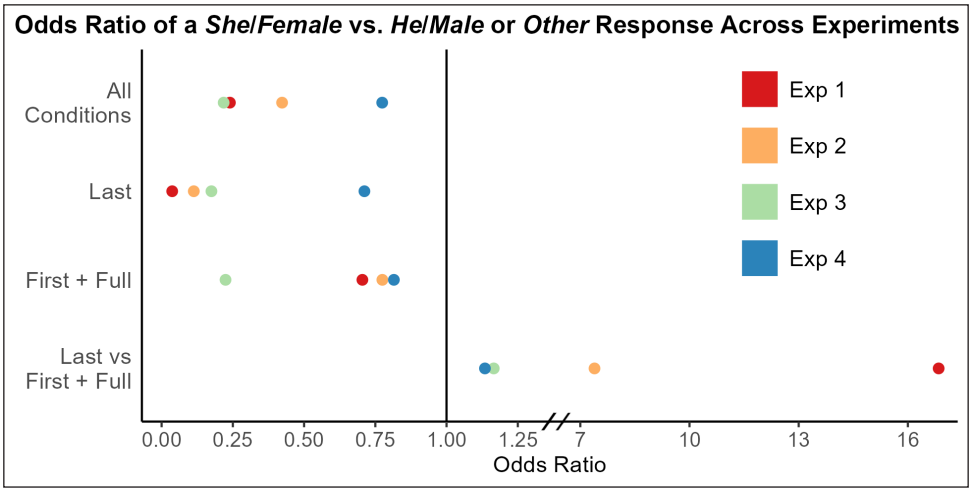


**Figure 8:** Comparing Experiments 1–4: Odds ratios of *she* vs. *he* + *other* and *female* vs. *male* + *other* responses averaged across all condititions, in the Last Name condition only, in the First + Full Name conditions only, and in the Last Name compared to the First + Full Name conditions. Values less than 1 indicate being less likely to produce a *she/female* response; values greater than 1 indicate being more likely to produce a *she/female* response, with a discontinuous X axis to include the 2 largest values. Odds ratios correspond to the exponentiated beta estimates reported in the models.

## 6.2 Choosing referential forms

A primary focus of research on pronoun production in English has been on when speakers use pronouns instead of other referential expressions. A common assumption is that there is a causal link between a referent's status in the discourse (e.g., whether it was the sentence topic, how often it was mentioned, and in what syntactic position, among others) and the form of reference the speaker selects. As a referent becomes more focused, salient, or prominent, the forms of reference used generally become more reduced, and pronouns are more likely (e.g., Gundel et al., 2012; Rodhe & Kehler, 2014; see Arnold & Zerkle, 2019, for discussion). For example, Schmitt et al.'s (1999) model of lexical access in pronoun production, tested in German, assumes that if a lexical concept is activated and sufficiently "in focus" in the discourse, the speaker will produce a pronoun instead of a full noun phrase. In this model, activating the lexical concept also activates the corresponding grammatical gender node (masculine, feminine, neuter), and if the speaker

uses a pronoun, the gender node is selected in order to produce the correct pronoun (Jescheniak & Levelt, 1994; Levelt et al., 1999; Roelofs, 1992). Lexical access models generally agree that grammatical gender is represented as a separate lexical-syntactic feature in the mental lexicon (Wang & Schiller, 2019), but models differ in the structure of, and time course with which, this feature is connected to other linguistic representations (serial, unidirectional connections, e.g., Jescheniak & Levelt, 1994; Levelt et al., 1999; Roelofs, 1992; or bidirectional connections, e.g., Dell, 1986, 1988, 1999; Dell & O'Seaghdha, 1992). When producing gender-marked pronouns, as well as determiners, competition between different forms can arise from the grammatical gender features of *other* lexical concepts that are also activated (Schiller & Caramazza, 2003). Generally, models of grammatical gender selection do not include competition between multiple gender features activated by the *same* lexical concept. The closest analogue is languages where the singular and plural forms of determiners vary for the same grammatical gender, and one approach argues that the singular form is activated by default, and can interfere with activating the plural form (Jescheniak et al., 2014; Schriefers et al., 2002).

Complexities arise, however, when we consider that gendered language talking about people reflects a social construct negotiated between speakers, not a discrete grammatical or semantic feature (Ackerman, 2019; Conrod, 2020; McConnell-Ginet, 2014). In contrast to grammatical gender, information about social gender carried by names is probabilistic. We know from experience that most people named Mary are referred to with *she*, most people named Brian are referred to with *he*, and people named Jordan are commonly referred to with *he* or *she*. But without knowing more about a particular person, speakers may be unsure of which pronouns are appropriate. Additionally, there are many contexts in which multiple choices of pronouns are available, such as using singular *they* instead of *he* or *she* to leave a referent's gender unspecified (e.g., *My friend$_i$ sent me a picture of their$_i$ cat*) and using singular *they* instead of *he* or *she* for people who use *they/them* pronouns. This means that models of pronoun production that include reference to people need to account for speakers' decisions about *which* pronouns to produce, in addition to decisions about *when* to produce pronouns.

Our findings offer insights into the mechanisms guiding inferences about gender and the processes by which people choose pronouns to refer to a person. We propose that the ways in which we refer to people are influenced by multiple factors, including speaker knowledge of gender distributions, speaker inference about a referent's gender, and speaker pronoun choice, in turn influencing the comprehender's inference about referent gender (**Figure 9**). To explain our findings, we discuss potential locations for bias to infer referents as male in this process. In particular, we focus on contexts like those in our experimental stimuli, where speakers have cues about gender in the form of names, but need to make an inference about which pronouns, if any, are appropriate to produce.

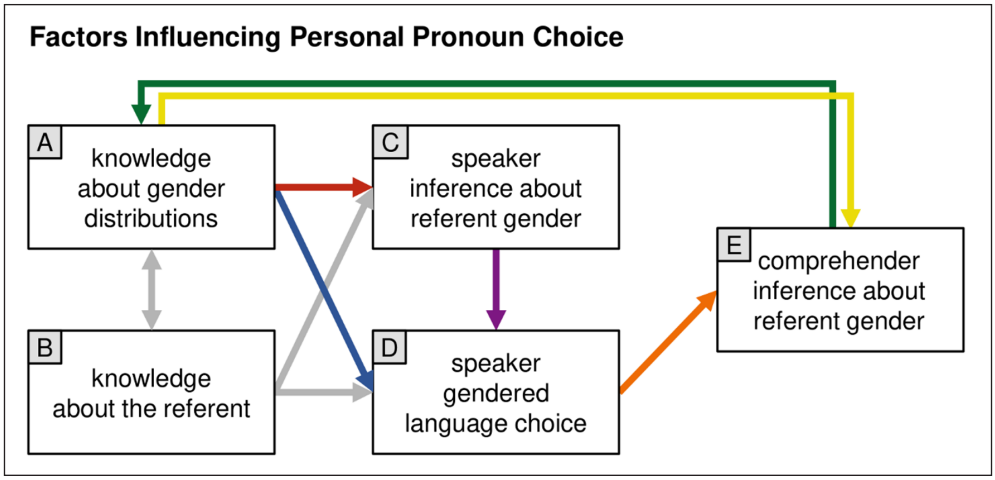


**Figure 9:** Factors influencing personal pronoun choice.

First, we assume that, based on world experience, speakers form and store estimates of gender distributions across contexts, including, at the most basic level, the knowledge that close to half of people are women or girls ([A] in **Figure 9**). Speakers also have information about the proportions of women in specific contexts: Estimates about the relative rates of women in various jobs showed a strong correlation with employment data, and when actual and estimated data diverged, participants were more likely to overestimate the proportion of men in a given occupation than to overestimate the proportion of women (Garnham et al., 2015; Misersky et al., 2014). Similarly, our norming study found a strong positive correlation between how feminine a first name was rated and the proportion at which it was given to children assigned female at birth. These findings indicate that people have reasonably well-calibrated knowledge of gender distributions in occupational contexts and based on first names, and that biases, when present, more frequently underestimate the proportion of women.

Speakers also have knowledge about a particular referent [B], and in many contexts, this includes information about what names, pronouns, titles, and other forms of reference are appropriate for them. In the present experiments, we focus on contexts where relatively little information is provided about the referent, requiring speakers to make inferences about their gender [C] and what language to use to refer to them [D]. Context-specific knowledge about gender distributions [A] contributes to the speaker's inferences about a referent's gender [C]. Previous findings show that this process [red] underestimates the prevalence of women, since people are biased to infer gender-unspecified referents as male (Davis Merritt & Kok, 1995; Davis Merritt & Wells Harrison, 2006; Silveira, 1980). In the present research, when no direct cues to gender were provided and participants were asked about their gender inferences (Last Name condition in Experiment 2), approximately 80% of responses inferred the referent to be male.

This bias persisted, albeit attenuated, when some gender information was given (First and Full Name conditions) and after repeated reference (Experiments 3 & 4).

The speaker's knowledge of gender distributions in general [A] is one contributor to their choice of what referring language to use [D], though speakers tend to use feminine language forms less frequently than their probabilistic knowledge about gender distributions would predict [blue] (Boyce et al., 2019; Hamilton, 1988; von der Malsburg et al., 2020). In the present research, participants were less likely to use *she* to refer to characters than the gender distribution of the first names would predict, after only one reference to the character and after repeated reference. One explanation is that the criterion for inferring referents as male is lower than the criterion for inferring them as other genders. This is one implication of the people = male hypothesis (Silveira, 1980): if the generic person is a man, then producing *he* (the unmarked category) might require a lower threshold of evidence than producing *she* (the marked category). Since pronouns are typically reserved to refer to the most salient, accessible, or in focus character (Arnold & Zerkle, 2019), a related possibility is that characters inferred as male are seen as more salient, and thus more likely to be referenced using a pronoun. This may explain why, in the First and Full Name conditions, *he* responses were dominant across the masculine half of the scale, whereas *she* and *other* responses (which typically did not use a pronoun) were both common in the feminine half of the scale.

Another factor in the choice of referring language [D] is the speaker's inference about that specific referent's gender [C]. A disconnect between the two is another potential source of bias [purple]: Recall that while around half of participants believed Hillary Clinton would win the 2016 US election, only 10% of responses used *she* to refer to the next president (von der Malsburg et al., 2020). When asked for both explicit and implicit measures about a character's gender, participants described generic referents as female at higher rates than they chose feminine names to refer to them (Hamilton, 1988).

These findings suggest that inferences about referent gender and choices about gendered language are distinct processes. Mappings between gender inferences and language choice also vary by dialect, such as uses of generic *he* and singular *they*. Moreover, these mappings are not always symmetric. This was clearest in Experiment 1, where participants still used *he* when the character had a feminine-leaning androgynous name, but were equally likely to use *she* or no pronouns for characters with masculine-leaning androgynous names. This suggests that a speaker's inference about a referent's gender may need to be more certain to prompt the use of *she* than to prompt the use of *he*.

It is important to note that the discrete choices involved in gendered language production do not preclude the underlying inference about a referent's gender being probabilistic. When participants in Experiments 1 & 3 used *he* or *she*, this did not necessarily reflect certainty about

a gender inference. This response pattern may be more common for speakers of dialects where forms that directly encode uncertainty, such as using singular *they* for a referent with an unknown or unspecified gender, are not available. Future work could explore how the same language produced may reflect underlying levels of confidence. The experiments here cannot distinguish between the contributions of distributional knowledge [blue] and inferences about a specific referent's gender [purple], only conclude that the resulting choices show a bias against using feminine language forms.

Comprehenders use speakers' gendered language choices [D], as well as their knowledge of gender distributions [A], to form their own inference about a new referent's gender [E]. One possibility is that comprehenders know that speaker pronoun choice is biased towards *he* and correct for this bias [weighting yellow over orange]. Although the experiments here do not address this question, Boyce et al. (2019) suggest that this is not the case. Participants read stories that included two repetitions of a role noun and one gendered pronoun. When asked to recall the character's gender, participants did not correct for masculine bias in pronoun use, and instead continued to recall the referents as feminine at lower rates than the normed gender distribution of the role nouns.

Another aspect at play here is comprehenders' knowledge of who and what speakers discuss. While the present experiments have focused on the probabilistic information about gender carried by names, the contexts in which a person is mentioned can also provide gender cues. This is particularly relevant in Experiments 3 & 4, where the stimuli were news stories highlighting a person's accomplishment, instead of sentences describing a person performing everyday activities. While a comprehender may know that people named Jordan are about equally likely to be male or female, they may also know that a person mentioned for their recent career accomplishment is more likely to be male. Several of the stories described more stereotypically-feminine accomplishments (i.e., charity work), but most were stereotypically more masculine (i.e., politics, sports). The gender stereotypes of the stories were counter-balanced within the experiment by having each story paired with masculine, feminine, and androgynous first names across lists. However, the present experiments did not attempt to measure or experimentally manipulate the fact that people may be making additional inferences about gender based on the fact that the character accomplished something and that their accomplishment was judged as newsworthy. The results here show a bias to infer people are male and to use masculine language forms when making inferences about a character in a brief story, leaving open questions about how these biases interact with speakers' original choices of who to discuss.

Finally, it is likely that inferences about gender in comprehension [E] influence underlying beliefs about gender distributions [A]. As such, speaker's choices about how to refer to entities in the world arguably drive patterns in language comprehension (MacDonald, 2013). Thus, if speakers consistently underuse feminine forms of reference and comprehenders do not correct

for this bias, beliefs about gender distributions in general and in specific contexts may then become biased to underestimate women [green].

## 6.3 Implications for talk about women

These results have potential implications for how we talk about women. When we refer to people, we choose between different combinations of forms, including pronouns, first names, last names, gendered titles (*Mr.*, *Mrs.*, *Ms.*), and nominally ungendered titles (*Doctor*, *Professor*). If certain forms of reference make feminine referents less likely to be inferred as feminine, should this influence which forms we choose? On the one hand, prior work suggests that people who are referred to with masculine-coded terms are judged more competent and successful (Atir & Ferguson, 2018; Rubin, 1981; Stewart et al., 2003; Takiff et al., 2001). Given these observations, a strategic speaker or writer could refer to a woman using masculine-coded forms to encourage a more masculine interpretation of the referent. This could mean reaping potential advantages (e.g., in perceived "eminence"), but potentially at the cost of having someone's femininity be diminished or unacknowledged, and of perpetuating language production patterns that in turn may shape biases in comprehension. Alternatively, it may be preferable to work to change the underlying tendency to underestimate the presence of women, by choosing language forms that make it more difficult to assume a person is a man by default, especially in contexts where women are less visible. An open question, then, is if using gendered language to refer to women in contexts where their presence is systematically underestimated (e.g., doctors) would change perceptions about the contributions of women to that sphere of life.

## 6.4 Conclusion

Across a series of four experiments, we asked if alternative forms of reference to individuals guide inferences about the gender of a referent introduced in a sentence or brief story. Using measures of personal pronoun choice (Experiments 1 & 3) and explicit queries about gender (Experiments 2 & 4), a persistent bias to assume that the referent was male was observed across all four studies. This bias was strongest when the character was introduced by last name alone, but persisted even when the character was referred to by their first name. We argue that the observed male bias in gender inference likely results from multiple processes, including biases in knowledge of gender distributions, inferences about a referent's gender, and pronoun choice.

## Data accessibility statement

The study preregistrations, materials, de-identified data, analysis code, and supplementary analyses all 4 experiments are available on OSF (10.17605/OSF.IO/AYPU2) and this project's Github repository (10.5281/zenodo.10293754).

## Ethics and consent

The protocol for this study was reviewed by the Vanderbilt University IRB (160023), and all participants gave informed consent.

## Acknowledgements

This work was supported in part by National Science Foundation 1556700 and 1921492 to Sarah Brown-Schmidt.

## Competing interests

The authors have no competing interests to declare.

## Authors' contributions

*Conceptualization:* Bethany Gardner, Sarah Brown-Schmidt

*Data curation:* Bethany Gardner

*Formal analysis:* Bethany Gardner, Sarah Brown-Schmidt

*Funding acquisition:* Sarah Brown-Schmidt

*Investigation:* Bethany Gardner

*Methodology:* Bethany Gardner, Sarah Brown-Schmidt

*Software:* Bethany Gardner

*Supervision:* Sarah Brown-Schmidt

*Visualization:* Bethany Gardner

*Writing (initial draft):* Bethany Gardner

*Writing (review & editing):* Sarah Brown-Schmidt

## References

Ackerman, L. (2019). Syntactic and cognitive issues in investigating gendered coreference. *Glossa: A Journal of General Linguistics*, *4*(1), 117. DOI: https://doi.org/10.5334/gjgl.721

American Psychological Association [APA]. (2019). *Singular* they. APA Style. https://web.archive.org/web/20211124212947/https://apastyle.apa.org/style-grammar-guidelines/grammar/singular-they

APA Publication Manual Task Force. (1997). Guidelines for nonsexist language in APA journals. *American Psychologist*, *32*(6), 487–494. DOI: https://doi.org/10.1037/0003-066X.32.6.487

Arnold, J. E., & Zerkle, S. A. (2019). Why do people produce pronouns? Pragmatic selection vs. Rational models. *Language, Cognition and Neuroscience*, *34*(9), 1152–1175. DOI: https://doi.org/10.1080/23273798.2019.1636103

Atir, S., & Ferguson, M. J. (2018). How gender determines the way we speak about professionals. *Proceedings of the National Academy of Sciences*, *115*(28), 7278–7283. DOI: https://doi.org/10.1073/pnas.1805284115

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bodine, A. (1975). Androcentrism in prescriptive grammar: Singular *they*, sex-indefinite *he*, and *he or she*. *Language in Society*, *4*(2), 129–146. DOI: https://doi.org/10.1017/S0047404500004607

Boyce, V., von der Malsburg, T., Poppels, T., & Levy, R. (2019). *Remember him, forget her: Gender bias in the comprehension of pronominal referents* [Conference Talk]. 32nd Annual CUNY Conference on Human Sentence Processing. https://osf.io/c8b3f/

Cameron, J. J., & Stinson, D. A. (2019). Gender (mis)measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, *13*(11), e12506. DOI: https://doi.org/10.1111/spc3.12506

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology*, *49A*(3), 639–664. DOI: https://doi.org/10.1080/713755647

Conrod, K. (2020). Pronouns and gender in language. In K. Hall & R. Barrett (Eds.), *Oxford handbook of language and sexuality*. DOI: https://doi.org/10.1093/oxfordhb/9780190212926.013.63

Davis Merritt, R., & Kok, C. J. (1995). Attribution of gender to a gender-unspecified individual: An evaluation of the people = male hypothesis. *Sex Roles*, *33*(3–4), 145–157. DOI: https://doi.org/10.1007/BF01544608

Davis Merritt, R., & Wells Harrison, T. (2006). Gender and ethnicity attributions to a gender-and ethnicity-unspecified individual: Is there a people = white male bias? *Sex Roles*, *54*, 787–797. DOI: https://doi.org/10.1007/s11199-006-9046-7

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. DOI: https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142. DOI: https://doi.org/10.1016/0749-596X(88)90070-8

Dell, G. S. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517–542. DOI: https://doi.org/10.1016/S0364-0213(99)00014-2

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*(1–3), 287–314. DOI: https://doi.org/10.1016/0010-0277(92)90046-K

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559. DOI: https://doi.org/10.3758/BF03195846

Files, J. A., Mayer, A. P., Ko, M. G., Friedrich, P., Jenkins, M., Bryan, M. J., Vegunta, S., Wittich, C. M., Lyle, M. A., Melikian, R., Duston, T., Chang, Y.-H. H., & Hayes, S. N. (2017). Speaker introductions at internal medicine grand rounds: Forms of address reveal gender bias. *Journal of Women's Health*, *26*(5). DOI: https://doi.org/10.1089/jwh.2016.6044

Flowers, A. (2015). *Unisex names data* [Data Set]. FiveThirtyEight. https://github.com/fivethirtyeight/data/tree/master/unisex-names

Garnham, A., Doehren, S., & Gygax, P. (2015). True gender ratios and stereotype rating norms. *Frontiers in Psychology*, *6*. DOI: https://doi.org/10.3389/fpsyg.2015.01023

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory & Cognition*, *30*(3), 439–446. DOI: https://doi.org/10.3758/BF03194944

Gastil, J. (1990). Generic pronouns and sexist language: The oxymoronic character of masculine generics. *Sex Roles*, *23*(11–12), 629–643. DOI: https://doi.org/10.1007/BF00289252

GLAAD. (2020). *GLAAD Media reference guide – transgender*. GLAAD. https://web.archive.org/web/20200522040917/https://www.glaad.org/reference/transgender

Gundel, J. K., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in Cognitive Science*, *4*(2), 249–268. DOI: https://doi.org/10.1111/j.1756-8765.2012.01184.x

Hamilton, M. C. (1988). Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex Roles*, *19*(11–12), 785–799. DOI: https://doi.org/10.1007/BF00288993

Hamilton, M. C. (1991). Masculine bias in the attribution of personhood: People = male, male = people. *Psychology of Women Quarterly*, *15*(3), 393–402. DOI: https://doi.org/10.1111/j.1471-6402.1991.tb00415.x

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(4), 824–843. DOI: https://doi.org/10.1037/0278-7393.20.4.824

Jescheniak, J. D., Schriefers, H., & Lemhöfer, K. (2014). Selection of freestanding and bound gender-marking morphemes in speech production: A review. *Language, Cognition and Neuroscience*, *29*(6), 684–694. DOI: https://doi.org/10.1080/01690965.2012.654645

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, *23*(3), 355–378. DOI: https://doi.org/10.1023/A:1023599719948

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75. DOI: https://doi.org/10.1017/S0140525X99001776

Lieberson, S., Dumais, S., & Baumann, S. (2000). The instability of androgynous names: The symbolic maintenance of gender boundaries. *American Journal of Sociology*, *105*(5), 1249–1287. DOI: https://doi.org/10.1086/210431

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*. DOI: https://doi.org/10.3389/fpsyg.2013.00226

McConnell-Ginet, S. (2014). Gender and its relation to sex: The myth of "natural" gender. In G. G. Corbett (Ed.), *The expression of gender* (pp. 3–38). DOI: https://doi.org/10.1515/9783110307337.3

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Öttl, A., Valdrova, J., Von Stockhausen, L., & Sczesny, S. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*, *46*(3), 841–871. DOI: https://doi.org/10.3758/s13428-013-0409-z

Moulton, J., Robinson, G. M., & Elias, C. (1978). Sex bias in language use: "Neutral" pronouns that aren't. *American Psychologist*, 1032–1036. DOI: https://doi.org/10.1037/0003-066X.33.11.1032

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2022). *Measuring sex, gender identity, and sexual orientation*. National Academies Press. DOI: https://doi.org/10.17226/26424

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory & Cognition*, *33*(6), 972–983. DOI: https://doi.org/10.3758/BF03193206

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

Pyykkönen, P., Hyönä, J., & Van Gompel, R. P. G. (2010). Activating gender stereotypes during online spoken language processing: Evidence from visual world eye tracking. *Experimental Psychology*, *57*(2), 126–133. DOI: https://doi.org/10.1027/1618-3169/a000016

Reynolds, D., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *The Quarterly Journal of Experimental Psychology*, *59*(5), 886–903. DOI: https://doi.org/10.1080/02724980543000088

Robertson, M. (2021). *Breaking, bending, and stretching the rules of singular* they [Conference talk]. 27th Annual Lavender Languages and Linguistics Conference.

Rodhe, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927. DOI: https://doi.org/10.1080/01690965.2013.854918

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107–142. DOI: https://doi.org/10.1016/0010-0277(92)90041-F

Rubin, R. B. (1981). Ideal traits and terms of address for male and female college professors. *Journal of Personality and Social Psychology*, *41*(5), 966–974. DOI: https://doi.org/10.1037/0022-3514.41.5.966

Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, *48*(1), 169–194. DOI: https://doi.org/10.1016/S0749-596X(02)00508-9

Schmitt, B. M., Meyer, A. S., & Levelt, W. J. M. (1999). Lexical access in the production of pronouns. *Cognition*, *69*(3), 313–335. DOI: https://doi.org/10.1016/S0010-0277(98)00073-0

Schriefers, H., Jescheniak, J. D., & Hantsch, A. (2002). Determiner selection in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 941–950. DOI: https://doi.org/10.1037/0278-7393.28.5.941

Silveira, J. (1980). Generic masculine words and thinking. *Women's Studies International Quarterly*, *3*(2–3), 165–178. DOI: https://doi.org/10.1016/S0148-0685(80)92113-2

Stewart, T. L., Berkvens, M., Engels, W. A. E. W., & Pass, J. A. (2003). Status and likability: Can the "mindful" woman have it all? *Journal of Applied Social Psychology*, *33*(10), 2040–2059. DOI: https://doi.org/10.1111/j.1559-1816.2003.tb01874.x

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562. DOI: https://doi.org/10.1016/S0749-596X(02)00536-3

Takiff, H. A., Sanchez, D. T., & Stewart, T. L. (2001). What's in a name? The status implications of students' terms of address for male and female professors. *Psychology of Women Quarterly*, *25*, 134–144. DOI: https://doi.org/10.1111/1471-6402.00015

United States Social Security Administration [USSSA]. (2019). *Top names over the last 100 years* [Data Set]. United States Social Security Administration. https://www.ssa.gov/oact/babynames/decades/century.html

United States Social Security Administration [USSSA]. (2020). *Beyond the top 1000 names* [Data Set]. United States Social Security Administration. https://www.ssa.gov/oact/babynames/limits.html

US Census Bureau. (2016). *Frequently occurring surnames from the 2010 census* [Data Set]. US Census Bureau. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Uscinski, J. E., & Goren, L. J. (2011). What's in a name? Coverage of Senator Hillary Clinton during the 2008 Democratic primary. *Political Research Quarterly*, *64*(4), 884–896. DOI: https://doi.org/10.1177/1065912910382302

Vincent, B. W. (2018). Studying trans: Recommendations for ethical recruitment and collaboration with transgender participants in academic research. *Psychology and Sexuality*, *9*(2), 102–116. DOI: https://doi.org/10.1080/19419899.2018.1434558

von der Malsburg, T., Poppels, T., & Levy, R. (2020). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 US and 2017 UK election. *Psychological Science*, *31*(2), 115–128. DOI: https://doi.org/10.1177/0956797619890619

Wang, M., & Schiller, N. O. (2019). A review on grammatical gender agreement in speech production. *Frontiers in Psychology*, *9*, 2754. DOI: https://doi.org/10.3389/fpsyg.2018.02754