Title:
Children's Perception of Gender in Voices: children are less certain than adults when making inferences about speaker gender identity and make less normative responses.

Kaitlin Lee
Committee: Dr. Sarah Creel, Dr. Marc Garellek, Dr. Will Styler, Benjamin Lang
Cognitive Science Honors Program, 2023-2024
University of California, San Diego
Email: klee67510@gmail.com

Abstract:
Children can make informed assumptions about speaker identity based on inferred roles associated with the speaker or learned information about the speaker. However, a child's ability to make an informed choice about a part of a speaker's identity, such as their gender presentation, is not as well known. In this study, we looked into whether preschool-aged children can infer gender presentation based on how someone sounds, and if this ability changes based on the voice gender ambiguity of the speaker. We found that children overall seem less certain in their inferences of speaker gender compared to adults and have less normative inferences about speaker gender, as compared to adults. These findings impact our understanding of how psycholinguistic mechanisms underlie social categorization of gender, and how this understanding is developed.

Cognitive Science and Linguistics Departments
University of California, San Diego

1.1 Introduction:

What Are Little Boys Made Of?

"What are little boys made of?
What are little boys made of?
Snips, snails
And puppy-dogs' tails
That's what little boys are made of.

What are little girls made of?
What are little girls made of?
Sugar and spice
And everything nice
That's what little girls are made of."

– Author Unknown, early 19th century

From a young age, we can identify characteristics of femininity and masculinity and physical differences between men and women. Trautner et al (2003) laid out three stages for how gender stereotyping and knowledge are formed in children:

1) Children learn about gender-related characteristics in the toddler and preschool years.
2) At around 5 to 7 years, this newly-formed gender knowledge is rigidly consolidated.
3) After this peak of rigidity in understanding gender, a phase of relative flexibility follows.

Voice quality of a person contributes to the perception and knowledge of gender. While gender is a complex combination of different characteristics, voice plays a key role in presenting our gender and how we want to be viewed. For example, transgender people in the first steps of their social transition will modulate their voice differently to reflect their gender identity. Regardless of whether they medically transition, many transgender individuals will attend professional voice training to adjust their register of delivery (Planned Parenthood 2024).

Different aspects of voice contribute to someone's gender presentation. Some of these are physiological attributes such as vocal tracts. Women have vocal tracts found to be around 20% smaller than vocal tracts in men, influencing different aspects of acoustic features, including fundamental frequency (Titze 1989). Fundamental frequency (F0) also varies distinctly between men and women: women typically have F0s approximately one octave greater than men's F0s (Klatt and Klatt 1990).

In children, in terms of acoustic measurements, vocal gender is differentiated mostly by formant frequencies. Perry et al (2001) found that the formant frequencies of 8-year-old boys are 9% lower than those of 8-year-old girls, while they have a large overlap of mean F0 until they're 12. Regarding perception, children have been found to exhibit different cue weights on F0 and vocal

tract length than adults to determine vocal gender (Nagels et al 2020).

Previous research suggests that children as early as toddlerhood not only match voices and faces based on gender but also make informed assumptions about someone's gender based on their voice. Bacon and Saffran (2021) had 22- to 24-month-old toddlers do a task where they would see two familiar objects, one coded to be made for boys and one for girls. An example of this would be a purple winter coat meant for girls, and a black and red winter hat meant for boys. The study found that when hearing either a male or female voice, toddlers would fixate their gaze on the object whose gender preference matched the speaker's gender voice.

Beyond gender, we know 3-to-5-year-old children can make informed inferences about speaker identity based on voice or informed inferences from voices based on speaker identity. Children can integrate talker-role-associated knowledge, even when not mentioned explicitly, and use their knowledge to inform their interpretations of speakers and what they say. Creel (2012) found that preschool children can use voice characteristics to infer what the speaker is likely to talk about and, thus, can predictively fixate their gaze on objects that speakers had indicated were their favorite colors. In Borovsky and Creel (2014), children aged 3-10 were tasked with identifying the correct object when asked for by a speaker assuming the role of a pirate or princess ("I want to hold the (sword/wand)". They found that children can fixate on the speaker-target image before the target object is mentioned by combining voice cues to speaker identity with verb information. In other words, children can use information about a speaker and determine aspects of their identity using voice characteristics or other elements of their voice.

However, sensitivity to gender-voice ambiguity has not been measured in children in any capacity. This may be due to a general lack of developmental studies focused on sociolinguistics, especially ones that look into gender perception. There is also a lack of studies that explore our understanding of gender as something that exists outside of a binary, especially in studies that focus on child development.

Going up the ages, we have done more research on adults' perception of voice gender. Overall, adults have been found to make inferences about people's gender based on a speaker's voice that are not always related to acoustic variables, but other perceptual elements.

Mullenix et al. (1995) assessed that adult perceivers regularly make informed assessments of people's gender based on voice. Instead of using abstract representations of voice, adults can interpret voices on a case-by-case basis, using phonetic representations. Other studies have found that gender perception is influenced by previously heard voices rather than by acoustic measurements like F0 (Schweinberger et al 2008).

Lang (2022) took 66 podcast clips, each with a unique speaker of a varied racial, gender, and

sexual orientation identity, and had adult participants infer each speaker's gender presentation, gender identity, and sexual orientation based on voice quality. Lang found that adults' inferences about identity were related to certain acoustic-phonetic properties in voices. However, Lang does not characterize whether this accurately reflects the speaker's actual sexuality or gender identity.

Still, there is a great lack of research learning more about how we understand and learn about identities that lay outside of the socially constructed gender binary of man and woman. We have only scratched the surface of what we know about adult humans' comprehension of non-binary or non-conforming presentations of gender. Still, to gain better insight into what humans, in general, can understand or process about gender presentation, we need to look more into the capabilities of children.

This leads to the question of whether preschool-aged children can infer speaker gender characteristics based on the gender voice properties of the speaker.

We hypothesize that preschool-aged children (ages three to five) should overall repeat the behavior of adults who take the same study. This means that they will make the same inferences about speakers and will pair them with a specific gender presentation. Depending on rated gender voice ambiguity, children will be able to pick up on this ambiguity, which will then induce a longer time to make an inference on the possible speaker.

Our null hypothesis is that three-to-five-year-olds will not detect subtleties in gender voice presentation as much as adults do, inducing similar inference-making times across different gender groupings, regardless of the pre-rated gender voice ambiguity these speakers may have.

2.1 Participants

We collected data by visiting private preschools in the general San Diego Metropolitan area. Participating children received a sticker at the end of their run, regardless of completion. Participating preschools received a Barnes and Noble gift card as compensation for participating. For children, we collected data from various language and cultural backgrounds. Child participants were either monolingual American English speakers or came from multilingual backgrounds. Home and educational spoken languages, so far with our preliminary study, included Mandarin Chinese, Bulgarian, French, Spanish, Russian, Hungarian, and Tagalog. They ranged in age from 2 years and 8 months old to five years old.

It should be noted that data collection of preschool children participants for this paper, turned in for the review of the UC San Diego Cognitive Science Honors board, has not been completed.

For adults, we recruited lab members and other college students/faculty through word of mouth from the University of California, San Diego. We collected data from 18 adults, with 1 drop due

to the participant's ocular motor dysfunction.

Adults came from diverse linguistic backgrounds. Most adult participants identified as multilingual, with home or educational backgrounds in Tagalog, Spanish, Mandarin Chinese, French, Hindi, Turkish, Korean, and Hebrew.

All participants gave their informed consent following the protocols approved by the University of California, San Diego Human Research Protections Program.

2.2 Stimuli / Norming

Audio clips were sourced from Ben Lang, who had previously used the clips for his 2023 study. Clips were taken from publicly available podcasts, with speakers of various genders, races, sexual orientations, and political alignments. Lang reported grouping the speakers by computing a k-means solution with k= 5 based on study participants' ratings of each speaker. He took that and found the speakers who were most aligned with each other in different groupings. We took Lang's methodology and found four speakers closest to the mean in each of the four groupings: [Male (M), Female (F), Ambiguous Male-Leaning (AMM), and Ambiguous Female-Leaning (AMF)]. Each of the resulting selected 16 audio clips contained a different speaker.

The podcasts that we sourced voice clips from were:

- *Queerified with Gigi Gorgeous and Mimi*
- *Homophilia*
- *The Matt Walsh Show*
- *Queery with Cameron Esposito*
- *Dyking Out - A Lesbian and LGBTQIA Podcast for Everyone!*
- *Why Won't You Date Me? With Nicole Byer*
- *The Candace Owens Podcast*
- *This Might Get Weird*
- *At Liberty*
- *Getting Curious with Jonathan Van Ness*
- *Latino USA*
- *Code Switch*
- *BRUNCH*

To determine if the phrases had any signifier of the gender identity of the speaker, we wrote out transcripts of each phrase, then had lab members rate them on a scale of 1 to 9 on a Google form; a rating of 1 was "man", 9 was "woman", and 5 was "uncertain". From all of the ratings, we determined that there was not a strong effect of noticeable gender signifiers.

We sourced our photos from the website https://thispersondoesnotexist.com/. The website uses

AI StyleGAN, a generative adversarial network using progressive growth mechanisms to generate deceptively realistic portraits of human faces (Karras et al). We took photos from the website and had adult lab members rank and categorize features based on these faces. Lab members categorized age range and ethnic background in an open-answer prompt. They chose from a list of possible gender identities for each face. Then, they ranked gender presentation on a Likert scale (1 for very masculine, 7 for very feminine, and 4 for somewhere in between).

From these rankings, we chose faces ranked very close to either 1 or 7. Then, we paired faces that matched in perceived age range and ethnic background (based on rankings given by lab research assistants).

For each face, we determined a target and competitor image by choosing the best image representing the speaker group to which the audio clip belonged. Male and ambiguous male-leaning voices were paired with faces labeled male, and female and ambiguous female-leaning voices were paired with faces labeled female. This should not be taken to imply that any face is a true “right” answer to the question of which voice was matched.

2.3 Procedure

We ran an adult pilot study alongside the children's study to determine possible developmental differences between children and adults. This adult study serves as a control group to compare the results of the children's study.

We used MatLab to present a binary-choice picture selection task. Participants listened to an audio clip of a speaker saying a phrase and were then engaged in a binary choice task between a male or female speaker. This was done by either clicking on a 400 X 400-pixel image (for adults) or pointing to the image to prompt the experimenter to click on an image (for children).

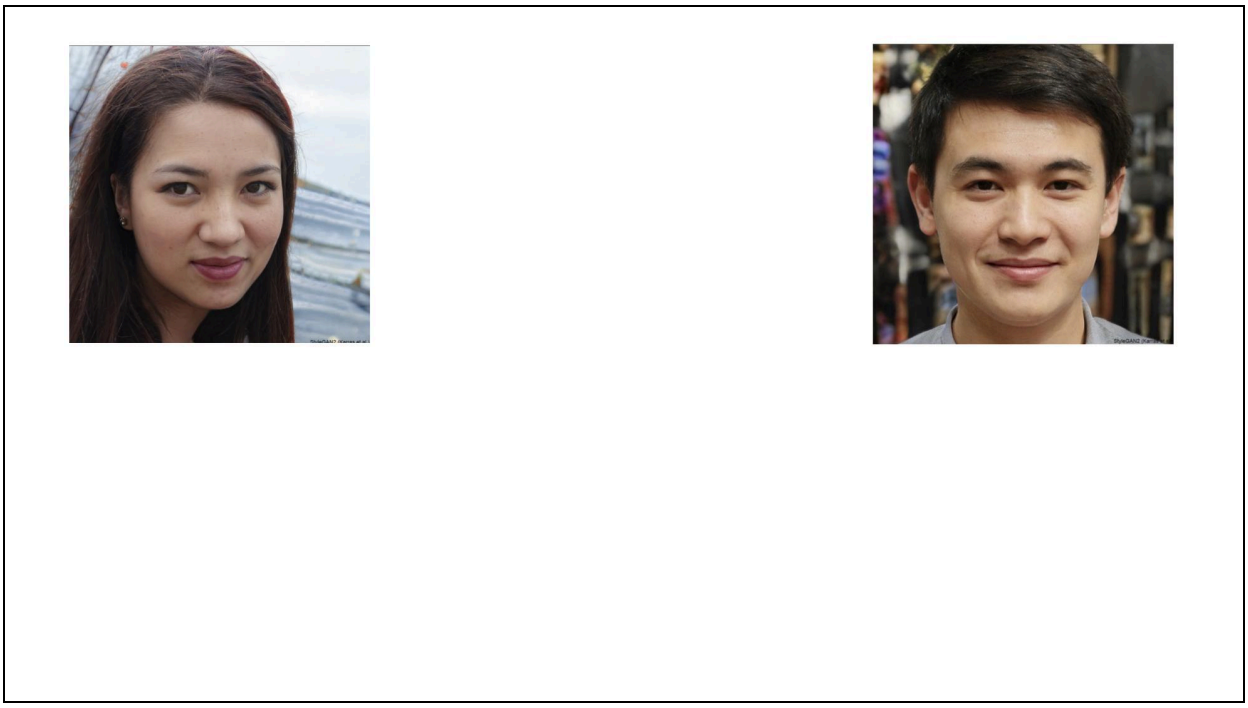

Fig 1.1: A recreation of a frame from a trial for the binary-choice task, with the two images included. Male and female figures would alternate between positions on screen.

For each trial, there were pictures of a male and a female on opposing sides. An audio clip would play, and the participant would either click on or point at a picture. There were 32 trials for 16 audio clips and 16 face pairings. Each audio clip would have one face pairing. Audio clips would play twice, but the male and female pictures would switch sides in the second trial with the audio clip. Ordering of the trials was done randomly.

Participants also had to complete a post-task face gender identification task to verify that they identified face gender as intended. In each trial, the participant saw two faces, one male and one female, and the audio asked for the participant to either click on or indicate where the man or the woman was. When we ran the post-task, we used the same face pairs from the first task. The speaker for the post-task was a female native Californian English speaker.

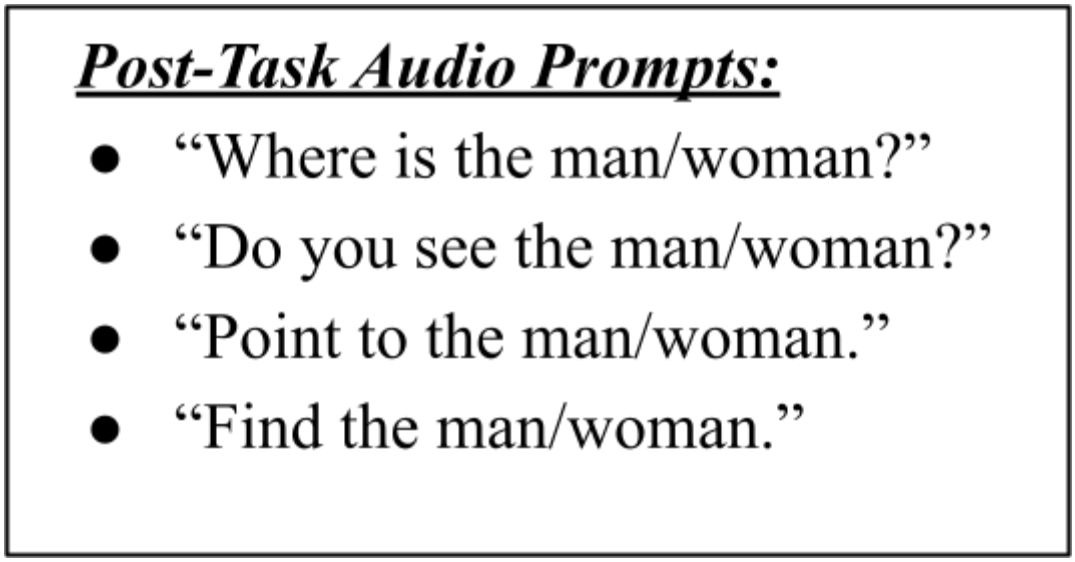

Fig 1.2: Phrases used in the post-task. Eye-tracking measurements began when the words "man" or "woman" were uttered. A female young adult native Californian English speaker read the phrases.

To measure how fast participants decided on speaker gender presentation, we used mouse-clicking data to measure reaction time. We also used eye gaze to measure visual reaction time with an Eyelink Portable Duo machine.

2.4 Analysis

For this study, we measured clicking speeds and eye movement speeds. For adults, we measured clicking speeds and eye movement to determine how fast they made an inference, and when they decided on the stimuli. We used clicks to measure the accuracy of inferences.

For children, since experimenters would click for them, we treated clicking speed as an inaccurate measurement of decision speed. To make up for this inaccuracy, we used mostly eye-tracking movement to measure the speed of inference-making. We also used clicks to measure the accuracy of inferences.

Ahead of data analysis, we preregistered our analysis to state that we would be looking at our data at two different times: when we gathered about 16 participants and after about 32 participants. This was to forestall any suspicions of data manipulation.

All data was analyzed using Python scripts, Excel, and RStudio.

3.1 Results: Adult Piloting

From adult piloting, we found that adults could connect voice clips with a speaker of a matching gender presentation accurately most of the time. We found that most adult participants (13 out of

17) made at least one non-normative answer. Most of the non-normative answers were for voice clips whose speakers were in the male-leaning ambiguous grouping.

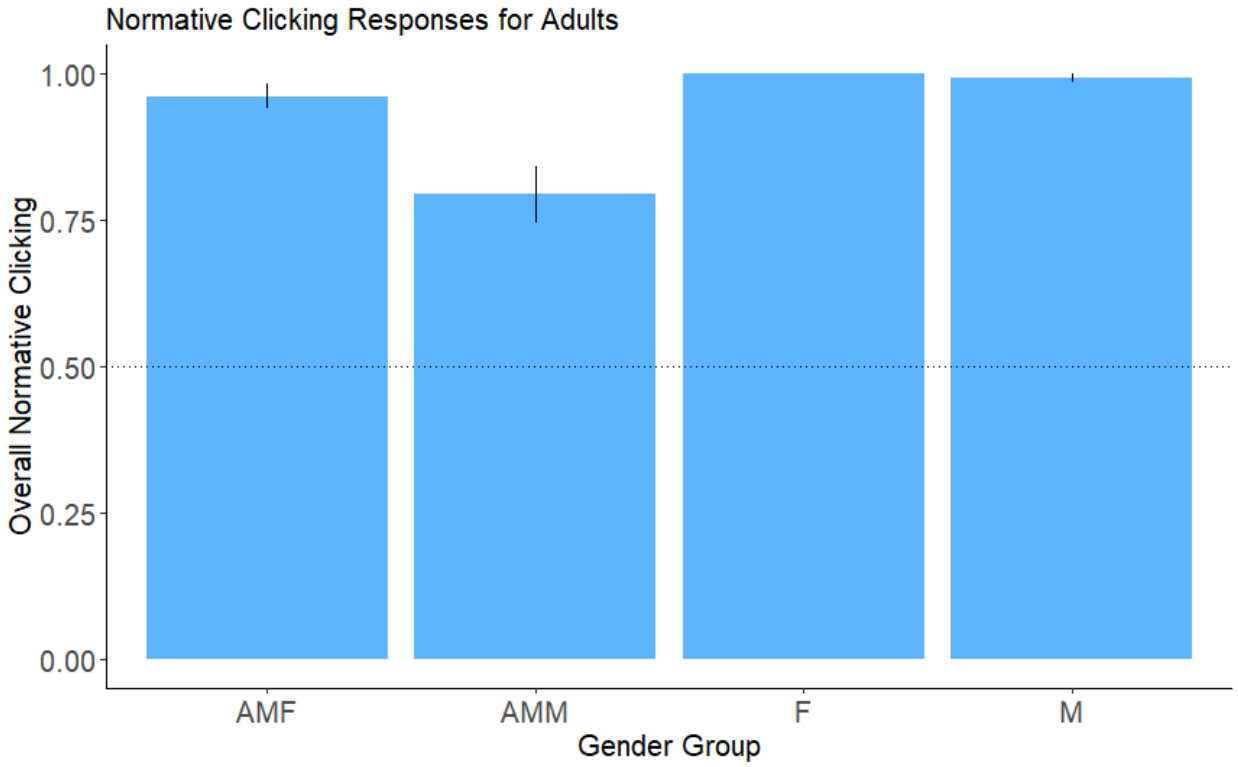

Fig 3.1: Normative clicking accuracy for adult participants. This showed whether adults overall made similar inferences about speaker gender. Adults made similar inferences for more typical male/female speakers, but there was more variance with ambiguous female and ambiguous male speakers.

Average reaction time from mouse clicks indicated that for typical male and female voice clips, reaction time was about the same. Participants reacted faster to voice clips that aligned more with the gender binary than the voice clips with gender-ambiguous voices. Voice clips with male-leaning ambiguous grouping had a greater reaction time than all of the other groupings. However, its difference from female-leaning ambiguous responses is less significant than its differences with the female or male groupings.

The average fixation time we found from eye-tracking data reflected the reaction time from clicking time. On average, overall participants fixated on the target image, reflecting results from Lang (2023). However, the average fixation on ambiguous female-leaning and ambiguous male-leaning voices was lower for target images, and it took longer for participants to fixate on the target image for ambiguous female-leaning and ambiguous male-leaning voices.

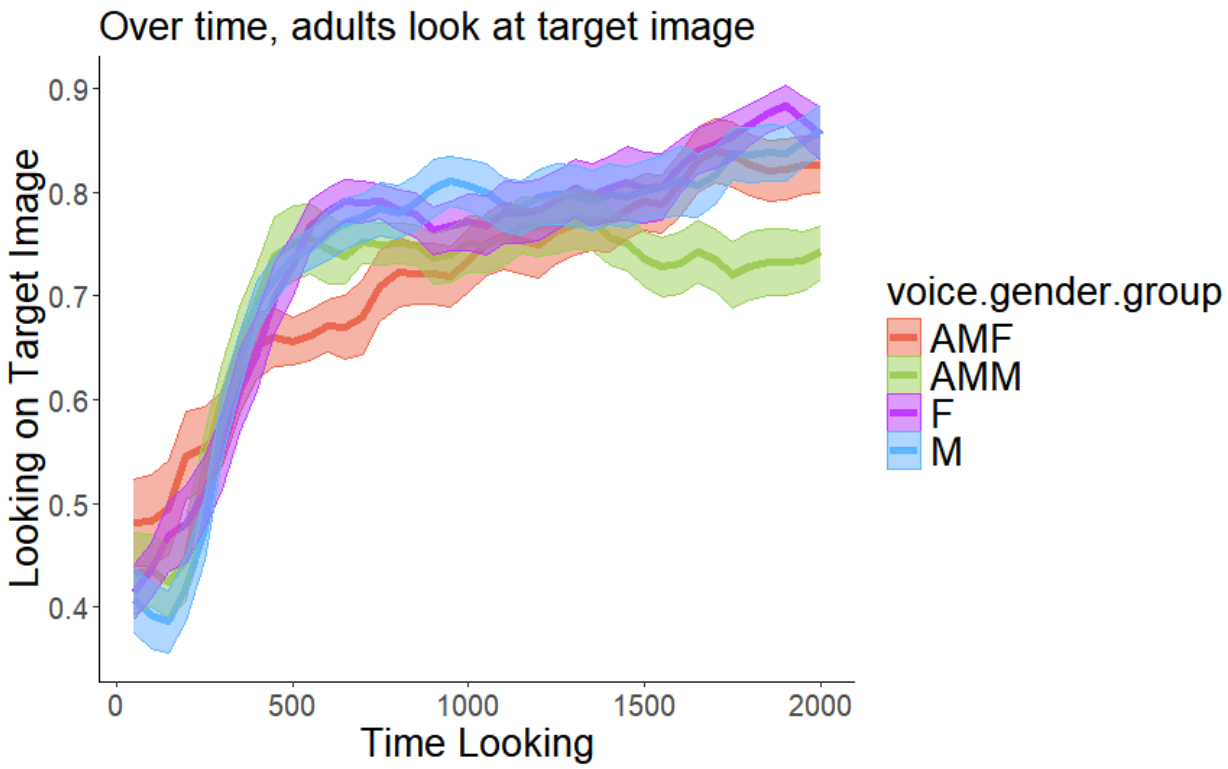

Fig 3.2: Chart of gaze fixation on the target image, depending on voice gender group condition. Here, we can see that overall, participants looked at the image that corresponded to what participants in Lang's study had determined for the voice clip classification. We can see overall that it took participants longer to fixate on target images for ambiguous-gender voices.

3.2 Results: Child Results

For children, we hypothesized that 3-to-5-year-olds would either replicate the results of adults or would not have non-significant differences between their response accuracy (or inference normativity level compared to adults) to different gender groupings.

Based on preliminary findings, we found that children do not replicate the inferences of adults. Children overall have inferences that are less normative compared to adults. Children do not have significantly different "accuracy" with inference-making with more ambiguously gendered voices versus more typically gendered voices.

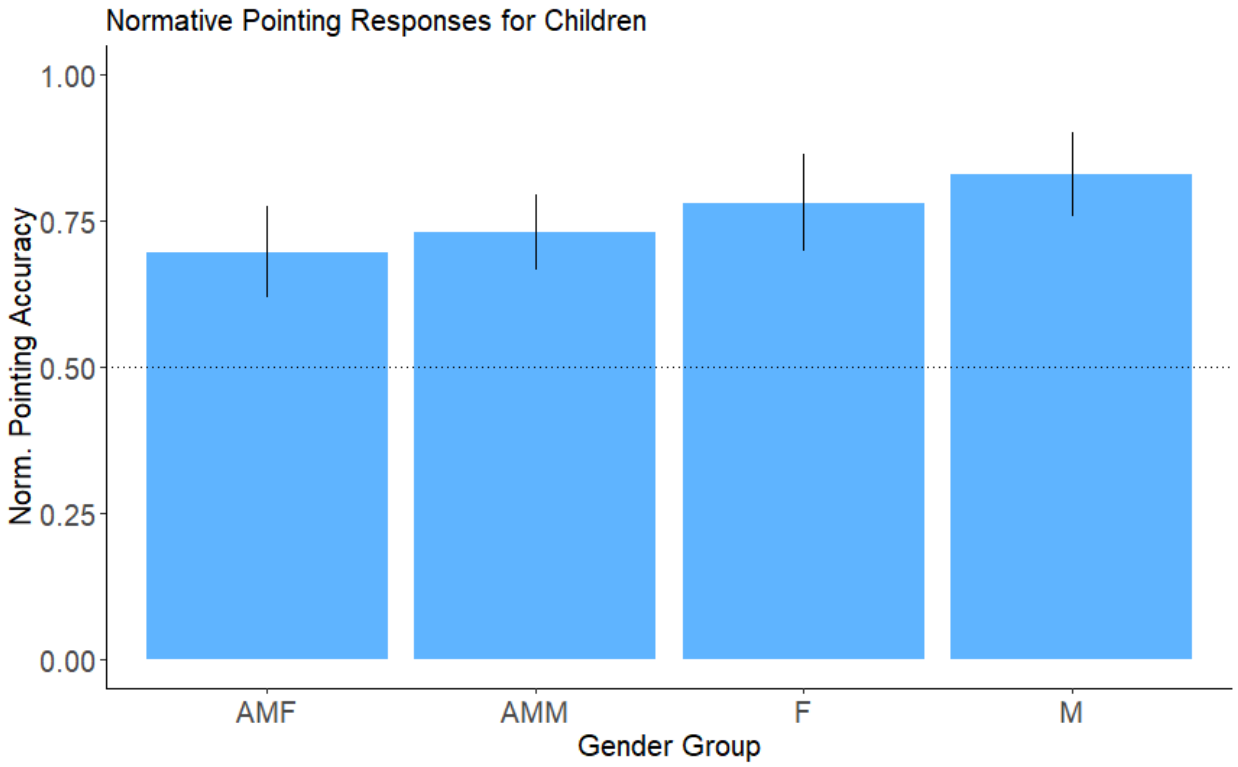

Fig. 3.3: Children on how "accurate" they are in matching adults' normative inferences. We can see overall that their "accuracy" is lower than adults, meaning that children do not align with adults on their inferences of speaker gender.

Additionally, we see that children were about equal in their normative "accuracy" with both ambiguous female and ambiguous male speakers. Compared to adults, children are more uncertain about their inferences of ambiguous speakers, regardless of inferred gender identity. Children are better at identifying typical speakers, but not in a significant way.

A t-test yielded a p-value of 0.144, suggesting that the difference in children's normative inferences between ambiguous and typical speakers is not statistically significant. However, since this is based on preliminary data, we cannot definitively conclude that this difference will remain statistically insignificant.

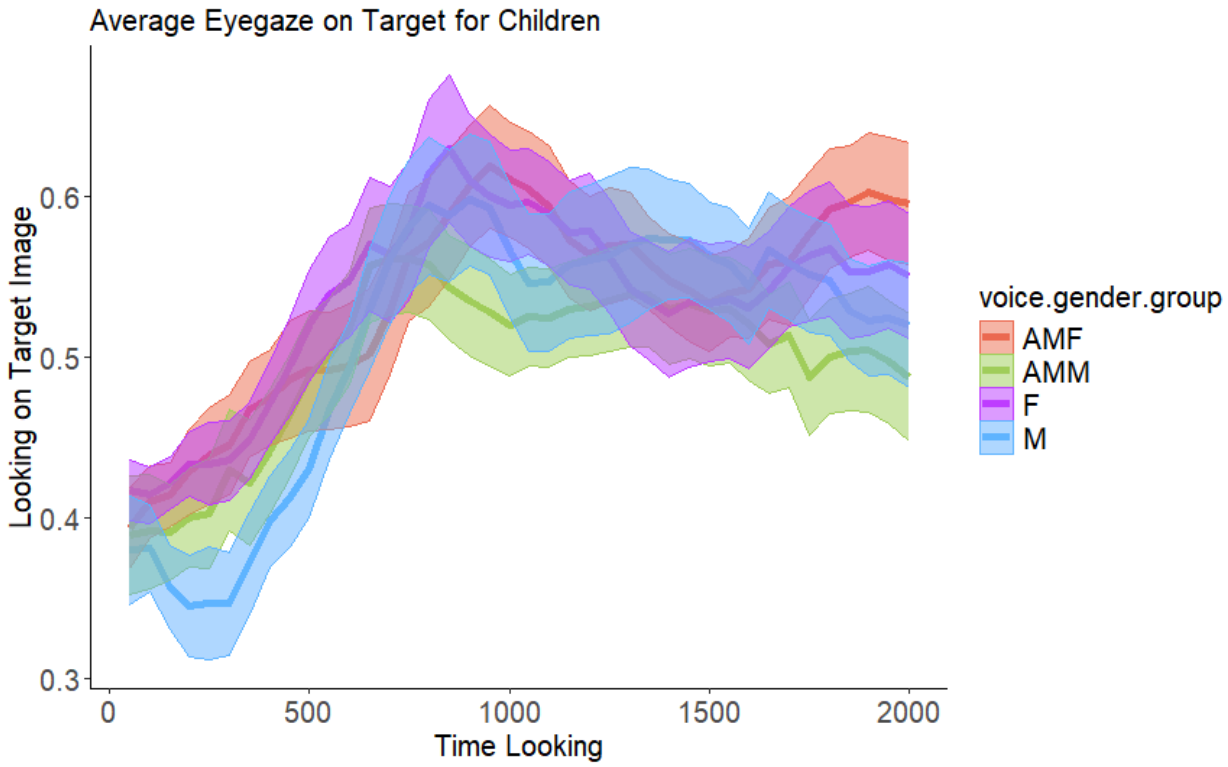

Fig 3.4: Children's fixation on target image. Overall, there seems to be no significant difference between the ambiguous speakers and the typical speakers.

4.1 Discussion

In this study, we investigated whether children recognize gender presentation based on how someone sounds, and if their ability to determine gender changes based on the gender voice ambiguity of the speaker.

For our adult participants, we found that listeners made inferences about a speaker's vocal gender presentation similar to those of the listeners in Lang (2023) and each other. Overall, adult speakers make similar perceptual inferences about speakers, as we saw our adult participants have inferences aligned with how participants in Lang (2023) ranked them. Additionally, listeners made more divergent inferences about speakers' identities for speakers rated as more ambiguous.

Overall, 3-to-5-year-old children diverged from adults' set normative responses and demonstrated lower certainty of gender identity. Children made fewer responses that were normative to adults and the differences between response "accuracy" for gender-ambiguous speakers and gender-typical speakers were statistically insignificant. Although children made adult-akin inferences about the gender of male speakers, on average not all children made these "accurate" inferences for male speakers. This suggests children overall made divergent inferences about speakers, regardless of how adults had determined the gender speaker identity

to be.

Furthermore, children's inferences about all gender groups had no statistical difference. This means that children may not have solidified ideas of what specific gender groups look like. These initial findings reflect previous research that found that children solidified their perception of gender around the ages of 5 to 7 (Trautner et al 2005). It appears that around the ages of three, four, and five, children still may not have established normative beliefs about gender, which may explain why their answers for both typical and ambiguous speakers overlap more greatly, and why answers are less normative overall for both typical and ambiguous speakers.

However, it is important to take note that, while perceivers make similar perceptions of gender identity based on voice as other perceivers, it does not mean that they are all making the *correct* inferences about gender. For instance, one speaker categorized consistently as female was actually a transgender man. Additionally, two speakers categorized as ambiguously male were women (one transgender, one cisgender), and one speaker was a non-binary person. While adult perceivers may be able to detect inconsistencies between appearance and assumed ambiguity in voice (as shown through longer times to make inferences), this does not mean that perceivers can always accurately reconstruct someone's gender identity.

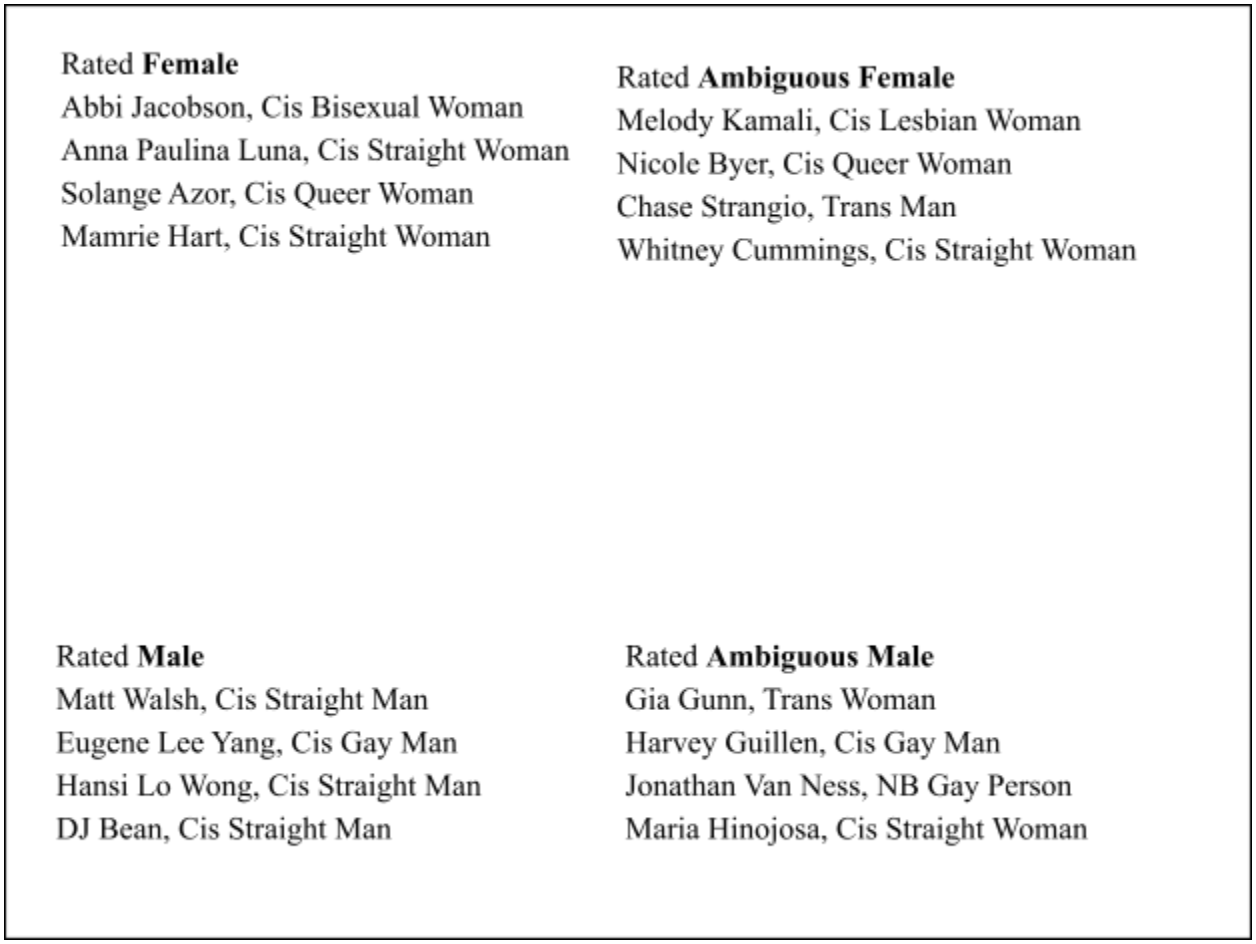

Fig 4.1: Speakers were chosen from publicly available podcasts (list included in 2.2) and were categorized into four groupings: female, male, ambiguous female, and ambiguous male. Here is the list of speakers in their respective groupings and their real publicly documented identities. Cis is short for Cisgender, Trans is short for Transgender, and NB is short for Non-binary.

4.2 Limitations

One limitation of this study is that there is no way to control the reasoning behind why the listener made the inference that they made. While making their inferences, children may have been randomly pointing due to uncertainty, while others may have been robust in their choices. Without this knowledge, we understand less the motivations behind children's inferences.

Moreover, some child participants may have been uncertain and confused about what they were supposed to do, potentially impacting their responses. However, preliminary data analysis revealed that these participants did not select speakers randomly overall.

A big limitation of this study is that it provides a highly simplified version of what it means to be gender non-conforming or queer. Gender presentation is a complex topic that requires a highly interdisciplinary explanation that a paper restricted to a year of research can not encompass.

Gender presentation encompasses a multifaceted spectrum of behaviors, mannerisms, and appearances that are socially linked to gender identity. Voice alone cannot fully convey someone's entire gender presentation, highlighting the study's narrow focus.

Additionally, a limitation arises from the disconnect between the perceived sexual orientation of a speaker and the perceived sexual orientation of a person in an image. If a sound file sounds like it came from a homosexual man, but the associated image does not look like a homosexual man, then this would induce greater hesitation from the participant to select the man in the binary choice.

Another limitation is the population from which we drew our sample size. Our sample size for adults consisted of college students and young children in southern California, historically a more diverse environment than other places in the United States. This limits the study's generalizability to less diverse populations in the United States.

4.3 Impact and Future Studies

The results illustrate a better understanding of how psycholinguistic mechanisms underlie social categorization of gender. Additionally, this study helps push for more research that better encompasses our understanding of gender presentation outside of the binary.

In the future, it would be interesting to study the perceptions of adults and children from different geographic regions in the US. More broadly, various countries hold different perceptions of masculinity and femininity in their voices, so it would be informative for a wider breadth of people to see how perception changes in different nations (Enaifoghe 2023).

To learn more about the development of gender perception, it would be interesting to study different age groups that are slightly older. Studying six to eight-year-olds, who would have more solidified concepts of gender, would lead us to a greater understanding of child development.

Additionally, this study could be replicated on a greater scale to see if the initial results hold up. With a larger sample size, we could see greater possible effects if they differ with age groups or social groups. For instance, would a 50-year-old homosexual man perform better than a 24-year-old heterosexual woman? Would a 27-year-old nonbinary queer individual perform about the same as a 27-year-old cisgender queer individual? Learning more about the intersections of age groups and LGBT+ identities in terms of perception of gender would contribute to a better understanding of how queer communities identify each other, and if this perception changes depending on age demographic.

We can also reexamine the study in different methods, such as applications to online or digital

versions, to see differences in online interaction and perception of gender. This could serve as a continuation of Creel (2012), which also looked at preschoolers' usage of talker information in online conversation and comprehension.

5.0 Acknowledgments:
Thank you to my advisor Dr. Sarah Creel for her support, guidance, and incredible patience in mentoring me and guiding the development of my research. Thank you to the rest of my honors advisory board: Dr. Marc Garellek, Dr. Will Styler, and Benjamin Lang. Furthermore, I would like to thank Lang for lending his data and stimuli to me. I would also like to thank Dr. Bradley Voytek for his advice and support as the Cognitive Science Honors advisor, and his instrumental help in developing my final presentation and paper.

Thank you to the Language Acquisition and Sound Recognition Lab for housing my research and thanks to all of the undergraduate research assistants who helped me collect my stimuli and data. Alex Farrow, Caroline Hall-Sherr, Carly Villongco, Chelsea Weisbrich, Advika Shankar, Kaitlyn Chou, and Mina Cicekoglu — without your help, I would not have this study running as smoothly as it did.

Last but not least, I'd like to thank my friends and family who supported and cheered me on. And a special thank you to my Connor, who would feed me when I forgot to eat.

References

Bacon, Desia, and Jenny Saffran. "Role of Speaker Gender in Toddler Lexical Processing." *Infancy*, vol. 27, no. 2, 22 Jan. 222, pp. 291–300, https://doi.org/10.1111/info.12455.

Borovsky, Arielle, and Sarah Creel. "Children and Adults Integrate Talker and Verb Information in Online Processing." *Developmental Psychology*, vol. 50, no. 5, 2014, pp. 1600-1613.

Creel, Sarah. "Preschoolers' Use of Talker Information in On-Line Comprehension." *Child Development*, vol. 83, no. 6, Nov. 2012, pp. 2042–2056.

Enaifoghe, Andrew. "The Influence of Culture and Gender Differences in Communication: Society's Perception." *International Journal of Research in Business and Social Science*, vol. 12, 2023, pp. 460-468, https://doi.org/10.20525/ijrbs.v12i7.2720.

Karras, Tero, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." *arXiv*, https://arxiv.org/abs/1812.04948.

Klatt, D. H., and L. C. Klatt. "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers." *The Journal of the Acoustical Society of America*, vol. 87, 1990, pp. 820–857.

Lang, Benjamin. "Reconstructing the Perception of Gender Identity, Sexual Orientation, and Gender Expression in American English." *International Congress of Phonetic Sciences,* Aug 2023, https://www.researchgate.net/publication/374472030_Reconstructing_the_perception_of_gender_identity_sexual_orientation_and_gender_expression_in_American_English

Mullenix, John W., et al. "The Perceptual Representation of Voice Gender." *The Journal of the Acoustical Society of America*, vol. 98, 01 Dec. 1995, pp. 3080-3095, https://doi.org/10.1121/1.413832.

Nagels, Arne, et al. "Development of Voice Perception is Dissociated Across Gender Cues in School-Age Children." *Scientific Reports*, vol. 10, 2020, article 5074, https://doi.org/10.1038/s41598-020-61732-6.

Planned Parenthood. "What Do I Need to Know About Transitioning?" *Planned Parenthood*, 2024, www.plannedparenthood.org/learn/gender-identity/transgender/what-do-i-need-know-about-transitioning.

Perry, Tracy L., Ralph N. Ohde, and Daniel H. Ashmead. "The Acoustic Bases for Gender Identification from Children's Voices." *The Journal of the Acoustical Society of America*, vol. 109, 2001, pp. 2988–2998.

Schweinberger, Stefan, et al. "Auditory Adaptation in Voice Perception." *Current Biology*, vol. 18, 06 May 2008, pp. 684-688, https://www.cell.com/current-biology/pdf/S0960-9822(08)00454-5.pdf.

Titze, Ingo R. "Physiologic and Acoustic Differences Between Male and Female Voices." *The Journal of the Acoustical Society of America*, vol. 85, 1989, pp. 1699–1707.

Trautner, H. M., Ruble, D. N., Cyphers, L., Kirsten, B., Behrendt, R., & Hartmann, P. (2005). Rigidity and flexibility of gender stereotypes in childhood: Developmental or differential? *Infant and Child Development, 14*(4), 365–381. https://doi.org/10.1002/icd.399