Graded gender perceptions in personal names predict difficulty in reference resolution

Hassan Khan (yetanotherhassan.khan@mail.utoronto.ca)[1], **Yoonjung Kang**[2,1] **& Dave Kush**[2,1]
[1]Department of Linguistics, University of Toronto
[2]Department of Language Studies, University of Toronto Scarborough

Abstract

We conducted two experiments to better understand how comprehenders use probabilistic knowledge of the associations between genders and personal names when inferring the gender of new referents during reading. Personal names are, surprisingly, under-investigated in studies on gender inference compared to role nouns (e.g., *nurse*) despite being strongly connected to gender. In a self-paced reading task involving variably gender-biased names and gendered reflexives, we observe a gender mismatch effect proportional to the incongruence between a name's perceived gender and that of the reflexive. In a novel variant of the Maze task involving forced ambiguity resolution between two gendered reflexives, we find decisions track the perceived gender of names and decisions take longer with more ambiguous names. We interpret the results to support a model in which comprehenders make graded, not categorical, inferences about gender during incremental comprehension, contrary to previous proposals.

Keywords: stereotypical gender; personal names; gender mismatch effect; Maze task; ambiguity resolution

Background

During incremental language comprehension, inferring the gender of new referents depends on the interaction of lexical knowledge, world knowledge and discourse context. Some nouns, like *queen* or *uncle*, are definitionally gendered, but many that are not still carry stereotypical gender associations (e.g. *nurse* is stereotypically female). Past research has used the distinction between definitional and stereotypical gender to investigate how language users represent gender and how they recruit different information sources to draw inferences about gender (e.g. Carreiras et al., 1996; Kreiner et al., 2008).

One question that studies have asked is whether comprehenders draw *categorical* inferences about the gender of referents introduced by stereotypically gendered NPs, as is assumed to be the case with definitional gender, or whether they draw *graded* inferences that represent probabilistic uncertainty about referents' gender. Categorical accounts, like that of Kreiner et al. (2008), assert that when a referent is referred to with a stereotypically gendered role noun, its gender is automatically assigned as either male or female and only reassessed when challenged by additional context. Categorical accounts can differ on whether they assume the categorical feature to be lexically encoded or inferred from world knowledge. They also differ on whether inferences are *deterministic*, based on the stereotypical gender of the noun, or *probabilistic*, based on the noun's gender distribution (see Kreiner et al., 2008, for a detailed discussion). In graded accounts, by contrast, gender may be represented in a non-categorical, probabilistic manner, either as a feature with a continuous value in a single representation, as Staub (2009) posits for number, or across multiple weighted representations held concurrently.

Against this backdrop, some studies have tested how participants respond to information, such as a gendered reflexive (*herself*/*himself*), that either confirms or disconfirms the definitional or stereotypical gender of a prior noun phrase. These studies use processing differences attributed to varying gender representations and their interaction with other information, i.e. Gender Mismatch Effects (GMMEs), as an implicit measure of gender assignments or activations. For example, in a sentence like (1), comprehenders show difficulty processing the masculine reflexive (*himself*) after seeing a feminine subject NP (*queen*), while they show no difficulty with the matching feminine reflexive (*herself*). Importantly, a stereotypically gendered subject NPs like *nurse* show a similar GMME, as in (2), suggesting similar inferences for definitional and stereotypical gender. GMMEs can manifest as increased reading times at or after the reflexive in reading studies (e.g. Carreiras et al., 1996; Duffy & Keir, 2004; Kreiner et al., 2008; Sturt, 2003) or P600-like components in event-related potential (ERP) studies (Osterhout et al., 1997).

(1) The queen looked at {himself/herself} in the mirror.

(2) The nurse looked at {himself/herself} in the mirror.

Such studies have asked (i) whether stereotypical GMMEs differ in size from definitional GMMEs (Carreiras et al., 1996; Doherty & Conklin, 2017; Osterhout et al., 1997) and (ii) whether stereotypical GMMEs can be attenuated with preceding discourse context (Carreiras et al., 1996; Duffy & Keir, 2004; Kreiner et al., 2008). A few studies have found suggestive differences between nouns with definitional or highly biased gender compared to stereotypical or less biased nouns (e.g. Doherty & Conklin, 2017; Osterhout et al., 1997), but in at least one study (Experiment 1 from Kreiner et al., 2008), size differences emerged in later measures after an initial stage where GMMEs were comparable. Potential differences in the size and contextual attenuation of effects are relevant to establishing whether inferences are categorical or graded, although the interpretation of such results depends on various processing assumptions (see Kreiner et al., 2008, and our General Discussion). A natural prediction is that if representations or

inferences are probabilistic or graded, the size of the GMME should vary as a continuous function of individual items' ambiguity. To our knowledge, no perception study has tested for this possibility, instead only looking for differences between noun types at the group level.

The aforementioned GMME studies have focused on role nouns. However, names are also conventionally associated with genders, and can vary in the degree of their association with one gender or another. Unlike most nouns, gender is often the most salient information carried by names (Alford, 1988), so names may offer a more direct insight into gender processing. While most names, like *David* and *Emma*, are strongly associated with male or female gender (Lieberson et al., 2000), many names are less categorical, but still biased (e.g. *Jordan*, *Avery*). Some names are more ambiguous than others (e.g. *Quinn*). Name-gender associations depend on individual experience, but Gardner and Brown-Schmidt (2024) found normed name ratings to be highly consistent with actual distributions.

To our knowledge, few studies have investigated the nature of gender inferences drawn from names, including whether inferences are categorical or graded.[1] A notable exception is Gardner and Brown-Schmidt (2024), which examined pronoun choice in participants' continuations of sentences containing variably gendered personal names. The authors found that pronoun choice (and explicit gender inferences) were continuously proportional to averaged name ratings on a 7-point scale from "definitely masculine" to "definitely feminine" collected in a separate norming study. As a production study, their results may reflect conscious offline decision-making processes different to those observable in online comprehension tasks. They also follow the convention of using averaged ratings from a norming study, which obscures the potential contribution of individual differences; we use individual ratings from participants for our experiments.

Our experiments fill gaps in the literature by investigating whether gender ambiguity in personal names influences the difficulty of processing a subsequent reflexive pronoun in a gradient manner. The first experiment, a self-paced reading (SPR) task, tested whether names' gender biases impacted the size of a GMME. Our second experiment, a new variant of the Grammar Maze (G-Maze; Forster et al., 2009) task, tested whether ambiguity-driven processing difficulties persist even when participants can choose which reflexive they prefer to continue the sentence.

Experiment 1

Past studies have found GMMEs on pronouns using names with strong conventional gender associations (e.g. Badecker & Straub, 2002; Kush & Dillon, 2021). We tested whether GMME size varies continuously as a function of participants' subjective assessment of gender-name associations.

Stimuli

30 names were selected from datasets containing baby names in Ontario from 1917 through 2022 (Government of Ontario, 2022a, 2022b). All selected names occurred at least 3,000 times. Although we analyze names' gender bias as a continuous variable, we selected names in five categories of six names each: strongly female (SF), female-leaning (FL), ambiguous (AM), male-leaning (ML), and strongly male (SM). SM and SF names were used at least 99.9% of the time with their associated gender. FL and ML names were used with their biased gender over 65% of the time, while AM names were more evenly split. Participants' gender ratings (discussed below) generally agreed with database statistics.

We created 30 test sentences similar to Kreiner et al.'s (2008) anaphora stimuli. Each contained a four-word main clause followed by a non-finite adjunct clause as in (3). The subject of the main clause was a name drawn from the five groups above. Adjunct clauses started with a subordinator (*after* in 3), a non-finite verb (*injuring*), and an independently manipulated gendered reflexive (*himself*/*herself*).

(3) {David/Charlie/Robin/Sydney/Emma} took time off after injuring {himself/herself} in the weight room.

Each participant saw all 30 names and all 30 sentences, each presented once with one of the two reflexives in a Latin square design. Test sentences were randomly intermixed among 46 filler sentences of varying length and complexity.

Participants

237 self-reported English-dominant participants were recruited through the Linguistics Participant Pool at the University of Toronto Scarborough and were compensated with course credit.

Procedure

Sentences were presented in a centered, word-by-word non-cumulative self-paced reading task (Just et al., 1982) on *PCIbex* (Zehr & Schwarz, 2018) in a controlled lab environment. Participants answered a forced-choice comprehension question after each sentence. After the reading task, participants rated the gender associations of each of the 30 names on a continuous sliding scale. The endpoints of the scale were marked "100% FEMALE" on the left (recorded as 0) and "100% MALE" on the right (recorded as 99).

Analysis

Data from 171 of the original 237 participants were analyzed. Participants were excluded for various reasons: 49 participants reported growing up outside of North America, potentially making the counts from the Ontario name corpus less representative; 10 had reaction times greater than 1,000 ms in 40% of observations; 5 had reaction times below 200 ms in

[1] Carreiras et al. (1996) mentions the prevailing assumption that the gender of the referent is automatically incorporated into the mental model for conventionally male or female personal names, likening them to role nouns with definitional gender, but offers no account of less categorical or ambiguously gendered personal names.

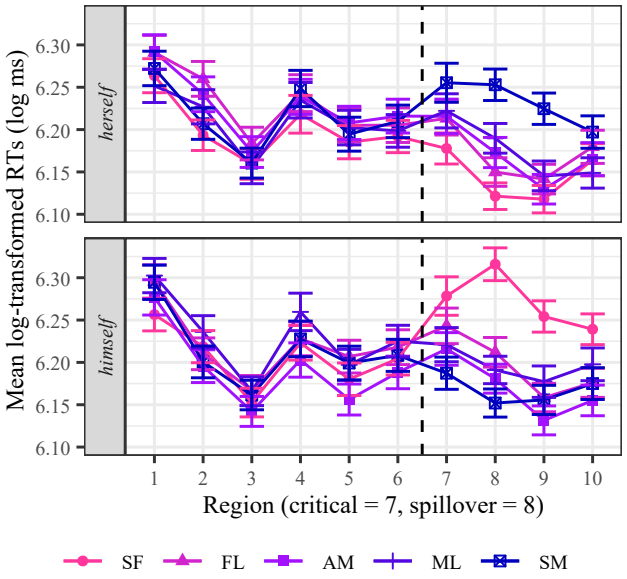

Figure 1: Log-transformed RTs by region and gender bias faceted by reflexive, with standard errors. The dashed line indicates the onset of the critical region.

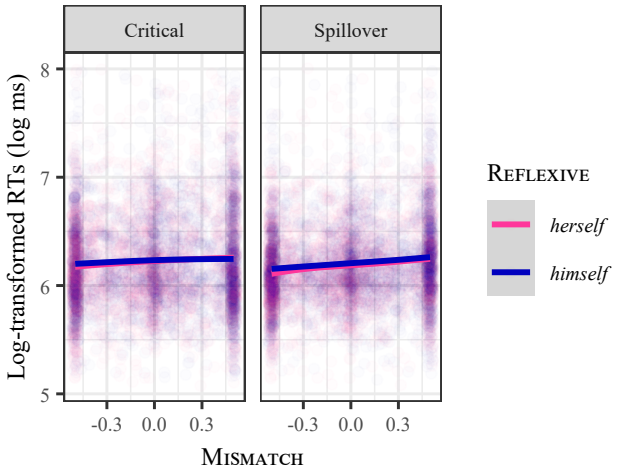

Figure 2: Log-transformed RTs by Mismatch and Reflexive at the critical and spillover regions, with Locally Estimated Scatterplot Smoothing (Loess) curves for each reflexive condition. Mismatch values are horizontally jittered.

one-third of observations; and one participant had an accuracy below 80% on comprehension questions. We also excluded observations with reaction times less than 150 ms or greater than 3,000 ms (0.50% of all observations).

For our analysis, we introduce a gradient measure Mismatch (complete match = -0.5, complete mismatch = 0.5) to capture the incongruence between a participant's rating of the name and the gender of the reflexive on each trial, treating ratings as reflective of the gender associations participants held while reading. For example, if a participant rated *Michael* as completely male, a trial where they see *Michael* presented with *himself* would receive a Mismatch of -0.5, while a presentation with *herself* would receive 0.5. If another participant rated it as somewhat but not completely male, a trial where they see *Michael* with *himself* would have a Mismatch between -0.5 and 0 whereas *herself* would have a Mismatch between 0 and 0.5.

Log-transformed reading times in the critical reflexive and spillover regions were analyzed using linear mixed effects models fit with *lme4* (Bates et al., 2015) in *R* (R Core Team, 2025), followed by *t*-tests of significance using Satterthwaite's degrees of freedom approximation with *lmerTest* (Kuznetsova et al., 2017). Models were fit with fixed effects of Reflexive (feminine = -0.5, masculine = 0.5), Mismatch, and their interaction; random intercepts by Participant, Sentence, and Name; and random slopes for Mismatch, Reflexive, and their interaction by Participant and Sentence.

Results and discussion

Region-by-region reading times are shown in Figure 1, binned by name group for ease of visualization. Our models revealed

Table 1: Fixed effects from the linear mixed-effects model predicting log-transformed RTs at the critical region.

	Coeff	SE	*t*	*df*	*p*
(Intercept)	6.225	0.024	257.52	176.0	<0.001
Mismatch	0.090	0.018	5.09	73.9	<0.001
Reflexive	0.008	0.011	0.71	44.6	0.478
Mis:Refl	0.001	0.033	0.02	34.3	0.985

a significant positive effect of Mismatch on log-transformed RTs at both the critical and spillover region, meaning participants were slower to read reflexives that were less congruent with the perceived gender of a name; see Tables 1 and 2. The average estimated penalty of a complete mismatch according to the model is 45 ms at the reflexive and 65 ms at the spillover region. Analysis of RTs at the pre-critical region showed no significant effects, confirming that the effect of Mismatch originated at the reflexive.

The effect of Reflexive was not reliable at the critical region, but we observe a significant positive effect of Reflexive in the spillover region, indicating *himself* resulted in a slowdown of 12 ms on average compared to *herself* (see also Figure 2). The interaction with Mismatch was not significant in either model. We return to this effect briefly in the General Discussion.

Overall, our results are consistent with the hypothesis that the GMME is gradient and proportional to the degree of mismatch. While a gradient GMME is consistent with a probabilistic inference (whether the representation is itself categorical or graded; we explore this in depth in the General Discussion), it could also be the result of participants initially drawing categorical and deterministic inferences and

Table 2: Fixed effects from the linear mixed-effects model predicting log-transformed RTs at the spillover region.

	Coeff	SE	*t*	*df*	*p*
(Intercept)	6.195	0.020	308.51	178.1	<0.001
MISMATCH	0.133	0.014	9.17	34.7	<0.001
REFLEXIVE	0.024	0.010	2.34	47.8	0.02
MIS:REFL	0.000	0.031	-0.01	31.3	0.989

having difficulty accommodating the unexpected pronoun proportional to how unlikely it is. To more closely observe the effects of ambiguity, we turn to an experiment that can distinguish reinterpretation from ambiguity resolution.

Experiment 2

To further probe effects of gender, we ran a novel variant of the Grammar Maze (G-Maze) task, which has been shown to replicate results from SPR with larger and more localized effects (Forster et al., 2009). The G-Maze task is a word-by-word forced choice sentence continuation task, in which participants choose between grammatical continuations and ungrammatical distractor words.

In our version of the task, participants read grammatical/ungrammatical pairs in the pre-and post-critical regions as usual, but receive two potentially grammatical, but mutually exclusive, alternatives at the critical region: the reflexives *himself* and *herself*. Choosing a reflexive requires disambiguating the gender of the subject if it is not yet disambiguated, but does not force a particular reading unlike the SPR task. Thus, there is no mismatch *per se*. We look at how name ratings affect reflexive selection and how the degree of ambiguity affects reaction times. If responses track the distributions of names, there is evidence for the use of graded probabilistic knowledge in gender inference, as opposed to deterministic categorical inferences of the biased gender that can be overridden by context as Duffy and Keir (2004) seem to suggest (see also Kreiner et al., 2008).

Additionally, longer response times at the reflexive after ambiguous names could be the result of a more continuous representation of referent gender (as suggested by Staub, 2009), or competition between two comparably likely options (as suggested by Boyce et al., 2020). In either case, participants maintain some representation of the ambiguity of the referent's gender, which excludes a naive account of deterministic or probabilistic categorical inferences that are only reassessed when the initial assumed gender is challenged by the context.

Stimuli

Stimuli were reused from Experiment 1, though reflexive gender was no longer factorially manipulated: alternatives at the critical region were always *himself* and *herself* in test items. Distractor words were automatically generated with the *A-Maze* script from Boyce et al. (2020), ensuring similar length and frequency with paired targets. We manually verified all distractors to be ungrammatical continuations.

Participants

82 self-reported English-dominant participants were recruited online through the Prolific Academic recruitment platform (prolific.co). Participation was restricted to IP addresses within Canada and the US. Participants received £3.75 for their participation, which lasted 24 minutes on average.

Procedure

Participants completed a modified G-Maze task (Forster et al., 2009). Items were shown word-by-word, with the target and distractor words appearing on random sides of the screen. Participants pressed 'e' or 'i' to select a continuation. An incorrect response outside the critical region ended the trial. At the critical region, neither *himself* nor *herself* were ever considered 'incorrect'. We recorded which reflexive participants chose and response times for each word in the sentence. After completing the Maze task, participants completed a name rating task identical to that of Experiment 1.

Analysis

Data from 73 of the original 82 participants were analyzed. 3 participants were excluded for successfully completing less than a pre-determined 70% of trials. 6 were excluded for giving highly unusual name ratings (e.g. predominantly using the midpoint of the scale, or mis-rating most strongly biased names), although our results stayed significant whether or not they were included. We also excluded 10 trials where participants spent less than 150 ms reading the name.

For our analysis, we introduce AMBIGUITY as a measure of how close to the centre of the scale a participants' rating of the name on a given trial was. AMBIGUITY (completely unambiguous = -0.5, completely ambiguous = 0.5) was taken as the absolute distance between the midpoint and the closest endpoint. A rating of completely female or male would receive an AMBIGUITY of -0.5, while a rating at the centre of the scale would receive an AMBIGUITY of 0.5. RATING was centred between -0.5 (completely female) and 0.5 (completely male).

Log-transformed RTs were analyzed at the reflexive. Choices of reflexive were analyzed using a logistic mixed-effects model followed by asymptotic Wald z-tests of significance. The first model had a fixed effect of AMBIGUITY; random intercepts by PARTICIPANT, SENTENCE, and NAME; and random slopes for AMBIGUITY by PARTICIPANT and SENTENCE. The second model had a fixed effect of RATING; random intercepts by PARTICIPANT, SENTENCE, and NAME; and random slopes for RATING by PARTICIPANT and SENTENCE.

Results and discussion

Participants' choice of reflexive was proportional to the perceived gender bias of names, as seen in Figure 3. This is confirmed by a significant positive effect of RATING on participants' choices in the model presented in Table 3. The lack

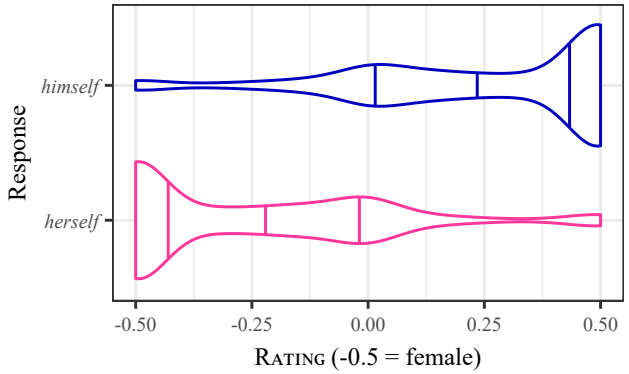

Figure 3: Violin plots showing the distribution of Ratings that elicited *herself* and *himself* responses. Lines indicate quartiles.

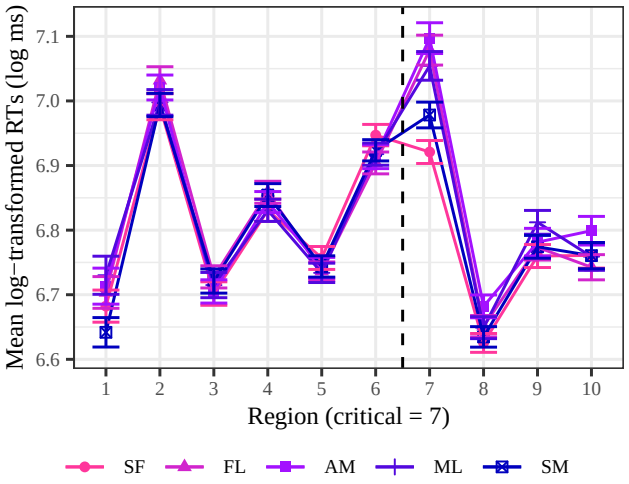

Figure 4: Log-transformed RTs by region and gender bias, with standard errors. The dashed line indicates the onset of the critical region.

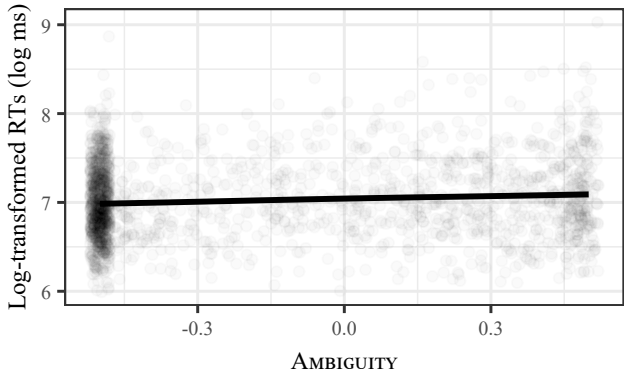

Figure 5: RTs at the critical region by Ambiguity, with a Loess curve. Ambiguity values are horizontally jittered.

Table 3: Fixed effects from the logistic mixed-effects model predicting reflexive choice.

	Coeff	SE	*z*	*p*
(Intercept)	-0.143	0.087	-1.64	0.102
Rating	5.938	0.370	16.05	<0.001

Table 4: Fixed effects from the linear mixed-effects model predicting log-transformed RTs at the critical region.

	Coeff	SE	*t*	*df*	*p*
(Intercept)	7.041	0.035	201.29	80.5	<0.001
Ambiguity	0.176	0.041	4.29	26.5	<0.001

of a significant intercept means we do not observe an overall bias towards *himself* or *herself*. By-region response times are shown in Figure 4. Response times at the reflexive increased for more ambiguously gendered names, a pattern borne out by a significant positive effect of Ambiguity in the model presented in Table 4. The effect is plotted in Figure 5.

The results are consistent with the hypotheses that greater gender ambiguity leads to lengthened decision times and that participants' choice of reflexives depends on the perceived distribution of a name in a gradient manner.

General discussion

Our experiments demonstrate that participants have access to, and maintain representations of, the continuously variable degree of gender associations in personal names during comprehension. Experiment 1 (SPR) showed a correlation between the size of a GMME and the continuous measure Mismatch, something previously uninvestigated beyond comparisons between binned groups. Experiment 2 (G-Maze) showed a correlation between participants' individual ratings and reflexive choice. Experiment 2 also demonstrated that sensitivity to the ambiguity of a name persists even when it is possible to choose either reflexive.

The results of Experiment 1 are straightforwardly consistent with a graded inference account. If graded inference involves weighted parallel representations or a continuous gender value assigned to a referent in a single representation, the size of the GMME should scale with the probability assigned to the mismatching gender. This is what we see.

The results of Experiment 1 could also be explained under a simple categorical inference account under certain assumptions. If gender is initially assigned categorically, but revised when confronted with new, incompatible information, then the size of a GMME could index the cost of computing or accommodating the new interpretation (as proposed by Kreiner et al., 2008; Osterhout et al., 1997, for stereotypical GMMEs). An alternative interpretation can be found when stereotypical gender is violated, but not when definitional gender is (Kreiner et al., 2008). Our results would motivate the additional as-

sumptions that (i) accommodation cost is continuous, scaling with strength of bias and (ii) accommodating dispreferred gender for personal names works the same way. Alternatively, the gradient GMMEs in Experiment 1 are in principle compatible with a model in which early categorical inferences are made probabilistically across trials. According to such a model, average processing difficulty reflects the proportion of trials on which participants experience a slowdown due to categorical mismatch with a reflexive. A strongly biased name like *Michael* would never mismatch *himself*, but a fully ambiguous name would mismatch *himself* about half of the time, resulting in a longer average RT.

Though it could explain Experiment 1's results, a simple deterministic categorical inference theory would fail to explain response patterns in Experiment 2, where the probability of choosing a reflexive was sensitive to ambiguity. If the biased gender is consistently assigned until contradicted, we should expect participants to always choose the reflexive consistent with the biased gender of a name, rather than the gradient pattern we observe. Unlike the simple account, probabilistic categorical assignment could explain the pattern of reflexive choice in Experiment 2, with response probabilities directly reflecting the probabilities of choosing a particular categorical gender. These patterns are, of course, also consistent with a graded representation account, where responses are probabilistically drawn from the representation.

Evidence in favour of graded inference over both versions of categorical inference comes from the fact that participants took longer to select a reflexive after seeing more ambiguous names. Firstly, the ambiguity-sensitive effect is similar to the pattern observed by Staub (2009) for agreement computation under conditions of number attraction. In a sentence continuation task using items like (4), Staub (2009) saw participants took longer to choose a verb (*is* vs. *are*) when the target and distractor nouns had mixed numbers (one plural, one singular) than when both were singular or plural. Importantly, the penalty was observed regardless of which form the participants ultimately chose.

(4) The key(s) to the cabinet(s) {is/are}...

Staub (2009) used the ambiguity penalty to argue that both correct and incorrect responses were drawn from the same graded representation, with less certain responses resulting in longer decision times. Secondly, discussion from Boyce et al. (2020) on the processing of forced-choice alternatives in the Maze task supports the contention. According to Boyce et al. (2020), the two alternatives could be processed independently and in parallel, or they could compete. If they are processed in parallel, response times should only depend on the likelier candidate. If they interact, competition should increase with ambiguity, resulting in increased processing time. In either scenario, if participants have already drawn a categorical inference, the matching reflexive should be overwhelmingly preferred, and finer-grained probabilistic knowledge of gender associations should not intrude on the decision process. Thus, both Staub's (2009) and Boyce et al.'s (2020) arguments lead us to conclude that the ambiguity effect on response times indicates a gradient representation.

A potential rebuttal to this reasoning is that participants' sensitivity to gradient ambiguity could be a task effect: Seeing both reflexives may reactivate discourse-independent probabilities rather than probing their gender inference directly. While we cannot exclude this possibility with our experimental set-up, we note that this should affect all names, while the response times at the reflexive for strongly gendered names in Figure 4 do not seem qualitatively different from those in the SPR task in Figure 1, suggesting there was no penalty for strongly gendered names. If reactivating discourse-independent probabilities is costly, this cost should be observable in even strongly gendered names.

An unexpected effect was the reading-time penalty for *himself* in Experiment 1. One might have expected a penalty for *herself* since it is less frequent, more marked, and disfavoured by the *people* = *male* bias for using masculine pronouns more than actual gender distributions would suggest (Gardner & Brown-Schmidt, 2024; Hamilton, 1991). Given that the difference is found even with strongly gendered names (see Figures 1 and 2), we speculate that the difference may be linked to some awareness of the tendency for names to undergo shifts in usage over time from male to female but not the reverse (Lieberson et al., 2000), as it would mean it is more acceptable and thus less surprising for a male name to be given to a woman (as implied by *herself*) than the reverse. While it appears from Figure 1 that the difference is compounded by the especially strong mismatch effect when a strongly female name is shown with *himself*, which could suggest that female names engender stronger expectations for a feminine pronoun than the reverse, we do not see the significant interaction this would imply. Future examinations of non-linear effects, like the apparent jumps at the extremes of gender ratings, might be able to disentangle these. It would also be worth exploring how a more sophisticated treatment of gender that treats nonbinarity and gender-neutrality in names and reference affects the observed patterns.

We end with two methodological take-aways: Firstly, our experiments show that individual ratings work well as a continuous predictor for work on effects of stereotypical gender associations. Individual ratings are superior to averaged norms, used in previous work like Gardner and Brown-Schmidt (2024), in that they are logistically simpler, require fewer assumptions, and, most importantly, offer finer-grained information about the effects of individual gender perceptions, rather than population-level distributions. Second, we introduced a novel variant of the Maze task, which allows us to observe both how ambiguity affects response times and how responses track gradient representations simultaneously. We suggest that the variant may be useful in investigating other aspects of gradience and ambiguity in the future.

Acknowledgements

The authors would like to thank the research assistants in the Eyelands Lab for their work on the experiment, as well as the anonymous peer reviewers for their feedback. This work was funded by the University of Toronto Excellence Award-Social Sciences and Humanities, SSHRC Insight Grant #435-2020-0209, and SSHRC Insight Development Grant #430-2022-0011.

References

Alford, R. D. (1988). *Naming and identity: A cross-cultural study of personal naming practices*. HRAF Press.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748–769. https://doi.org/10.1037/0278-7393.28.4.748

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

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*(104082), 1–31. https://doi.org/10.1016/j.jml.2019.104082

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 Section A*, *49*(3), 639–663. https://doi.org/10.1080/713755647

Doherty, A., & Conklin, K. (2017). How gender-expectancy affects the processing of "them". *The Quarterly Journal of Experimental Psychology*, *70*(4), 718–735. https://doi.org/10.1080/17470218.2016.1154582

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. https://doi.org/10.3758/BF03195846

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/BRM.41.1.163

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

Government of Ontario. (2022a). Ontario top baby names (female). https://data.ontario.ca/dataset/ontario-top-baby-names-female

Government of Ontario. (2022b). Ontario top baby names (male). https://data.ontario.ca/dataset/ontario-top-baby-names-male

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

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, *58*(2), 239–261. https://doi.org/10.1016/j.jml.2007.09.003

Kush, D., & Dillon, B. (2021). Principle b constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language*, *120*(104254), 1–16. https://doi.org/10.1016/j.jml.2021.104254

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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. https://doi.org/10.1086/210431

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

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, *60*(2), 308–327. https://doi.org/10.1016/j.jml.2008.11.002

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

Zehr, J., & Schwarz, F. (2018). Penncontroller for internet based experiments (ibex). https://doi.org/10.17605/OSF.IO/MD832