Supplementary materials for: **Biased inferences about gender through names**

Bethany Gardner & Sarah Brown-Schmidt

OSF: 10.17605/OSF.IO/AYPU2
Github: https://github.com/bethanyhgardner/gender-bias-names

Contents

1 Norming study & name stimuli

Table S1. [A] First names used as stimuli, listed from most masculine to most feminine, with the mean and SD of each name's gender rating (1 as "definitely masculine" and 7 as "definitely feminine") from the norming study (N = 50). [B] Last names used as stimuli.

First Names

Name	*Mean*	*SD*
Matthew	1.21	0.74
Brian	1.24	0.75
James	1.28	0.61
Chris	2.12	1.27
Tommie	2.41	1.63
Emerson	2.61	1.44
Stevie	3.16	1.53
Quinn	3.75	1.60
Reese	3.87	1.67
Taylor	4.22	1.14
Riley	4.34	1.35
Jessie	4.39	1.27
Kerry	4.73	1.29
Blair	5.22	1.53
Jackie	5.34	1.13
Jodie	5.59	1.22
Elisha	5.86	1.83
Ashley	6.24	1.15
Mary	6.73	0.86
Rebecca	6.78	0.85
Emily	6.82	0.73

Last Names

Baker	Cooper	Smith
Bell	Green	Turner
Brooks	Hill	Walker
Brown	King	Ward
Campbell	Miller	White
Collins	Moore	Wright
Cook	Parker	Young

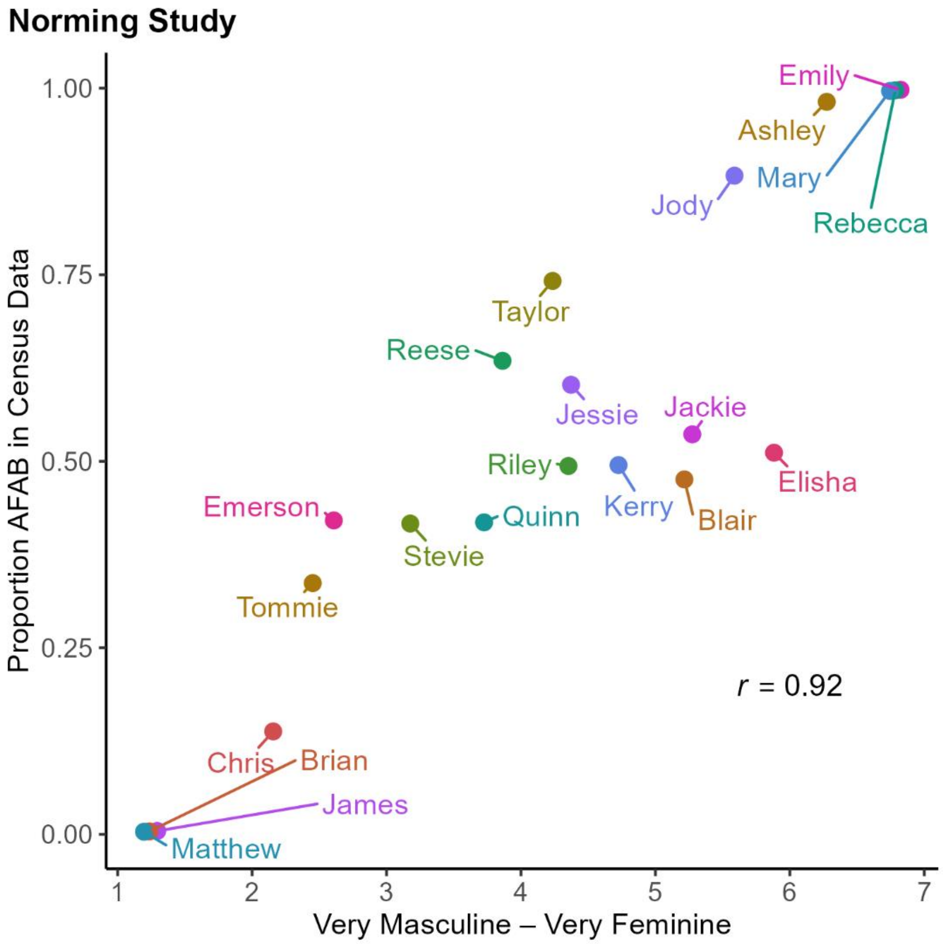

Figure S1. Correlation between gender ratings of first names from norming study and US census data (proportion of birth certificates for children assigned female at birth). AFAB = Assigned female at birth.

2 Experiment 1

2.1 Participant demographics and exclusions

Table S2. Experiment 1: Participant demographics. Gender was asked about using a free response box, with "did not provide" typically indicating an unrelated answer (e.g., age). Race has a different total, as participants could select multiple options.

Experiment 1: Participant Demographics

Age	**457**
18–24	47
25–34	201
35–44	109
45–54	56
55–64	33
65–74	10
75–84	1
Gender	**457**
Male	244
Female	196
Did not provide	15
Genderfluid	1
Nonbinary	1
Race	**485**
White	388
Asian	40
Black or African American	37
Other	10
American Indian or Alaska Native	9
Native Hawaiian or Pacific Islander	1
Education	**457**
High school graduate	45
Some college	106
2-year degree	68
4-year degree	195
Professional degree	40
Doctorate	3
Total Participants	**457**

Table S3. Experiment 1: Rationale for participant exclusions.

Experiment 1: Participant Exclusions

Included	**457**
Excluded	**113**
Indicated that they guessed the goal of the study was about names and gender	43
Completed study before, due to error in Mechanical Turk	36
Did not follow instructions for sentence completion task, e.g., entering *good* for question	23
Reported that they were not native English speakers	11
Total Collected	**570**

2.2 Gender rating centering

In the norming study, participants rated the first names on a 1–7 scale, with 7 as "very feminine." The 21 names selected as stimuli aren't perfectly centered on this scale, partially because androgynous names that lean feminine are much less common than androgynous names that lean masculine (Lieberson et al., 2000). A reviewer noted that mean centering Gender Rating on the items (*M* = 4.21), instead of on the original scale (= 4), may underestimate the size of the bias towards *he* responses. Here, we redo the model testing the effects of Condition and Gender Rating in the First and Full Name conditions (Table 3 in the main manuscript) with Gender Rating centered at 4 (Table S4). This produced a similar pattern of results as reported in the text, but with larger absolute values for the beta estimates for the intercept (-0.84 vs. -0.51) and the condition effect (0.57 vs. 0.53).

Table S4. Experiment 1: Model results for the effects of Condition and Gender Rating—now mean centered at 4—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 (Recentered)

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.843	0.123	-6.829	**<0.001**
Condition (First = -.5, Full = +.5)	0.568	0.247	2.305	0.021†
Gender Rating (Centered at 4; Masc -, Fem +)	1.593	0.073	21.966	**<0.001**
Condition × Gender Rating	-0.175	0.139	-1.257	0.209
Random Effects				
τ_{00} Participant	0.889			
τ_{00} Item	0.501			
N Participant	305			
N Item	83			
Observations	6372			

†Bonferroni corrected α = .0125

2.3 Excluding *other* responses

Other responses totaled 7.12% of all responses. While these responses were not numerous enough to analyze using inferential statistics, we describe them here to characterize the dataset. *Other* responses fell into several categories (Figure S2): Repeated Name responses repeated the name of the character and thus did not provide further information about the participant's inference about the character's gender (e.g., *Jordan woke up early to walk the dog. After making coffee...Jordan sat down to read the news*). Null Subject responses had no grammatical subject (e.g., *…sat down to read the news*). Other Subject responses talked about other characters or the environment (e.g., *...it started to rain*). Singular *They* responses used they/them pronouns to refer to the named character (e.g., *...they sat down to read the news*). Singular *They* responses were distinguished from uses of plural *they* (Other Subject) by the context of the prompt and were most common in the Last Name condition. Overall, repeated names were the most common response in this category.

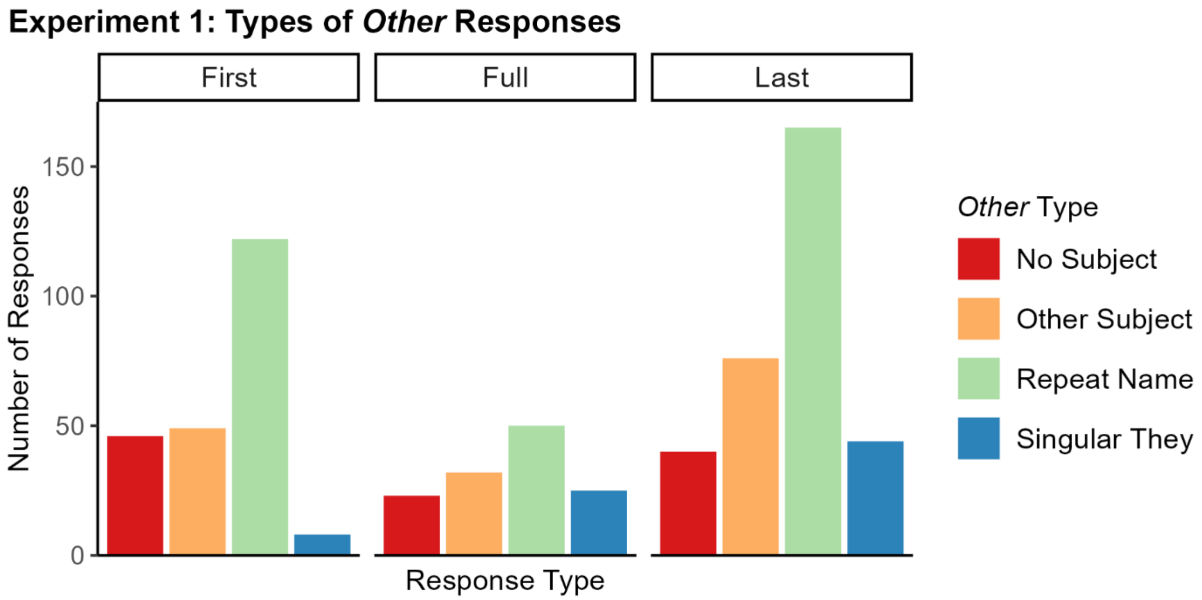

Figure S2. Experiment 1: Number of other *responses (N = 681) by type across conditions.*

To evaluate whether the results were impacted by grouping *other* responses with *he* responses, we repeated the primary analyses predicting the log odds of a *she* response (=1) as opposed to a *he* response (=0), with *other* responses excluded (Table S5, Table S6). The results revealed the same pattern of findings as reported in the main text, with the following exceptions: In the model testing the effects of Condition and Gender Rating (Table S6), the intercept (which measures response bias) and the difference between the First and Full Name conditions were not significant. Note that in the primary analyses, the difference between the First and Full Name conditions was marginally significant after correction for multiple comparisons in the Condition model (Table 2) and nonsignificant in the Condition and Gender Rating model (Table 3).

Table S5. Experiment 1: Model results for the effect of Condition on the likelihood of she *(=1) as opposed to* he *(=0) responses, with* other *responses excluded.*

Experiment 1: Condition (Without *Other* Responses)

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.130	0.343	-3.294	**<0.001**
Condition: Last (-.66) **vs. First** (+.33) **+ Full** (+.33)	2.990	0.784	3.816	**<0.001**
Condition: First (-.5) vs. Full (+.5)	0.555	0.782	0.709	0.478
Random Effects				
τ_{00} Participant	0.823			
τ_{00} Item	9.099			
N Participant	456			
N Item	104			
Observations	8883			

†Bonferroni corrected α = .0167

Table S6. Experiment 1: Model results for the effects of Condition and Gender Rating on the likelihood of she *(=1) as opposed to* he *(=0) responses in the First and Full Name conditions, with* other *responses excluded.*

Experiment 1: Condition and Gender Rating (Without *Other* Responses)

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.224	0.128	-1.751	0.080
Condition (First = -.5, Full = +.5)	0.410	0.256	1.602	0.109
Gender Rating (Mean-centered; Masc -, Fem +)	1.740	0.084	20.612	**<0.001**
Condition × Gender Rating	-0.251	0.161	-1.565	0.118
Random Effects				
τ_{00} Participant	0.581			
τ_{00} Item	0.627			
N Participant	304			
N Item	83			
Observations	6016			

[†]Bonferroni corrected α = .0125

2.4 Quadratic gender rating

An exploratory analysis added a quadratic effect of Gender Rating to evaluate the hypothesis that the effect of Gender Rating would be nonlinear, possibly with a larger *he* response bias at the midpoint of Gender Rating (androgynous names) than at the endpoints (strongly-gendered names) (Table S7). The quadratic effect was not significant, nor did it significantly interact with Condition, indicating that the magnitude of the Condition effect was not significantly magnified or tempered at the midpoint.

Table S7. Experiment 1: Model results for the effects of Condition, Gender Rating, and Quadratic 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 Quadratic Gender Rating

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.600	0.150	-3.992	**<0.001**
Condition (First = -.5, Full = +.5)	0.385	0.300	1.283	0.200
Gender Rating (Mean-centered; Masc -, Fem +)	1.597	0.072	22.214	**<0.001**
Quadratic Gender Rating	0.037	0.039	0.940	0.347
Condition × Gender Rating	-0.161	0.137	-1.169	0.242
Condition × Quadratic Gender Rating	0.060	0.078	0.773	0.440
Random Effects				
τ_{00} Participant	0.891			
τ_{00} Item	0.480			
N Participant	305			
N Item	83			
Observations	6372			

†*Bonferroni corrected α = .0083*

2.5 Participant gender

An exploratory analysis tested effects of participant gender, investigating if male participants showed a larger *he* response bias. Participants were asked about their gender in a free-response box, and female (N = 196), nonbinary (N = 1), and genderfluid (N = 1) participants were grouped together, as there were not enough responses to create more than 2 categories. Participants who did not indicate their gender (N = 15) were excluded from this analysis. Participant Gender was mean-center effects coded, comparing non-male participants (N = 198) to male participants (N = 244). Adding Participant Gender to the model testing effects of Condition (Table S8) showed that male participants were somewhat less likely to respond *she* across all three conditions ($\beta = -.26$, $z = -2.19$, $p < .05$). Adding Participant Gender to the model testing Condition and Gender Rating (Table S9) showed significant interactions with Condition ($\beta = .55$, $z = 1.99$, $p < .05$) and Gender Rating ($\beta = .15$, $z = 2.02$, $p < .05$). However, neither effect of Participant Gender remained significant after Bonferroni corrections for multiple comparisons.

Table S8. Experiment 1: Model results for the effects of Condition and Participant Gender (comparing non-male to male participants) on the likelihood of she *responses (=1) as opposed to* he *and* other *responses (=0).*

Experiment 1: Condition and Participant Gender

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.424	0.307	-4.637	**<0.001**
Condition: Last (-.66) **vs. First** (+.33) **+ Full** (+.33)	2.808	0.701	4.003	**<0.001**
Condition (First = -.5, Full = +.5)	0.586	0.699	0.839	0.402
Participant Gender (Non-Male = -.5, Male = +.5)	-0.264	0.121	-2.181	0.029[†]
Condition (Last vs. First + Full) × Participant Gender	0.396	0.270	1.466	0.143
Condition (First vs. Full) × Participant Gender	0.431	0.283	1.523	0.128
Random Effects				
$\tau_{00\ \text{Participant}}$	1.018			
$\tau_{00\ \text{Item}}$	7.202			
$N_{\text{Participant}}$	442			
N_{Item}	104			
Observations	9251			

[†] *Bonferroni corrected α = .0083*

Table S9. Experiment 1: Model results for the effects of Condition, Gender Rating, and Participant Gender (comparing non-male to male participants) 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, Gender Rating, and Participant Gender

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.531	0.119	-4.478	**<0.001**
Condition (First = -.5, Full = +.5)	0.515	0.237	2.169	0.030†
Gender Rating (Mean-centered; Fem +, Masc -)	1.596	0.072	22.251	**<0.001**
Participant Gender (Non-Male = -.5, Male = +.5)	-0.189	0.138	-1.370	0.171
Condition × Gender Rating	-0.170	0.137	-1.235	0.217
Condition × Participant Gender	0.550	0.277	1.990	0.047†
Gender Rating × Participant Gender	0.150	0.074	2.024	0.043†
Condition × Gender Rating × Participant Gender	-0.285	0.149	-1.919	0.055
Random Effects				
$\tau_{00\ \text{Participant}}$	0.880			
$\tau_{00\ \text{Item}}$	0.472			
$N_{\text{Participant}}$	295			
N_{Item}	83			
Observations	6164			

†*Bonferroni corrected α = .0063*

3 Experiment 2

3.1 Participant demographics and exclusions

Table S10. Experiment 2: Participant demographics. Gender was asked about using a free response box, with “did not provide” indicating an unrelated answer (e.g., age). *Race has a different total, as participants could select multiple options.*

Experiment 2: Participant Demographics	
Age	**1351**
18–24	137
25–34	619
35–44	344
45–54	157
55–64	76
65–74	17
75–84	1
Gender	**1351**
Male	694
Female	566
Did not provide	88
Nonbinary	2
Genderqueer	1
Race	**1396**
White	1051
Black or African American	191
Asian	98
Other	30
American Indian or Alaska Native	17
Native Hawaiian or Pacific Islander	9
Education	**1351**
High school graduate	180
Some college	276
2-year degree	166
4-year degree	572
Professional degree	146
Doctorate	5
Less than high school	6
Total Participants	**1351**

Table S11. Experiment 2: Rationale for participant exclusions.

Experiment 2: Participant Exclusions	
Included	**1351**
Excluded	**183**
Completed study before, due to error in Mechanical Turk	73
Did not follow instructions and answered every recall question with a name instead of a gender	54
Did not follow instructions and responded nonsensically, e.g., entering *good* for each question	40
Reported that they were not native English speakers	16
Total Collected	**1534**

3.2 Gender rating centering

As in Experiment 1, we reran the model testing the effects of Condition and Gender Rating in the First and Full Name conditions with Gender Rating mean centered relative to the scale (= 4) (Table S12) instead of relative to the items (= 4.22) (Table 6). This produced the same pattern of results, but with a larger absolute value for the beta estimate of the intercept (-0.35 vs. -0.18).

Table S12. Experiment 2: Model results for the effects of Condition and Gender Rating—now mean centered at 4—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 (Recentered)

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.351	0.060	-5.90	**<0.001**
Condition (First = -.5, Full = +.5)	-0.208	0.119	-1.75	0.079
Gender Rating (Centered at 4; Masc -, Fem +)	0.783	0.035	22.338	**<0.001**
Condition × Gender Rating	-0.066	0.069	-0.961	0.336
Random Effects				
τ_{00} Participant	0.114			
τ_{00} Item	0.141			
N Participant	903			
N Item	83			
Observations	6321			

†Bonferroni corrected α = .0125

3.3 Excluding *other* responses

The rate of *other* responses (3.72%) was lower than in Experiment 1, resulting in too few observations to analyze separately. Conducting the primary analyses excluding *other* responses (Table S13, Table S14) revealed a similar pattern of results, with the exception of the intercept in the model testing Condition and Gender Rating becoming marginally significant after Bonferroni correction (*β* = -.13, *z* = -2.19, p < .05).

Table S13. Experiment 2: Model results for the effect of Condition on the likelihood of responding female *(=1) as opposed to* male *(=0), with* other *responses excluded.*

Experiment 2: Condition (Without *Other* Responses)

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.785	0.150	-5.221	**<0.001**
Condition: Last (-.66) **vs. First** (+.33) **+ Full** (+.33)	1.922	0.342	5.616	**<0.001**
Condition: First (-.5) vs. Full (+.5)	-0.200	0.344	-0.582	0.560
Random Effects				
τ_{00} Participant	0.102			
τ_{00} Item	1.780			
N Participant	1322			
N Item	105			
Observations	9105			

†Bonferroni corrected α = .0167

Table S14. Experiment 2: Model results for the effects of Condition and Gender Rating on the likelihood of responding female *(=1) as opposed to* male *(=0) in the First and Full Name conditions, with* other *responses excluded.*

Experiment 2: Condition and Gender Rating (Without *Other* Responses)

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.126	0.058	-2.187	0.029†
Condition (First = -.5, Full = +.5)	-0.192	0.115	-1.669	0.095
Gender Rating (Mean-centered; Masc -, Fem +)	0.784	0.035	22.423	**<0.001**
Condition × Gender Rating	-0.061	0.069	-0.893	0.372
Random Effects				
τ_{00} Participant	0.029			
τ_{00} Item	0.139			
N Participant	897			
N Item	83			
Observations	6201			

†*Bonferroni corrected α = .0125*

3.4 Quadratic gender rating

An exploratory analysis included a quadratic effect of Gender Rating, again testing the hypothesis that the effect of Gender Rating would be different at the midpoint of Gender Rating (androgynous names) than at the endpoints (strongly-gendered names) (Table S15). The quadratic effect of Gender Rating was not significant, nor did it significantly interact with Condition, indicating that the magnitude of the Condition effect was not significantly magnified or tempered at the midpoint of Gender Rating.

Table S15. Experiment 2: Model results for the effects of Condition, Gender Rating, and Quadratic 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 Quadratic Gender Rating

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.139	0.081	-1.723	0.085
Condition (First = -.5, Full = +.5)	-0.258	0.161	-1.598	0.110
Gender Rating (Mean-centered; Masc -, Fem +)	0.779	0.036	21.895	**<0.001**
Quadratic Gender Rating	-0.013	0.020	-0.661	0.509
Condition × Gender Rating	-0.061	0.070	-0.878	0.380
Condition × Quadratic Gender Rating	0.011	0.039	0.286	0.775
Random Effects				
$\tau_{00\ \text{Participant}}$	0.114			
$\tau_{00\ \text{Item}}$	0.141			
$N_{\text{Participant}}$	903			
N_{Item}	83			
Observations	6321			

[†]Bonferroni corrected α = .0083

3.5 Participant gender

An exploratory analysis tested effects of participant gender to investigate if male participants showed a larger bias to recall the characters as male. Participants were asked about their gender in a free-response box. Female (N = 566), nonbinary (N = 2), and genderqueer (N = 1) participants were grouped together, as there were not enough responses to create more than 2 categories. Participants who did not indicate their gender (N = 88) were excluded from this analysis. Participant gender was mean-center effects coded, comparing non-male participants (N = 569) to male participants (N = 694). Adding participant gender to the model testing the effects of Condition (Table S16) showed that male participants were somewhat less likely to respond *female* than non-male participants across all three conditions ($\beta = -.13$, $z = -2.02$, $p < .05$), but this difference was not significant after correction for multiple comparisons. The interaction between Participant Gender and the Last vs. First + Full contrast was significant ($\beta = -.42$, $z = -2.93$, $p < .01$), such that the effect of Participant Gender was significant in the First and Full Name conditions ($\beta = -.26$, $z = -3.75$, $p < .001$), but not in the Last Name condition ($\beta = .15$, $z = 1.23$, $p = .22$). Adding Participant Gender to the model testing Condition and Gender Rating (Table S17) showed a significant effect of Participant Gender ($\beta = -.23$, $z = -3.34$, $p < .001$). Both results indicate that male participants were less likely than non-male participants to respond *female* in the First and Full Name conditions, but that there was no effect of Participant Gender in the Last Name condition. The interaction between Participant Gender and Gender Rating was significant ($\beta = -0.16$, $z = -2.664$, $p < .01$), such that the effect of Gender Rating was smaller for male participants than non-male participants.

Table S16. Experiment 2: Model results for the effects of Condition and Participant Gender (comparing non-male to male participants) on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 2: Condition and Participant Gender

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.851	0.153	-5.568	**<0.001**
Condition: Last (-.66) **vs. First** (+.33) **+ Full** (+.33)	2.023	0.347	5.824	**<0.001**
Condition (First = -.5, Full = +.5)	-0.200	0.349	-0.573	0.567
Participant Gender (Non-Male = -.5, Male = +.5)	-0.125	0.062	-2.018	0.044†
Condition (Last vs. First + Full) × Participant Gender	-0.418	0.142	-2.932	**0.003**
Condition (First vs. Full) × Participant Gender	0.092	0.140	0.658	0.511
Random Effects				
τ_{00} Participant	0.183			
τ_{00} Item	1.826			
N Participant	1263			
N Item	105			
Observations	8841			

†*Bonferroni corrected α = .0083*

Table S17. Experiment 2: Model results for the effects of Condition, Gender Rating, and Participant Gender (comparing non-male to male participants) 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, Gender Rating, and Participant Gender

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.160	0.062	-2.997	0.008†
Condition (First = -.5, Full = +.5)	-0.203	0.121	-1.680	0.093
Gender Rating (Mean-centered; Masc -, Fem +)	0.808	0.037	21.981	**<0.001**
Participant Gender (Non-Male = -.5, Male = +.5)	-0.228	0.068	-3.345	**0.001**
Condition × Gender Rating	-0.058	0.072	-0.811	0.418
Condition × Participant Gender	0.128	0.136	0.935	0.418
Gender Rating × Participant Gender	-0.149	0.045	-3.320	**0.001**
Condition × Gender Rating × Participant Gender	-0.116	0.090	-1.293	0.196
Random Effects				
$\tau_{00\ \text{Participant}}$	0.099			
$\tau_{00\ \text{Item}}$	0.149			
$N_{\text{Participant}}$	840			
N_{Item}	83			
Observations	5880			

†*Bonferroni corrected α = .0063*

4 Experiment 3

4.1 Participant demographics and exclusions

Table S18. Experiment 3: Participant demographics. Gender was asked about using a free response box, with "did not provide" indicating an unrelated answer (e.g., age). Race has a different total, as participants could select multiple options.

Experiment 3: Participant Demographics	
Age	**1272**
18–24	105
25–34	477
35–44	343
45–54	163
55–64	131
65–74	47
75–84	5
85 or older	1
Gender	**1272**
Female	638
Male	514
Did not provide	115
Nonbinary	2
Agender	1
Asexual	1
Prefer not to say	1
Race	**1319**
White	1000
Black or African American	154
Asian	98
Other	39
American Indian or Alaska Native	21
Native Hawaiian or Pacific Islander	7
Education	**1272**
High school graduate	111
Some college	276
2-year degree	137
4-year degree	550
Professional degree	172
Doctorate	22
Less than high school	4
Total Participants	**1272**

Table S19. Experiment 3: Rationale for participant exclusions.

Experiment 3: Participant Exclusions	
Included	**1272**
Excluded	**403**
Did not follow instructions, entering nonsense text or a repetition of the prompt	316
Completed study before, due to error in Mechanical Turk	72
Reported that they were not native English speakers	15
Total Collected	**1675**

4.2 Gender rating centering

We reran the primary analysis with Gender Rating mean centered relative to the scale (= 4) (Table S20) instead of relative to the items (= 4.21) (Table 8). As before, this produced the same pattern of results, but with a larger absolute value for the beta estimate of the intercept (-1.76 vs. -1.52) and slightly different estimates for the Condition effects.

Table S20. Experiment 3: Model results for the effects of Condition and Gender Rating—now mean centered at 4—on the likelihood of she *responses (=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition and Gender Rating (Recentered)

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.761	0.105	-16.728	**<0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.132	0.097	1.358	0.174
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.103	0.123	0.837	0.403
Gender Rating (Centered at 4; 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				
τ_{00} Participant	0.793			
τ_{00} Item	0.421			
N Participant	1272			
N Item	63			
Observations	8904			

†*Bonferroni corrected α = .0083*

4.3 Excluding *other* responses

Because the sentence completion prompts were less constrained than in Experiment 1, *other* responses represented a larger proportion of the data (31%). To keep analyses parallel between experiments and avoid post-hoc changes in analysis plans, we conducted the primary analysis on the log odds of *she* as opposed to *he* and *other* responses and report the analysis excluding *other* responses as a supplementary analysis (Table S21). When excluding *other* responses, 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. This follows the pattern observed in Experiments 1 and 2. The intercept ($\beta = -0.42$, $z = -3.42$, p < .001) and the interaction between Condition and Gender Rating ($\beta = 0.42$, $z = 5.46$, p < .001) both remained significant.

Table S21. Experiment 3: Model results for the effects of Condition and Gender Rating on the likelihood of she *(=1) as opposed to* he *(=0) responses, with* other *responses excluded.*

Experiment 3: Condition and Gender Rating (Without *Other* Responses)

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.424	0.124	-3.423	**0.001**
Condition: Last (-.6) **vs. First** (+.4) **+ Full** (+.4)	0.257	0.098	2.627	**0.009**
Condition: First (-.48) vs. Full (+.52); Last (.02)	-0.015	0.128	-0.114	0.910
Gender Rating (Mean-centered; Masc -, Fem +)	1.677	0.084	20.034	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.420	0.077	5.455	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.149	0.112	-1.330	0.183
Random Effects				
$\tau_{00\ \text{Participant}}$	0.539			
$\tau_{00\ \text{Item}}$	0.681			
$N_{\text{Participant}}$	1223			
N_{Item}	63			
Observations	6137			

†Bonferroni corrected α = .0083

4.4 Quadratic gender rating

An exploratory analysis included a quadratic effect of Gender Rating, again testing the hypothesis that the effect of Gender Rating would be different at the midpoint of Gender Rating (androgynous names) than at the endpoints (strongly-gendered names). The quadratic effect of Gender Rating was significant (β = -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 than at the end points. Additionally, the interaction between the quadratic effect of Gender Rating and the Last vs. First + Full contrast was significant (β = -0.10, z = -3.24, p < .001), such that the quadratic effect of Gender Rating was significant in the First and Full Name conditions (β = -0.15, z = -4.28, p < .001) but not in the Last Name condition (β = -0.06, z = -1.67, p = .09). This analysis also indicated a significant Condition effect for the Last vs. First + Full contrast (β = 0.24, z = 3.00, p < .01), such that participants were more likely produce *she* in the First and Full Name conditions compared to the Last Name condition.

Table S22. Experiment 3: Model results for the effects of Condition, Gender Rating, and Quadratic Gender Rating on the likelihood of she *responses (=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition and Quadratic Gender Rating

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.096	0.111	-9.876	**<0.001**
Condition: Last (-.6) **vs. First** (+.4) **+ Full** (+.4)	0.238	0.079	2.998	**0.003**
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.056	0.100	0.559	0.576
Gender Rating (Mean-centered; Masc -, Fem +)	1.070	0.056	19.263	**<0.001**
Quadratic Gender Rating	-0.114	0.031	-3.667	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.222	0.061	3.627	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.113	0.088	-1.280	0.200
Condition (Last vs. First + Full) × Quadratic Gender Rating	-0.096	0.030	-3.238	**0.001**
Condition (First vs. Full) × Quadratic Gender Rating	0.039	0.042	0.924	0.355
Random Effects				
$\tau_{00\ \text{Item}}$	0.300			
N_{Item}	63			
Observations	8904			

†*Bonferroni corrected α = .0056*

4.5 Participant gender

An exploratory analysis tested effects of participant gender (Table S23). Participants were asked about their gender in a free-response box. Female (N = 638), nonbinary (N = 2), agender (N = 1), and asexual (N = 1) participants were grouped together, as there were not enough responses to create additional categories; participants who did not indicate their gender (N = 117) were excluded from this analysis. Participant gender was mean-center effects coded, comparing non-male participants (N = 642) to male participants (N = 514). *Other* responses were included in the analysis. Results showed that male participants were less likely to use *she* than non-male participants across all three conditions ($\beta = -.33$, $z = -3.53$, $p < .001$). No interactions of Participant Gender with Condition or Gender Rating were significant.

Table S23. Experiment 3: Model results for the effects of Condition, Gender Rating, and Participant Gender (comparing non-male to male participants) on the likelihood of she *responses (=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition, Gender Rating, and Participant Gender

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.580	0.105	-14.983	**<0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.195	0.098	1.985	0.047†
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.136	0.122	1.116	0.265
Gender Rating (Mean-centered; Fem +, Masc -)	1.148	0.063	18.248	**<0.001**
Participant Gender (Non-Male = -.5, Male = +.5)	-0.339	0.096	-3.532	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.113	0.053	2.153	0.031†
Condition (First vs. Full) × Gender Rating	-0.079	0.067	-1.188	0.235
Condition (Last vs. First + Full) × Participant Gender	0.120	0.197	0.610	0.542
Condition (First vs. Full) × Participant Gender	0.047	0.243	0.192	0.848
Gender Rating × Participant Gender	-0.017	0.052	-0.335	0.737
Condition (Last vs. First + Full) × Gender Rating × Participant Gender	0.094	0.105	0.896	0.370
Condition (First vs. Full) × Gender Rating × Participant Gender	-0.046	0.133	-0.345	0.730
Random Effects				
$\tau_{00\ \text{Participant}}$	0.753			
$\tau_{00\ \text{Item}}$	0.448			
$N_{\text{Participant}}$	1156			
N_{Item}	63			
Observations	8092			

†*Bonferroni corrected α = .0042*

4.6 Character ratings

The Accomplishment, Likeability, and Importance ratings were measured on a scale of 1–7, with 1 indicating a more favorable rating. These ratings were clustered near the positive ends of the scales (Accomplishment: *M* = 2.10, *SD* = 1.23; Likeability: *M* = 2.23, *SD* = 1.28; Importance: *M* = 2.55, *SD* = 1.34). To simplify model interpretation, this scale was reverse-coded, so that higher numbers indicated more favorable ratings, and all 3 ratings were mean-centered. Likeability (Table S24), Accomplishment (Table S25), and Importance (Table S26) were added to 3 separate models predicting *she* as opposed to *he* and *other* responses. The main effect of Likeability was significant (β = 0.08, z = 2.86, p < .01), such that more Likeable characters were more likely to be referred to with *she*. No other main effects of character ratings or their interactions with Condition or Gender Rating were significant after Bonferroni corrections for multiple comparisons. An alternative analysis is to include all 3 character ratings in the same model with Condition and Gender Rating. The maximal model that converged resulted in a model with fixed effects up through a subset of the 3-way and 4-way interactions, and no random intercepts by item or by participant (Table S27). Again, the only significant effect of character rating was *she* responses increasing with Likeability (β = 0.11, z = 3.43, p < .001).

Table S24. Experiment 3: Model results for the effects of Condition, Gender Rating, and Character Likeability rating on the likelihood of she *(=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition, Gender Rating, and Character Likeability

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.375	0.089	-15.400	**<0.001**
Gender Rating (Mean-centered; Masc -, Fem +)	1.032	0.055	18.803	**<0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.139	0.071	1.950	0.051
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.069	0.090	0.765	0.444
Likeability (Mean-centered; Less -, More +)	0.080	0.028	2.854	**0.004**
Condition (Last vs. First + Full) × Gender Rating	0.088	0.045	1.943	0.052
Condition (First vs. Full) × Gender Rating	-0.064	0.058	-1.097	0.273
Condition (Last vs. First + Full) × Likeability	-0.064	0.057	-1.125	0.261
Condition (First vs. Full) × Likeability	0.070	0.069	1.010	0.313
Gender Rating × Likeability	0.027	0.017	1.554	0.120
Condition (Last vs. First + Full) × Gender Rating × Likeability	0.062	0.035	1.777	0.076
Condition (First vs. Full) × Gender Rating × Likeability	-0.036	0.044	-0.811	0.417
Random Effects				
$\tau_{00\ \text{Item}}$	0.352			
N_{Item}	63			
Observations	8904			

†*Bonferroni corrected α = .0042*

Table S25. Experiment 3: Model results for the effects of Condition, Gender Rating, and Character Accomplishment rating on the likelihood of she *(=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition, Gender Rating, and Character Accomplishment

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.372	0.090	-15.261	**<0.001**
Gender Rating (Mean-centered; Masc -, Fem +)	1.034	0.055	18.721	**<0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.139	0.071	1.969	$0.049^{\dagger}$
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.074	0.089	0.824	0.410
Accomplishment (Mean-centered; Less -, More +)	0.072	0.028	2.556	$0.011^{\dagger}$
Condition (Last vs. First + Full) × Gender Rating	0.090	0.045	1.993	$0.046^{\dagger}$
Condition (First vs. Full) × Gender Rating	-0.057	0.058	-0.978	0.328
Condition (Last vs. First+ Full) × Accomplishment	-0.084	0.059	-1.431	0.152
Condition (First vs. Full) × Accomplishment	-0.070	0.070	-0.999	0.318
Gender Rating × Accomplishment	0.030	0.017	1.706	0.088
Condition (Last vs. First + Full) × Gender Rating × Accomplishment	0.085	0.036	2.355	$0.019^{\dagger}$
Condition (First vs. Full) × Gender Rating × Accomplishment	0.003	0.044	0.064	0.949
Random Effects				
$\tau_{00\ \text{Item}}$	0.360			
N_{Item}	63			
Observations	8904			

†*Bonferroni corrected α = .0042*

Table S26. Experiment 3: Model results for the effects of Condition, Gender Rating, and Character Importance rating on the likelihood of she *(=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition, Gender Rating, and Character Importance

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.375	0.090	-15.296	**<0.001**
Gender Rating (Mean-centered; Masc -, Fem +)	1.033	0.055	18.713	**<0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.137	0.071	1.930	0.054
Condition: First (-.48) vs. Full (+.52); Last (.02)	0.076	0.089	0.853	0.394
Importance (Mean-centered; Less -, More +)	0.042	0.026	1.597	0.110
Condition (Last vs. First + Full) × Gender Rating	0.087	0.045	1.930	0.054
Condition (First vs. Full) × Gender Rating	-0.055	0.058	-0.941	0.346
Gender Rating × Importance	0.003	0.017	0.208	0.836
Condition (Last vs. First + Full) × Importance	-0.060	0.052	-1.143	0.253
Condition (First vs. Full) × Importance	0.078	0.064	1.232	0.218
Condition (Last vs. First + Full) × Gender Rating × Importance	0.066	0.033	1.967	0.049†
Condition (First vs. Full) × Gender Rating × Importance	-0.029	0.041	-0.693	0.488
Random Effects				
$\tau_{00\ \mathrm{Item}}$	0.359			
N_{Item}	63			
Observations	8904			

†*Bonferroni corrected α = .0042*

Table S27. Experiment 3: Model results for the effects of Condition, Gender Rating, Character Accomplishment, Character Importance, and Character Likeability rating on the likelihood of she *(=1) as opposed to* he *and* other *responses (=0).*

Experiment 3: Condition, Gender Rating, and Character Ratings

	Refer to using she			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-1.18	0.04	-30.97	**<0.001**
Gender Rating (Mean-centered; Masc -, Fem +)	0.90	0.02	42.33	**<0.001**
Condition: Last (-.6) **vs. First** (+.4) **+ Full** (+.4)	0.19	0.07	2.77	0.006†
Condition: First (-.48) vs. Full (+.52); Last (.02)	-0.06	0.09	-0.70	0.486
Accomplishment (Mean-centered; Less -, More +)	0.05	0.03	1.54	0.123
Importance (Mean-centered; Less -, More +)	-0.07	0.03	-2.38	0.017†
Likeability (Mean-centered; Less -, More +)	0.11	0.03	3.43	**0.001**
Condition (Last vs. First + Full) × Accomplishment	0.04	0.07	0.60	0.546
Condition (First vs. Full) × Accomplishment	-0.24	0.09	-2.66	0.008†
Condition (Last vs. First + Full) × Importance	-0.04	0.06	-0.60	0.551
Condition (First vs. Full) × Importance	0.10	0.08	1.29	0.198
Condition (Last vs. First + Full) × Likeability	-0.05	0.07	-0.80	0.424
Condition (First vs. Full) × Likeability	0.02	0.08	0.27	0.791
Accomplishment × Importance	0.00	0.02	0.02	0.987
Accomplishment × Likeability	0.01	0.02	0.37	0.710
Importance × Likeability	0.01	0.02	0.26	0.793
Condition (Last vs. First + Full) × Acc. × Importance	0.02	0.05	0.49	0.622
Condition (First vs. Full) × Acc. × Importance	0.03	0.06	0.51	0.613
Condition (Last vs. First + Full) × Acc. × Likeability	0.01	0.04	0.18	0.858
Condition (First vs. Full) × Acc. × Likeability	0.02	0.06	0.30	0.766
Condition (Last vs. First + Full) × Imp. × Likeability	0.01	0.05	0.18	0.860
Condition (First vs. Full) × Imp. × Likeability	0.08	0.06	1.28	0.200
Accomplishment × Importance × Likeability	0.01	0.01	1.21	0.227
Condition (Last vs. First + Full) × Acc.* Imp × Likeability	0.02	0.02	1.43	0.153
Condition (First vs. Full) × Acc. × Imp. × Likeability	0.06	0.02	2.53	0.012†
Observations	8904			

†Bonferroni corrected α =0.002

5 Experiment 4

5.1 Participant demographics and exclusions

Table S28. Experiment 4: Participant demographics. Gender was asked about using a free response box, with "did not provide" indicating an unrelated answer (e.g., age). Race has a different total, as participants could select multiple options.

Experiment 4: Participant Demographics

Age	**1253**
18–24	101
25–34	515
35–44	341
45–54	169
55–64	89
65–74	34
75–84	4
Gender	**1253**
Male	602
Female	555
Did not provide	91
Nonbinary	3
Transgender female	1
Transgender male	1
Race	**1295**
White	990
Black or African American	146
Asian	106
Other	24
American Indian or Alaska Native	21
Native Hawaiian or Pacific Islander	8
Education	**1253**
High school graduate	122
Some college	292
2-year degree	138
4-year degree	528
Professional degree	150
Doctorate	15
Less than high school	8
Total Participants	**1253**

Table S29. Experiment 4: Rationale for participant exclusions.

Experiment 4: Participant Exclusions	
Included	**1253**
Excluded	**108**
Completed study before, due to error in Mechanical Turk	46
Did not follow instructions and answered every recall question with a name instead of a gender	11
Did not follow instructions and responded nonsensically, e.g., entering *good* for each question	32
Reported that they were not native English speakers	19
Total Collected	**1361**

5.2 Gender rating centering

We reran the primary analysis with Gender Rating mean centered relative to the scale (= 4) (Table S30) instead of relative to the items (= 4.21) (Table 10). As before, this produced a similar pattern of results, but with a larger absolute value for the beta estimate of the intercept (-0.41 vs. -0.26) and slightly different estimates for the Condition effects.

Table S30. Experiment 4: Model results for the effects of Condition and Gender Rating—now mean centered at 4—on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition and Gender Rating (Recentered)

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.413	0.082	-5.017	**<0.001**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.099	0.063	1.576	0.115
Condition: First (-.49) vs. Full (+.51)	0.090	0.074	1.212	0.226
Gender Rating (Mean-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				
τ_{00} Participant	0.201			
τ_{00} Item	0.360			
$N_{\text{Participant}}$	1253			
N_{Item}	63			
Observations	8771			

[†] *Bonferroni corrected α = .0083*

5.3 Excluding *other* responses

As in Experiment 2, the rate of *other* responses (2.99%) was too small to analyze separately. Conducting the primary analysis (Table 10) with *other* responses excluded (Table S31) showed a similar pattern of results.

Table S31. Experiment 4: Model results for the effects of Condition and Gender Rating on the likelihood of female *responses (=1) as opposed to* male *responses (=0), with* other *responses excluded.*

Experiment 4: Condition and Gender Rating (Without *Other* Responses)

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.164	0.082	-2.002	$0.045^{\dagger}$
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.135	0.058	2.337	$0.019^{\dagger}$
Condition: First (-.49) vs. Full (+.51)	0.113	0.068	1.653	0.098
Gender Rating (Mean-centered; Masc -, Fem +)	0.770	0.046	16.554	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.137	0.035	3.890	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.092	0.043	-2.130	$0.033^{\dagger}$
Random Effects				
$\tau_{00\ \text{Participant}}$	0.050			
$\tau_{00\ \text{Item}}$	0.369			
$N_{\text{Participant}}$	1232			
N_{Item}	63			
Observations	8509			

†*Bonferroni corrected α = .0083*

5.4 Quadratic gender rating

An exploratory analysis included a quadratic effect of Gender Rating, again testing the hypothesis that the effect of Gender Rating would be different at the midpoint of Gender Rating (androgynous names) than at the endpoints (strongly-gendered names) (Table S32). The quadratic effect of Gender Rating was not significant, nor did it significantly interact with Condition, indicating that the magnitude of the Condition effect was not significantly magnified or tempered at the midpoint of Gender Rating.

Table S32. Experiment 4: Model results for the effects of Condition, Gender Rating, and Quadratic Gender Rating on the likelihood of female *responses (=1) opposed to* male *and* other *responses (=0).*

Experiment 4: Condition and Quadratic Gender Rating

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.369	0.116	-3.189	**0.001**
Condition: Last (-.6) vs. First (+.4) + Full (+.4)	0.161	0.080	2.006	0.045†
Condition: First (-.49) vs. Full (+.51)	-0.076	0.093	-0.825	0.409
Gender Rating (Mean-centered; Masc -, Fem +)	0.780	0.046	16.814	**<0.001**
Quadratic Gender Rating	0.034	0.026	1.306	0.192
Condition (Last vs. First + Full) × Gender Rating	0.132	0.035	3.800	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.092	0.043	-2.157	0.031†
Condition (Last vs. First + Full) × Quadratic Gender Rating	-0.014	0.019	-0.737	0.461
Condition (First vs. Full) × Quadratic Gender Rating	0.060	0.024	2.539	0.011†
Random Effects				
τ_{00} Participant	0.204			
τ_{00} Item	0.348			
$N_{\text{Participant}}$	1253			
N_{Item}	63			
Observations	8771			

†*Bonferroni corrected α = .0056*

5.5 Participant gender

An exploratory analysis tested effects of participant gender (Table S33). Participants were asked about their gender in a free-response box. Female (N = 556) and nonbinary (N = 3) participants were grouped together, as there were not enough responses to create three categories. Participants who described themselves as transgender female (N = 1) and transgender male (N = 1) were grouped correspondingly. Participants who did not indicate their gender (N = 91) were excluded from this analysis. Participant gender was mean-center effects coded, comparing non-male participants (N = 559) to male participants (N = 603). Male participants were less likely to recall the character as female across all three conditions (β = -.20, z = -3.27, p < .001). No interactions of Participant Gender with Condition or Gender Rating were significant.

Table S33. Experiment 4: Model results for effects of Condition, Gender Rating, and Participant Gender (comparing non-male to male participants) on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition, Gender Rating, and Participant Gender

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.250	0.083	-3.025	**0.002**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.150	0.064	2.357	0.018†
Condition: First (-.49) vs. Full (+.51)	0.078	0.075	1.040	0.298
Gender Rating (Mean-centered; Masc -, Fem +)	0.765	0.047	16.409	**<0.001**
Participant Gender (Non-Male = -.5, Male = .5)	-0.199	0.061	-3.270	**0.001**
Condition (Last vs. First + Full) × Gender Rating	0.096	0.036	2.662	0.008†
Condition (First vs. Full) × Gender Rating	-0.099	0.043	-2.269	0.023†
Condition (Last vs. First + Full) × Participant Gender	-0.024	0.128	-0.187	0.852
Condition (First vs. Full) × Participant Gender	-0.144	0.149	-0.967	0.333
Gender Rating × Participant Gender	-0.020	0.035	-0.570	0.569
Condition (Last vs. First + Full) × Gender Rating × Participant Gender	0.041	0.073	0.564	0.573
Condition (First vs. Full) × Gender Rating × Participant Gender	-0.053	0.087	-0.606	0.545
Random Effects				
$\tau_{00\ \text{Participant}}$	0.182			
$\tau_{00\ \text{Item}}$	0.367			
$N_{\text{Participant}}$	1162			
N_{Item}	63			
Observations	8134			

†*Bonferroni corrected α = .0042*

5.6 Character ratings

As in Experiment 3, the ratings for Likeability, Accomplishment, and Importance were largely clustered at the positive end of the scale (Accomplishment: *M* = 2.45, *SD* = 1.37; Likeability: *M* = 2.56, *SD* = 1.39; Importance: *M* = 2.85, *SD* = 1.46). To simplify model interpretation, this scale was reverse-coded, so that higher numbers indicated more favorable ratings, and all 3 ratings were mean-centered. Likeability (Table S34), Accomplishment (Table S35), and Importance (Table S36) were added to 3 separate models predicting *female* as opposed to *male* and *other* responses. In addition to the effects of Condition and Gender Rating reported in the main text, and after Bonferroni corrections for multiple comparisons, a main effect of Likeability was observed ($\beta = 0.09$, $z = 4.51$, p < .001), such that more likeable characters were more likely to be recalled as female. In addition, significant interactions of Gender Rating with Likeability ($\beta = 0.10$, $z = 8.72$, p < .001), Accomplishment ($\beta = 0.11$, $z = 9.22$, p < .001), and Importance ($\beta = 0.08$, $z = 6.70$, p < .001) were due to the effects of Gender Rating increasing as characters were rated more favorably. An alternative analysis includes all 3 character ratings in the same model with Condition and Gender Rating. Like in Experiment 3, the maximal model that converged included all two-way interactions between fixed effects, some three- and four-way interactions, and no random intercepts by item or by participant (Table S37). The main effects of Likeability ($\beta = 0.18$, $z = 6.46$, p < .001) and Importance ($\beta = -0.12$, $z = -4.91$, p < .001) were significant, such that characters rated more likeable and less important were less likely to be recalled as female. The four-way interactions between the Last vs. First + Full Name conditions, Gender Rating, Accomplishment, and Importance and between Last vs. First + Full Name conditions, Gender Rating, Accomplishment, and Likeability were also significant, but are difficult to interpret meaningfully.

Table S34. Experiment 4: Model results for the effects of Condition, Gender Rating, and Character Likeability rating on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition, Gender Rating, and Character Likeability

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.244	0.078	-3.118	**0.002**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.119	0.055	2.155	0.031[†]
Condition: First (-.49) vs. Full (+.51)	0.067	0.065	1.026	0.305
Gender Rating (Mean-centered; Masc -, Fem +)	0.752	0.044	16.896	**<0.001**
Likeability (Mean-centered; Less -, More +)	0.087	0.019	4.510	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.125	0.035	3.549	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.099	0.043	-2.312	0.021[†]
Condition (Last vs. First + Full) × Likeability	0.045	0.038	1.169	0.242
Condition (First vs. Full) × Likeability	0.091	0.046	1.975	0.048[†]
Gender Rating × Likeability	0.103	0.012	8.719	**<0.001**
Condition (Last vs. First + Full) × Gender Rating × Likeability	0.019	0.023	0.820	0.412
Condition (First vs. Full) × Gender Rating × Likeability	0.002	0.029	0.063	0.950
Random Effects				
τ_{00} Item	0.337			
N Item	63			
Observations	8771			

[†] *Bonferroni corrected α = .0042*

Table S35. Experiment 4: Model results for the effects of Condition, Gender Rating, and Character Accomplishment rating on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition, Gender Rating, and Character Accomplishment

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.243	0.079	-3.096	**0.002**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.113	0.055	2.046	0.041[†]
Condition: First (-.49) vs. Full (+.51)	0.071	0.065	1.087	0.277
Gender Rating (Mean-centered; Masc -, Fem +)	0.749	0.045	16.815	**<0.001**
Accomplishment (Mean-centered; Less -, More +)	0.047	0.019	2.472	0.013[†]
Condition (Last vs. First + Full) × Gender Rating	0.126	0.035	3.585	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.101	0.043	-2.342	0.019[†]
Condition (Last vs. First + Full) × Accomplishment	0.012	0.038	0.321	0.748
Condition (First vs. Full) × Accomplishment	0.080	0.047	1.701	0.089
Gender Rating × Accomplishment	0.108	0.012	9.224	**<0.001**
Condition (Last vs. First + Full) × Gender Rating × Accomplishment	0.056	0.023	2.408	0.016[†]
Condition (First vs. Full) × Gender Rating × Accomplishment	-0.009	0.029	-0.318	0.750
Random Effects				
τ_{00} Item	0.338			
N Item	63			
Observations	8771			

[†] *Bonferroni corrected α = .0042*

Table S36. Experiment 4: Model results for the effects of Condition, Gender Rating, and Character Importance rating on the likelihood of female *responses (=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition, Gender Rating, and Character Importance

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.249	0.082	-3.052	**0.002**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.125	0.062	2.021	0.043[†]
Condition: First (-.49) vs. Full (+.51)	0.080	0.073	1.097	0.273
Gender Rating (Mean-centered; Masc -, Fem +)	0.772	0.046	16.824	**<0.001**
Importance (Mean-centered; Less -, More +)	0.010	0.020	0.487	0.626
Condition (Last vs. First + Full) × Gender Rating	0.133	0.035	3.810	**<0.001**
Condition (First vs. Full) × Gender Rating	-0.098	0.043	-2.304	0.021[†]
Condition (Last vs. First + Full) × Importance	0.010	0.039	0.252	0.801
Condition (First vs. Full) × Importance	0.077	0.047	1.647	0.100
Gender Rating × Importance	0.078	0.012	6.703	**<0.001**
Condition (Last vs. First + Full) × Gender Rating × Importance	0.016	0.023	0.694	0.487
Condition (First vs. Full) × Gender Rating × Importance	-0.008	0.028	-0.292	0.770
Random Effects				
$\tau_{00\ \text{Participant}}$	0.205			
$\tau_{00\ \text{Item}}$	0.358			
N_{Item}	63			
$N_{\text{Participant}}$	1253			
Observations	8771			

[†] *Bonferroni corrected α = .0042*

Table S37. Experiment 4: Model results for the effects of Condition, Gender Rating, Character Accomplishment, Character Importance, and Character Likeability rating on the likelihood of female *(=1) as opposed to* male *and* other *responses (=0).*

Experiment 4: Condition, Gender Rating, and Character Ratings`

	Recall as female			
Predictors	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	-0.24	0.03	-7.96	**<0.001**
Gender Rating (Mean-centered; Masc -, Fem +)	0.79	0.02	37.90	**<0.001**
Condition: Last (-.67) vs. First (+.33) + Full (+.33)	0.05	0.06	0.88	0.379
Condition: First (-.49) vs. Full (+0.51)	0.12	0.07	1.66	0.096
Accomplishment (Mean-centered; Less -, More +)	0.02	0.03	0.66	0.512
Importance (Mean-centered; Less -, More +)	-0.12	0.03	-4.91	**<0.001**
Likeability (Mean-centered; Less -, More +)	0.18	0.03	6.46	**<0.001**
Condition (Last vs. First + Full) × Gender Rating	0.12	0.04	2.89	0.004†
Condition (First vs. Full) × Gender Rating	-0.12	0.05	-2.37	0.018†
Accomplishment × Gender Rating	0.05	0.02	2.90	0.004†
Importance × Gender Rating	-0.01	0.02	-0.43	0.669
Likeability × Gender Rating	0.03	0.02	1.39	0.166
Condition (Last vs. First + Full) × Accomplishment	0.01	0.06	0.19	0.850
Condition (First vs. Full) × Accomplishment	-0.01	0.06	-0.15	0.883
Condition (Last vs. First + Full) × Importance	-0.01	0.05	-0.23	0.816
Condition (First vs. Full) × Importance	0.04	0.06	0.62	0.538
Condition (Last vs. First + Full) × Likeability	0.07	0.06	1.31	0.191
Condition (First vs. Full) × Likeability	0.03	0.07	0.42	0.673
Accomplishment × Importance	0.02	0.02	1.39	0.166
Accomplishment × Likeability	0.00	0.02	0.04	0.966
Importance × Likeability	-0.02	0.02	-1.27	0.205
Gender Rating × Accomplishment × Importance	0.01	0.01	0.72	0.473
Gender Rating × Accomplishment × Likeability	-0.02	0.01	-2.10	0.036†
Gender Rating × Importance × Likeability	-0.02	0.01	-2.08	0.037†
Condition (Last vs. First + Full) × Gender Rating × Acc.	0.07	0.04	1.91	0.056
Condition (First vs. Full) × Gender Rating × Accomplishment	0.01	0.05	0.25	0.801
Condition (Last vs. First + Full) × Gender Rating × Importance	0.02	0.04	0.42	0.678
Condition (First vs. Full) × Gender Rating × Importance	-0.02	0.04	-0.54	0.589
Condition (Last vs. First + Full) × Gender Rating × Likeability	-0.06	0.04	-1.55	0.120
Condition (First vs. Full) × Gender Rating × Likeability	0.03	0.05	0.68	0.498
Condition (Last vs. First + Full) × Accomplishment × Importance	0.03	0.03	0.93	0.350
Condition (First vs. Full) × Accomplishment × Importance	-0.09	0.04	-2.35	0.019†
Condition (Last vs. First + Full) × Accomplishment × Likeability	0.01	0.03	0.25	0.805
Condition (First vs. Full) × Accomplishment × Likeability	0.03	0.04	0.78	0.433

Condition (Last vs. First + Full) × Gender Rating × Acc. × Imp.	0.09	0.02	4.39	**<0.001**
Condition (First vs. Full) × Gender Rating × Acc. × Imp.	-0.03	0.03	-1.28	0.202
Condition (Last vs. First + Full) × Gender Rating × Acc. × Lik.	-0.10	0.02	-4.54	**<0.001**
Condition (First vs. Full) × Gender Rating × Acc. × Lik.	0.04	0.03	1.36	0.174
Observations	8771			

[†]*Bonferroni corrected α = .0013*