Gender Typicality in Children's Speech: A comparison of the Speech of Boys with and without Gender Identity Disorder

Benjamin Munson, Laura Crocker, Janet B Pierrehumbert, Allison Owen-Anderson, and Kenneth J. Zucker.

Journal of the Acoustical Society of America
to appear May 2015

Running Head: Gender Typicality in Children's Speech

Gender Typicality in Children's Speech: A comparison of Boys with and without Gender Identity Disorder

Benjamin Munson

Laura Crocker

Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin Cities

Janet B. Pierrehumbert

Department of Linguistics, Northwestern University

Allison Owen-Anderson

Kenneth J. Zucker

Centre for Addiction and Mental Health, Toronto, Ontario

Please Direct Correspondence to:

Benjamin Munson
Department of Speech-Language-Hearing Sciences
University of Minnesota
115 Shevlin Hall
164 Pillsbury Drive, SE
Minneapolis, MN55455
(612) 624-0304
Fax: (612) 624-7586
Munso005@umn.edu

Abstract

This study examined whether boys with Gender Identity Disorder (GID) produced less prototypically male speech than control boys without GID, a possibility that has been suggested by clinical observations. Two groups of listeners participated in tasks where they rated the gender typicality of single words (group 1) or sentences (group 2) produced by 15 5-13 year old boys with GID and 15 age-matched boys without GID. Detailed acoustic analyses of the stimuli were also conducted. Boys with GID were rated as less boy-like than boys without GID. In the experiment using sentence stimuli, these group differences were larger than in the experiment using single-word stimuli. Listeners' ratings were predicted by a variety of acoustic parameters, including ones that differ between the two groups and ones that are stereotypically associated with adult men's and women's speech. Future research should examine how these variants are acquired. [PACS codes 43.70.Ep, 43.71.Bp]

I. Introduction

Listeners can reliably identify many attributes of talkers from patterns of pronunciation. One such attribute is a talker's gender. Gender is so robustly encoded in spoken language that talkers' genders can be identified by automatic classification algorithms nearly perfectly from the acoustic characteristics of short segments of the vowel /ɛ/ (Bachorowski & Owren, 1999). The majority of gender differences in speech are arguably the result of sex differences in the speech-production mechanism. For example, men, on average, have longer vocal tracts and longer, thicker vocal folds than women (Fitch & Giedd, 1999; Titze, 1989). These result in lower-frequency resonance peaks in men's vowels than in women's and in a lower-pitched voice. Other differences, however, are clearly not the consequence of anatomical variation between the sexes. These include the differences between men and women's speech that occur only in certain age groups and socioeconomic strata, such as those documented in sociolinguistic studies. For example, Stuart-Smith (2007) showed that the extent of differences in /s/ production between men and women in Glasgow varied as a function of the speakers' age and social class. Even some gender differences in speech that appear to be the consequence of anatomical differences are arguably exaggerated or attenuated in language- and culture-specific ways. Johnson (2006) showed that sex differences in vowels' formant frequencies varied across a range of typologically diverse languages, even when population differences in height (which is correlated with vocal-tract size, Chern, Wong, Chu, & Ho, 2002) were controlled. If sex differences were solely the consequence of anatomical differences, then we would predict them to be equivalent across languages and cultures.

Taken together, these findings suggest that gender differences in speech are, in part, the result of learned, culturally and linguistically specific behaviors. This hypothesis is further

supported by the finding that boys and girls speak differently well in advance of the anatomical differences that account for some gender-specific speech differences in adults. This was illustrated by Perry, Ohde, and Ashmead's (2001) study of adults' ratings of the gender typicality of children's speech. Perry et al. found that naïve listeners rated the speech of boys and girls as young as four years old as sounding different. Perry et al. also noted that there were no significant differences in measurements of body size (height, weight, sitting height, and neck circumference) between the 4-year-old girls and boys in the study. These measurements of physical size have been shown to correlate with vocal tract size (Bennett, 1981) and therefore with a speaker's resonant frequencies (Fitch & Giedd, 1999). While early studies of vocal-tract development found no significant differences in size or shape between prepubescent boys and girls (Fitch & Giedd, 1999), more recent anatomical studies have shown some gender differentiation in vocal-tract morphology prior to puberty (Vorperian, Kent, Lindstrom, Kalina, Gentry, & Yandell, 2005), leaving open the possibility that some early gender differences are indeed the consequence of anatomical differences. However, gender-linked anatomical differences are less likely to be the cause for other early gender differences in speech, such as those documented in a sociolinguistic study of the acquisition of phonetic variation in and around Newcastle, England reported in Foulkes, Docherty, and Watt (2005; see also Docherty, Foulkes, Tillotsen, & Watts, 2005). Foulkes et al. (2005) found that, for the 24 children aged 2 to 3, there was no significant sex difference in the use of a gender-marked feature in that dialect, preaspiration. In contrast, for children between the ages of 3.5 and 4, girls produced significantly more preaspiration than did boys. This pattern mirrors the gender differences observed by Docherty et al. (2005) in the adult community and indicates that children increasingly produce adult-like gender-specific ways of speaking as their development progresses. These findings also

suggest that, even at very young ages, children have implicit knowledge that adult men and women speak differently. In part, they may acquire variants by emulating a subset of the individuals that they encounter during language acquisition.

The focus of this investigation was on gendered speech in boys with Gender Identity Disorder (GID) (American Psychiatric Association, 2000). A diagnosis of GID is typically made when a child shows distress or discomfort at her or his biological status as a female or male, evidenced by a stated desire to be of the opposite sex or by other signs of gender dysphoria (Zucker & Bradley, 1995). However, children can also be diagnosed with GID by showing gender-nonconforming preferences in terms of their cross-gendered behavioral interests and preferences, the sex composition of their peer group, or their choice of clothing. The GID population comes to clinical attention because of concern about the children's well-being on the part of adults (parents, teacher, family doctor, another mental health professional, etc.). The participants with GID in this study were identified because their behaviors were sufficiently troubling to their parents/caregivers or others to lead to a formal evaluation and, potentially, diagnosis and counseling services. Long-term investigations of boys with GID (e.g., Green, 1987; Money & Russo, 1979; Singh, 2012; Wallien & Cohen-Kettenis, 2008; Zuger, 1984) show that a considerable majority of them identify as GLB or transgendered as adults or to elect gender reassignment surgery. Clinical impressions of boys with GID suggest that their speech is less gender typical than that of their peers without GID (Zucker, 1992). This observation was one impetus for this study.

The purpose of this study was threefold. First, it provided a rigorous experimental test of the clinical observation that the speech of boys with GID is less gender-typical than that of boys without GID. Second, it examined whether these ratings varied as a function of the age of the

child being rated. Any differences between the groups would provide insight into the extent to which learning of gender-typical speech styles progresses in the age range we investigated.. Third, this study examined the perceptual cues that listeners used when rating the gender-typicality of children's speech. It did so by comparing the perception of stimuli that varied in their linguistic structure, both in their phonemic content and their overall linguistic complexity. The inclusion of words as stimuli allowed us to assess whether judgments were linked to the pronunciation of specific vowels and consonants. The inclusion of sentence stimuli allowed us to examine whether boys with GID also produced distinctive patterns of more global, sentence-level prosodic features, like patterns of tempo and fundamental frequency. While GID can be identified in both girls and boys, the focus in this study was on boys with GID. Our impetus for this was twofold. First, examining the speech of boys with GID allowed us to examine whether their distinctive pronunciation patterns mirrored those of adult men who identify as gay. In this way, our examination of boys with GID is an indirect examination of the early acquisition of gay male speech styles. Second, there is a higher incidence of GID in boys than in girls, allowing for the recruitment of a larger number of participants (Wood et al., 2013).

II. Method

A. Talkers

Speech data were obtained from 30 boys living in Toronto, ON, Canada. The first group consisted of 15 boys who were referred for clinical assessment and were diagnosed with GID based on information provided by parents (American Psychiatric Association, 2000). The second group consisted of 15 control boys with typical gender development. For comparative purposes, control males will henceforth be referred to as Typical Gender Development (henceforth TGD). The control participants were recruited from a day-care facility and were tested at the Center for

Addiction and Mental Health (CAMH) in the same environment as the boys with GID. They were selected as controls because they did not have a diagnosis of GID. In general, participants were chosen because they were native speakers of English without any history of speech or language disorders. All talkers spoke the dialect of English common to southern Ontario, which is the dialect represented in Canadian national broadcast media. This dialect shares many features with the variety of American English spoken by the listeners in the perception study, that of the upper Midwestern US (Labov, Ash, & Boberg, 2006).

Talkers were all between 5.7 and 12.8 years of age and had all achieved a Verbal IQ within normal limits, as measured by the Weschler Intelligence Scales for Children-III (WISC-III) (Wechsler, 1991). GID boys did not differ significantly in age from TGD boys (M_{GID}= 8.4 years, SD_{GID} = 2.0 years; M_{TGD}= 8.5 years, SD_{TGD}= 2.0 years) nor was there a significant difference in Verbal IQ as a function of diagnosis (M_{GID}= 101.87, SD_{GID} = 13.5; M_{TGD} = 99.8, SD_{TGD} = 14.13). In the perception task, stimuli from a group of younger girls were included as control items. These talkers were 5 to 7 years old, i.e., similar in age to the younger group of GID and TGD boys. Their speech samples were collected in the Minneapolis-St. Paul area for another unrelated study (Munson, Baylis, Krause, & Yim, 2010). The single-word perception task was designed and run before the sentence-perception task was designed. In the design of the latter experiment, it was found that four of the participants (3 TGD, 1 GID) who produced an acceptable number of real words had either refused to participate in the sentence-perception task, or produced utterances that were not sufficiently fluent or accurate to be included as stimuli. Sentence productions from four other talkers with similar ages and full-scale IQs were used instead.

B. Speech Materials

Two sets of speech materials were collected from the children: single words and sentences. A subset of these materials was used as stimuli in a perception experiment. The target words elicited were *bee*, *bird*, *boot*, *bus*, *cake*, *fish*, *foot*, *hat*, *rope*, *ship*, *sock*, *spoon*, *sun*, and *tent*. These 14 total target words are all monosyllabic and 11 have CVC structure. They were selected to contain a variety of vowels, as well as many words with the sibilant fricative /s/. The inclusion of /s/ was motivated by previous research indicating that the acoustic characteristics of these sounds differ as a function of gender and are significantly associated with variation in judgments of gender typicality (Linville, 1998; Munson, McDonald, DeBoe, & White, 2006;). The target words were all high frequency common nouns and were also chosen for the ease with which they could be represented pictorially on large cards.

The second set of productions consisted of 17 short (5-10 word) audio sentences that the children were instructed to repeat. The set included ones that were composed of primarily sonorant consonants (n=4, *I owe you a yo-yo, The yellow lion roared*), ones that had many instances of /s/ (n=9, *The squirrel sat on the see-saw*, *The salad had celery on it*, *The snowman wore a sweater*), and an additional four sentences with obstruent consonants other than /s/ (i.e., *I bumped my arm on a nail*). The sonorant consonant set was included under the assumption that the f0 would be present throughout the productions of these, and hence would be easiest to perceive and measure. The /s/ sentences were included to test the hypothesis that the /s/ characteristics cue judgments of gender. The 'other' sentences were designed to ensure that a variety of vowels would be elicited in words in prosodically strong positions. In this article, we compared the listeners' responses as a function of the three sentence types. We also compared the listeners' ratings with measures of f0 and temporal variation to examine the extent to which

judgments were influenced by sentence-level prosodic structure. Multiple tokens of these sentences were recorded by a speaker of general Canadian English. The most fluent and natural-sounding token of each was selected as a stimulus for the sentence-repetition task.

The 15 talkers in the GID group completed a picture-naming task onsite at CAMH. The task was administered by a psychometrist who had been trained by author KJZ to administer the task protocol. The psychometrist was familiar with the children and was not blind to their diagnoses. For the word-production task, large pictures on pages of a flip-book were presented to the talkers one at a time. There were 14 separate pages, for a total of 14 different pictures, and each talker repeated the series approximately 5 times. Talkers were instructed to say only the name of each object aloud (i.e., introductory remarks such as *that is a…, that's a picture of a…* were to be omitted). If a talker produced an unexpected word (for example, 'boat' instead of 'ship'), they were given a prompt to elicit the desired target word.

For the sentence repetition task, sentences were played from a portable CD player. Children were instructed to repeat the sentence exactly as they heard it as soon as it played. Five separate randomizations of the set of sentences were played. For the GID boys, both the word- and sentence-repetition tasks took place in a standard consultation room in the CAMH building. Talkers wore an AKG C420 head-mounted micro-mic attached through a Rolls phantom power source to a Marantz CDR330 CD recorder. Recordings were made at 44.1 kHz sampling rate with 16-bit quantization and were processed through a low-pass filter with an upper cutoff of 22.05 kHz to prevent aliasing. The control talkers were also recorded at CAMH.

A subset of the children's productions was used as stimuli in the perception tasks. From each of the 30 talkers, one production of each of the 14 single words was chosen quasi-randomly. A subset, rather than the full set of approximately 70 words spoken by each talker, was used in

order to make the experiment reasonable in length and to ensure that asymmetries in the number of usable tokens across talkers did not influence the results. The tokens were free from extraneous noise, such as a page turning or microphone rustling. Similarly, a production of each of the sentences was chosen quasi-randomly. A research assistant listened to all of the sentence tokens and picked one fluent token with no repetition errors or added words. This resulted in the exclusion of one especially long sentence (*I brought my suitcase and skis on vacation*) that had no usable tokens from many of the younger participants.

A variety of acoustic measures were made of the stimuli. For the single-word productions, the onset and offset of the initial consonant and the vowel were hand-labeled. The Bark-scaled F1 and F2 center frequencies of each of the vowels were extracted automatically using the LPC formant tracker in Praat (Boersma & Weenink, 2001). The f0 frequency in ERB units (Hermes & van Gestel, 1991) was extracted automatically at the vowel midpoint using the pitch-track function in Praat. Vowel duration was also logged. Selected acoustic characteristics of the word-initial /s/ from the *sock* and *sun* productions were also measured. These were the first three spectral moments of the 40 ms interval of frication noise centered at the fricative's midpoint. These values were chosen because they have been shown to differentiate /s/ from other fricatives of American English and because they differ as a function both of talker sex (Jongman, Wayland, & Wong, 2000) and listener-identified voice gender typicality (Munson et al., 2006; Munson, 2007). These values are shown in Table 1.

Three measures were taken of the sentence stimuli. The first of these was raw duration. The second and third were measures of f0. An f0 track of the entire sentence was generated in Praat, then was manually inspected to ensure that there were no mistrackings. The median f0 and

interquartile range of f0 in ERB were logged. These are shown in Table 2. Table 2 also shows the average speaking rate in syllables per second.

A series of statistical analyses examined whether the acoustic measures differed as a function of group and age. For the single-word stimuli, these were linear mixed-effects regressions in which the dependent measures were the eight acoustic measures of the words. The fixed effects were group and age in months. There were random intercepts for talker and word. The results of the subset of statistical models where either age or group were statistically significant are shown in Table 3. Significant effects of group were found for F2 frequency and f0 of vowels, and for the second spectral moment (m2) of /s/. The words produced by the boys with GID had a higher F2 frequency, a lower f0, and a higher m2 (i.e., a more-diffuse /s/ spectrum) than those of TGD boys. High m2 values are characteristic of frontally misarticulated (i.e., /θ/-like) tokens of /s/ (Baum & McNutt, 1990). Hence, this finding might indicate that the boys with GID are more likely than boys with TGD to misarticulate /s/. Informal inspection of the F2 differences between groups for individual word suggested that the F2 effect was not due to the production of a single vowel: large group differences were found for the vowels in the words *bird*, *foot*, and *sock*, but not for the featurally similar vowels in the words *boot* and *bus*. This finding makes it unlikely that the group differences were a consequence of differences in vocal-tract size or shape (a possibility suggested by the finding that boys with GID with older siblings have a lower birth weight than TGD boys with a similar birth order, Blanchard, Zucker, Cavacas, Allin, Bradley, & Schachter, 2002), which would lead to similar-magnitude differences for similar vowels. A related set of analyses examined three acoustic measures of the sentences: duration, median F0 and f0 interquartile range, with age in months and group as fixed effects and

talker and sentence as random intercepts. There were no effects of age or of group on any of these three analyses.

C. Listeners

Listeners were recruited via fliers posted on the University of Minnesota campus advertising for native speakers of English, 18-50 years old, with no current or past speech, language, and hearing disorders. There were 21 listeners in the word perception experiment and 17 in the sentence-perception experiment. The University of Minnesota student body comes largely from Minnesota, Western Wisconsin, and Eastern North and South Dakota. These areas are part of the North dialect region, as described by Labov, Ash, and Boberg (2006). This dialect region shares many features with the dialect spoken in Toronto, where the children were recorded. Moreover, studies of less-masculine-sounding male speech conducted in Minneapolis (e.g., Avery & Liss, 1996; Munson et al., 2006) and Toronto (e.g., Smyth, Jacobs, & Rogers, 2003) found that similar acoustic parameters characterize the speech of gay-sounding men and cue judgments of sexual orientation. This suggests that the perception study taking place in different locations likely had only a minimal effect on the results of the current study.

D. Procedure

The perception experiment was conducted in a sound-proof booth. Both experiments were written and carried out using the E-prime experiment management software (Schneider, Eschman, & Zuccolotto, 2002). Speech tokens were RMS normalized and presented to the listeners at approximately 65 dB HL via headphones. On each trial, one word was presented over headphones and simultaneously displayed on a 17" computer screen in 36-point courier font. After hearing a word via headphones, listeners selected one of six possible categories by pressing the corresponding key on the numeric keypad. The 6-point scale developed by Perry et al. (2001)

was used: 1=*positively a female*, 2=*appeared to be a female*, 3=*unsure; may have been a female*, 4=*unsure; may have been a male*, 5=*appeared to be a male*, 6=*positively a male*. Listeners responded by selecting a key on a numeric keypad, and the responses were recorded automatically by E-prime. Both experiments were preceded by 5 fully randomized practice trials using items that were taken from Munson et al. (2010). After every trial, a copy of the 6-point scale in 36-point courier font appeared on the computer screen so that memorization of the scale would not be a confounding factor in the listeners' responses. In the word perception task, listeners heard all 30 talkers (15 GID males and 15 TGD males) say each of the 14 target words once. In addition, the female talkers described earlier spoke a total of 10 words for which listeners also provided judgments. These stimuli were included to anchor the 6-point scale with examples of speech that were in fact spoken by females. The 430 tokens were presented randomly to each listener, both in terms of word order within the experiment and of word order from one listener to the next. The design of the sentence-perception experiment was similar, though it was somewhat longer, both because there were more tokens (n=485) and because the tokens themselves were longer.

III. Results

A. Group Differences

The first analysis examined whether boys with GID and TGD were rated differently from one another. This analysis addressed our first research question of whether the clinical impression that boys with GID have less prototypically masculine sounding voices would be confirmed in a controlled experimental study.

For this analysis, a linear mixed-effects regression (LMER) was used. Separate LMERs were calculated for word- and sentence-perception data. The dependent measure was the rating

on the 1-6 scale. The independent measures were group (dummy coded using contrast coding), age in months, and their interaction. Random intercepts were included in the model for subject and for stimulus. Given the restricted range of responses, the model used a Poisson distribution for the dependent measure. The model was fit using the lme4 package in R. The significance of individual terms was assessed using the lmerTest package in R. Results of the models are shown in Table 4. As these models show, both group and age affected ratings. Boys with GID were rated as less prototypically boy-like than TDG controls for both word and sentence stimuli. Moreover, age in months affected ratings: ratings for both groups of children became more boy-like with age.

Group differences in gender typicality ratings are shown in Figures 1 and 2. As these figures show, both groups elicited the full range of ratings from 1 to 6 for both stimulus types. Ratings for the participants are pooled across age because of the lack of a statistically significant interaction between group and age in months in the statistical models presented in Table 3. Figures 1 and 2 show that there were bigger group differences in ratings for sentence stimuli than for ratings of single words. The filler items from girls' speech were not included in the analysis because they were not matched in age to both groups of boys being rated. However, we can compare the ratings of the girls to the ratings of the younger boys. The average rating for the girls' voices was 2.58. The mean for the comparable-aged boys with GID in this study was 3.47. The mean for ratings of comparable-aged TGD control boys was 3.94. We can also compare both groups' ratings of the GID and TGD boys in this study to the values from Perry et al.'s (2001) study of the perception of girls' and boys' voices. In that study, 4-year-old girls' voices were given a rating of approximately 3, and boys' voices were given a rating of approximately 4. Eight-year-old girls' voices were rated approximately 3 and boys' voices were rated

approximately 4.5. Hence, the separation between the GID and TGD boys' voices was smaller than that between TGD boys and TGD girls in other, comparable studies of TGD boys' and TGD girls' speech.

B. Predictors of Ratings

The last set of analyses aimed to determine the acoustic characteristics of the stimuli that predicted variation in ratings. In this analysis, the data from the GID and TGD boys' ratings are pooled. The first of these was a linear mixed-effects regression predicting each word's perceived gender rating from various acoustic measures. Prior to this analysis, ratings for individual words were examined. The differences between the groups in average ratings for each word ranged from 1.1 for the word *sun* to 0.07 for the word *cake*. Six words had differences greater than 0.5: *sun*, *sock*, *bus*, *fish*, *spoon*, and *bird*. This list includes five of the six words with sibilant fricatives in them, and the word *bird*, which had one of the largest group differences in F2 frequency. The first LMER used all of the listeners' ratings of all of the words as the dependent measure. The fixed effects were vowel duration, F1 and F2 frequency, and f0 in ERB. Listener, talker, and word were included as random effects. The dependent measure was modeled as a Poisson distribution, and the lmerTest package was used to assess significance. The result of this model is shown in the upper portion of Table 5. The coefficients showed that tokens were more likely to be rated lower (i.e., less boy-like) when they had high f0, high formant frequencies, and a long duration, the latter perhaps reflecting a percept of greater overall precision in articulation.

A second model examined the ratings for only the two words that contained word-initial /s/, *sock* and *sun*. This model had four variables in addition to the four used in the model for all of the words: /s/ duration, first spectral moment, second spectral moment, and third spectral moment. Given the relatively large number of predictor measures and the small number of

dependent measures, this analysis might be subject to overfitting, and hence should be considered exploratory rather than confirmatory. The coefficients for these additional four variables are shown on the bottom of Table 5. The coefficients for the four vocalic variables are not shown because they were qualitatively similar to those in the full model for all stimuli. The coefficients for /s/ show that higher centroid frequency (m1), greater spectral diffuseness (m2), and a more negatively skewed spectrum (m3) of the fricative are associated with judgments of less boy-like speech. The high m1 of /s/ differentiates it from the most similar fricative, /ʃ/. Hence, tokens of /s/ that were especially clearly articulated were associated with judgments of less boy-like speech. The m2 of /s/ is generally low, and high m2 values are associated with frontally misarticulated, /θ/-like tokens of /s/. Hence, tokens that resembled these frontally misarticulated /s/ were rated as sounding less boy-like. This pattern is similar to the findings of Mack and Munson (2012) on the relationship between fricative quality and the perceived sexual orientation of adult men's voices. In that study, men were rated as gayer-sounding if their speech contained tokens of /s/ with high peak frequencies or highly distributed spectra. This may reflect the combined activation of two seemingly conflicting stereotypes about less-masculine speech: that it is hyperarticulated and that it exhibits a distinctive articulation of /s/, which is colloquially described as "lisped." The relationship between m3 and judgments is similar to that seen in adult men's voices, as reported by Munson, McDonald, DeBoe, and White (2006).

The inclusion of word as a random effect in the two LMERs minimized the possibility that the effects seen in these analyses were due to the outlying behavior of a subset of stimuli. However, given our interest in the specific perceptual features that cue judgments of gender identity, we conducted separate linear mixed-effects regressions for individual words. The fixed effects were the same as those in the models grouped by word. Talker and listener were random

effects; word was not, as each model examined only one word. The F1 frequency was a significant factor (as determined by having an $p < 0.05$ using the lmerTest package to assess statistical significance) in the models for *bee*, *foot*, and *ship*. The F2 frequency was a significant factor in the models for *bee*, *bird*, *bus*, *cake*, *foot*, *rope*, and *tent*. Vowel duration was significant in the models for *bus* and *hat*, and m2 was significant in the models for *sun* and *sock*.

The next set of analyses examined sentence ratings. In the analysis of the sentences, the predictors were the stimulus sentence's length in syllables, the duration of the stimulus sentence in milliseconds, the median f0, and the f0 interquartile range. Listener, talker, and sentence stimulus were random effects. The results of this model are shown in Table 6. Stimuli were rated as less boy-like if they had a high f0 and were spoken slowly.

As with the analyses of the words, acoustic analyses of subsets of the stimuli were done to better understand the cues that listeners might use when rating gender from sentences. The sentences were grouped into three types: those that contained primarily sonorant consonants, those with many instances of /s/, and the 'other' sentences. The difference in ratings between the groups was largest for the sentences containing many instances of /s/ (M_{TGD}=4.15, M_{GID}=3.13, difference=1.02). The group differences for the other two sentence types were smaller and comparable to one another (M_{TGD}=4.15, M_{GID}=3.35, difference=0.80 for the sonorant sentences; M_{TGD}=4.1, M_{GID}=3.23, difference=0.87 for the other sentences). The structure of the statistical model used to examine the subsets of sentences was identical to that used for all 17 sentences. The model for the sentences with many instances of /s/ was similar to the model for the entire set of sentences: sentence duration and median f0 were significant predictors of ratings. The model for the 'other' sentences was similar, though the coefficient for median f0 achieved statistical significance only when the less strict $\alpha < 0.10$ level was used. In the model for the sonorant

sentences, f0 IQR was the only significant factor. Sentences with more f0 modulation were rated as sounding less boy-like than those with less modulation.

IV. Discussion

There were three important findings in this study. The first was that the clinical impression that boys with GID sound less prototypically masculine than boys without GID was verified in a controlled experimental study. This was true both for an experiment using single words and one using sentences as stimuli. The second finding was that these group differences were larger when using sentences as stimuli than when using single word productions. The third finding was that the acoustic cues that listeners appeared to be using when rating gender included both variables that differed between the boys with and without GID in this study, and ones that did not. In the case where listeners used variables that did not differ between the groups, they generally labeled features associated with adult women's speech as more girl-like. The mix of these two types of cues is illustrated by the relationship between the acoustic characteristics of /s/ and ratings. Listeners rated /s/ tokens with diffuse spectra (consistent with the production of frontally misarticulated /s/) as less boy-like, consistent with group differences. They also rated tokens of /s/ with high peak frequencies as less boy-like, consistent with differences between adult men and women, and between heterosexual and gay men (Munson et al., 2006). Critically, none of the distinctive features used by the boys with GID were consistent with the notion that they have a global speech-sound disorder. While they did, on average, produce tokens of /s/ that were more consistent with a frontally articulated /s/, there were no was no evidence from the acoustic and perception tasks that their speech was less advanced than that of boys without GID. Indeed, some of the features that listeners associated with GID were associated with more accurate speech, like hyperarticulated /s/.

The results of this study have implications for our understanding of the development of distinctive speech styles more generally. As stated earlier, many boys with GID adopt a gay, bisexual, or transgendered identity as an adult, or elect to have sex reassignment surgery. A retrospective analysis by author KJZ of a study of a sample from CAMH, where these data were collected, found that 70% of boys with GID identified as gay men as adults. Other research has shown that at least a subset of adult gay men produce a distinctive speech style that allows listeners to identify their sexuality at greater than chance levels from phonetic variation alone (Linville, 1998; Mack, 2010; Munson et al., 2006, Pierrehumbert, Bent, Munson, Bradlow, & Bailey, 2004; Smyth et al., 2003). The acquisition of this distinctive speech style is a topic of great interest, as the age at which gay variants are acquired provides a window into the broader cognitive mechanisms that underlie language acquisition.

Early theories of linguistic and social plasticity predicted that a GLB identity could have no effect on patterns of pronunciation, given the age at which GLB identities were thought to be acquired. Lenneberg (1967) claimed that most aspects of spoken language are learned early in life during the so-called critical period for language acquisition, which he believed to end at puberty. This claim has continued to influence debate about language acquisition up to very recently (Komarova & Nowak, 2001; Long, 2005). As of 1992, it was reported that individuals do not identify as GLB until some years later, when adolescence is complete (Remafedi, Resnick, Blum, & Harris, 1992). However, more recent evidence shows that plasticity in both phonetics and gender identity does not follow as rigid a time course as was previously imagined. Studies of second-language acquisition (such as those reviewed in Newport, 1990) have identified declines in learning by individuals as young as eight years old. But, on the other hand, longitudinal sociolinguistic studies have found a sizable minority of individuals who modify

their speech patterns after puberty, particularly with regard to linguistic properties that have social significance (Sankoff, 2004; Sankoff & Blondeau, 2007). In the area of gender identity, same-sex attraction and experiences during adolescence may precede and do not always lead to a GLB identity (Savin-Williams, 2006; Savin-Williams & Ream, 2006). At the same time, however, increasing social tolerance has meant that some people already identify as GLB or transgender by mid- to late adolescence, as can be readily seen from newspaper articles on controversies about school bathrooms, dress codes, and prom dates.

Our youngest talkers with GID fall well within what Lenneberg claimed to be the critical period for language development. These boys produce speech that is perceptibly different from that of their TGD peers. This means that the rudiments of a gay male speech style can be acquired before adopting a gay identity or joining an adult gay male social network. With 10 years being a typical age for the onset of puberty in boys (Windle et al., 2008), it can be assumed that the older boys with GID included some boys who were in early puberty. The fact that the group differences between GID and TGD boys are stable across the age range we examined, while boys in both groups continue to sound more like males than like females, suggests that the phonetic correlates of social identity are learned early, and are maintained through the beginning of adolescence. Even the oldest boys in our study have developed a distinctive speech style before there is any realistic possibility that gay male adult speech is statistically dominant in their everyday experience. Although some adult gay men may have the option of choosing a predominately gay social and work environment, this level of personal freedom is not available to children, who live with their families and attend school. The inescapable conclusion is that language learning is not determined entirely by passive exposure to the ambient language, but involves preferences in attending to or imitating the variety of speech patterns that the child

observes. Crucially, this imitation need not be the imitation of a gay male role model. As argued by Munson and Babel (2007), Podesva (2007), and others, distinctive characteristics of adult gay male speech styles are a subset of the features of heterosexual men and women's speech, and of speech produced with varying degrees of formality and clarity in different social circumstances. Each one of the features (such as the more diffuse /s/ and the raised F2) could be acquired by children emulating it in the speech of one or more adults that provide input. Eckert (2014) discusses this same proposal with respect to differences between male and female speech styles, arguing that they are assemblies of speech variants that convey constellations of social meanings that are associated with men and women.

The findings in this study suggest that a larger-scale study of the speech characteristics of boys with and without GID is warranted. Such a study could answer some of the key questions that the current study raises. One clear finding from this paper is that the speech of boys with GID appears not to be a whole-scale approximation of adult female speech. If it were, we would expect consistently higher f0 in the boys with GID (as opposed to the lower f0 found in the single-word stimuli) and higher average F1 and F2 values. There is evidence that speakers do manipulate all of these variables when actively attempting to sound like the opposite sex: Indeed, these were the patterns found in a recent study of adult men's imitation of female voice gender by Cartei, Cowles, and Reby (2012).

The second question raised by this study concerns the mechanism that underlies the learning of this styles we observed. Research on other aspects of language learning has shown that children selectively learn words from some speakers over others. For example, Koenig and Harris (2005) showed that children were less likely to learn words from a speaker whom the child has observed acting untruthfully. Kinzler, Schutts, DeJesus, and Spelke (2009) showed that

children's social preferences were guided by a speaker's race, and, more strongly, by whether the speaker had an unfamiliar accent. Similar social preferences and tendencies for preferential learning may explain how boys with and without GID learn the speech styles shown in this study. The boys with GID might selectively learn phonetic variants from peers and adults who respond to them most positively or from those adults or peers with whom they share the strongest social bond. Again, this need not be an adult whose sexuality is known to the child. Rather, this could be an adult whose speech projects more general personal attributes that the child wishes to emulate.

A larger-scale study might also help us to overcome two of the weaknesses of this study. The first is that the GID population is self-selected and as such was neither a random sample of the population of boys nor even a random sample of boys who display the behaviors that lead to a diagnosis of GID. Debates about GID in the popular media often feature cases where the parents of children demonstrating GID-like behaviors (such as cross-dressing and gender dysphoria) treat this as normal variation rather than seeking services (e.g., Associated Press, 2013). The boys in this study were identified because their behaviors were sufficiently troubling to their parents/caregivers or to others to lead their family to seek counseling for the child and themselves. This non-random sampling limits the generalizations we can make from this study either about the population of boys with GID or the development of gendered speech behaviors more generally. It prevents us from assessing whether the speech differences between boys with and without GID are the result cumulative learning of phonetic variants or speech styles across development. No study using a clinically referred sample of boys with GID can overcome this weakness entirely. However, a larger sample with a larger set of demographic measures and a large, randomly sampled group of boys and girls without GID could at least control for more

variables than was possible in this study. The second weakness concerns the questions of which group, GID or TGD, is actively acquiring a distinctive style. The arguments in this article imply that the boys with GID are acquiring a distinctive speech style, akin to the distinctive style spoken by gay-identified adult men. However, there is a plausible alternative scenario in which the onus for acquisition is on the TGD control boys. The speech of the boys with GID in this study was more similar to the speech of adult women than is the speech of the boys with TGD. This parallels the difference between the speech of TGD boys and girls: TGD children in general sound more like adult women than adult men. It is logically possible that the onus for acquisition of phonetic markers of gender identity is on the TGD control boys. A study with a larger sample, including both boys and girls with and without GID, could more fully address the question of which children face the biggest challenges in learning gendered speech variants.

Acknowledgments.

This research was funded by the Northwestern University Vice President for Research and by a McKnight Presidential Fellowship from the University of Minnesota. We thank Eden Kaiser for assistance in testing participants. We thank J. Michael Bailey for the idea of this research.

References

Associated Press (2013, May 13). For transgender kids and their parents, finding support at school can be a particular challenge. http://talkingpointsmemo.com/news/for-transgender-kids-and-their-parents-school-challenge. php (Last accessed November 17, 2014).

Avery, J., & Liss, J. (1996). Acoustic characteristics of less-masculine-sounding male speech. *Journal of the Acoustical Society of America, 99*, 3738-3748.

Bachorowski, J., & Owren, M. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America, 106*, 1054-1063.

Baum, S., & McNutt, J. (1990). An acoustic analysis of frontal misarticulation of /s/ in children. *Journal of Phonetics, 18*, 51-63.

Bennett, S. (1981). Vowel formant frequency characteristics of preadolescent males and females. *Journal of the Acoustical Society of America, 69*, 321-328.

Blanchard, R., Zucker, K.J., Cavacas, A., Allin, S., Bradley, S.J., & Schachter, D. (2002). Fraternal birth order and birth weight in probably prehomosexual feminine boys. *Hormones and Behavior, 41*, 321-327.

Boersma, P., & Weenik, D. (2005). *Praat v. 4. 3. 27* [computer software]. Amsterdam: Institute of Phonetic Sciences.

Bradlow, A., Torretta, G., & Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker differences. *Speech Communication, 20*, 255-272.

Cartei, V., Cowles, H., & Reby, D. (2012) Spontaneous voice gender imitation abilities in adult speakers. *PLoS ONE, 7*, e31353

Cherng, C.-H. , Wong, C.-S. , Hsu, C.-H. , & Ho, S.-T. (2002). Airway length in adults: Estimation of the optimal endotracheal tube length for orotracheal intubation. *Journal of Clinical Anesthesia, 14*, 271-274.

Docherty, G. J., Foulkes, P., Tillotson, J., & Watt, D. J. L. (2005). On the scope of phonological learning: Issues arising from socially structured variation. In C. T. Best, L. Goldstein, & D. H. Whalen (eds.), *Laboratory Phonology 8* (pp. 393-421). Berlin: Mouton de Gruyter.

Eckert, P. (2014). The Problem with binaries: Coding for gender and sexuality. *Language and Linguistics Compass, 8*, 529-535.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *Journal of the Acoustical Society of America, 106*, 1511-1522.

Foulkes, P., Docherty, G., & Watt, D. (2005). Phonological variation in child-directed speech. *Language, 81*, 177-206.

Green, R. (1987). *The "sissy boy" syndrome and the development of homosexuality* (p. 1-409). New Haven: Yale University Press.

Hermes, D. J., & van Gestel, J. C. (1991). The frequency scale of speech intonation. *Journal of the Acoustical Society of America, 90*, 97-102.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252-1263.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *Journal of Phonetics, 34*, 485-499.

Kinzler, K., Shutts, K., DeJesus, J., & Spelke, E. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition, 27*, 623-634.

Koenig, M., & Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76,* 1261-1277.

Komarova, N.L., & Nowak, M.A. (2001) Natural selection of the critical period of language acquisition. *Proceedings of the Royal Society B, 268,* 1189-1196.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change* (p. 1-318). Berlin: Mouton de Gruyter.

Lenneberg, E. (1967). *Biological foundations of language* (p. 1-489)New York: Wiley.

Linville, S. (1998). Acoustic correlates of perceived versus actual sexual orientation in men's speech. *Folia Phoniatrica et Logopaedica, 50,* 35-48.

Long, M. (2005) Problems with supposed counter-evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics in Language Teaching, 43*, 287-317.

Mack, S. (2010). A sociophonetic analysis of perception of sexual orientation in Puerto Rican Spanish. *Laboratory Phonology, 1,* 41-63.

Money, J., & Russo, A. J. (1979). Homosexual outcome of discordant gender identity/role in childhood: Longitudinal follow-up. *Journal of Pediatric Psychology, 4,* 29-41.

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech, 50,* 125-142.

Munson, B., & Babel, M. (2007). Loose lips and silver tongues, or, projecting sexual orientation through speech. *Language and Linguistics Compass, 1*, 416-449.

Munson, B., Baylis, A., Krause, M., & Yim, D. (2010). Representation and access in phonological impairment. In C. Fourgeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology 10* (pp. 381-404). New York: Mouton de Gruyter.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics, 34*, 202-240.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science 14*, 11-28.

Perry, T. L. , Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America, 109*, 2988-2998.

Pierrehumbert, J., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The influence of sexual orientation on vowel production. *Journal of the Acoustical Society of America, 116*, 1905-1908.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics, 11*, 478-504.

Remafedi, G., Resnick, M., Blum, R., & Harris, L. (1992). Demography of sexual orientation in adolescents. *Pediatrics, 89*, 714-721.

Sankoff, G. (2004). Adolescents, young adults and the critical period: two case studies from *Seven Up*. In C. Fought (Ed.), *Sociolinguistic variation* (pp. 121-139). Oxford University Press

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 83, 560-588.

Savin-Williams, R. C. (2006). Who's gay? Does it matter? *Current Directions in Psychological Science, 15*, 40-44.

Savin-Williams, R. C., & Ream, G. L. (2006). Prevalence and stability of sexual orientation components during adolescence and young adulthood. *Archives of Sexual Behavior, 36*, 385-394.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime user's guide* (p. 1-208)Pittsburgh: Psychology Software Tools, Inc.

Singh, D. (2012). *A follow-up study of boys with gender identity disorder* (p. 1-324) Unpublished doctoral dissertation, University of Toronto.

Smyth, R., Jacobs, G., & Rogers, H. (2003). Male voices and perceived sexual orientation: An experimental and theoretical approach. *Language in Society, 32*, 329-350.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/in Glaswegian. In J. Cole & J. Hualde (Eds.), *Laboratory Phonology 9* (pp. 65-86). Berlin: Mouton de Gruyter.

Titze, I. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America, 85*, 1699-1707.

van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech, 38*, 253-265.

Vorperian, H., Kent, R., Lindstrom, M., Kalina, C., Gentry, L., & Yandell, B. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America, 117*, 338-350.

Wallien, M. S. C., & Cohen-Kettenis, P. T. (2008). Psychosexual outcome of gender dysphoric children. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 1413-1423.

Wechsler, D. (1991). *The Wechsler intelligence scale for children—third edition*. San Antonio, TX: The Psychological Corporation.

Windle, M., Spear, L., Fuligni, A., Angold, A., Brown, J., Pine, D., Smith, D., Giedd, J., & Dahl, R.(2008) Transitions into underage and problem drinking: developmental processes and mechanisms between 10 and 15 years of age. *Pediatrics, 121,* S273 -S289.

Wood, H, Sasakia, S., Bradley, S.J., Singha, D., Fantusa, S. Owen-Anderson, A., Di Giacomoa, A., Bain, J., & Zucker, K.J. (2013). Patterns of Referral to a Gender Identity Service for Children and Adolescents (1976–2011): Age, Sex Ratio, and Sexual Orientation. *Journal of Sex and Marital Therapy*, *1*, 1-6.

Zucker, K. J. (1992). "It ain't the meat, it's the motion": commentary on Rekers' and Morey's (1989) "Sex-typed body movements as a function of severity of gender disturbance in boys". *Journal of Psychology and Human Sexuality*, *5*, 69-73.

Zucker, K. J., & Bradley, S. J. (1995). *Gender identity disorder and psychosexual problems in children and adolescents* (p. 1-440) New York: Guilford Press.

Zuger, B. (1984). Early effeminate behavior in boys: Outcome and significance for homosexuality. *Journal of Nervous and Mental Diseases, 172*, 90-97.

Table 1. Acoustic measures of the stimuli used in the single-word perception task.

Sound(s)	Measure	GID		TGD	
		M	SD	M	SD
Vowels	F1 (bark)	6.26	1.68	6.20	1.58
	F2 (bark)	12.93	1.74	12.68	1.95
	duration (ms)	182	78	190	65
	f0 (ERB)	6.14	0.72	6.31	0.80
/s/	duration (ms)	161	34	169	43
	m1 (Hz)	5949	1502	5966	1213
	m2 (Hz)	2283	663	1912	443
	M3	-0.87	0.95	-0.88	0.85

*Table 2.*Acoustic measures of the stimuli used in the sentence perception task.

Measure	GID		TGD	
	M	SD	M	SD
Duration (ms)	1749	362	1780	383
Mean Keyword duration (ms)	295	74	306	73
Median f0 (ERB)	6.03	0.75	5.90	0.76
f0 IQR (ERB)	0.78	0.82	0.72	0.79

Table 3. The results of the linear mixed-effects model predicting acoustic measures of the word productions from age in months and group (dummy coded using contrast coding), with talkers as random effects.

Dependent Measure	Measure	Estimate	SE	*t*-value	*p*-value
F1 Frequency (Bark)	(Intercept)	6.75006	0.43078	15.669	<0.001
	Age	-0.06143	0.01892	-3.247	0.001
	Group[1]	0.03009	0.03756	0.801	0.423
F2 Frequency (Bark)	(Intercept)	13.81726	0.48302	28.606	<0.001
	Age	-0.11934	0.02192	-5.443	<0.001
	Group	0.11977	0.04353	2.752	0.006
F0 (ERB)	(Intercept)	6.65764	0.16233	41.013	<0.001
	Age	-0.05113	0.01854	-2.758	0.006
	Group	-0.09063	0.03680	-2.463	0.014
/s/ m2	(Intercept)	2786.41	313.08	8.900	<0.001
	Age	-81.41	36.01	-2.261	0.028
	Group	182.81	71.48	2.557	0.013

[1]Contrast coding: 1=GID, -1=TGD

Table 4. The results of the linear mixed-effects models predicting ratings from group (dummy-coded using contrast coding) and age in months (standardized).

Dependent Measure	Independent Variable	Estimate	SE	*t*-value	*p*-value
Ratings of Single Words	(Intercept)	1.27890	0.02003	63.84	<0.001
	Group[1]	-0.06642	0.01159	-5.73	<0.001
	Age	0.10639	0.01149	9.26	<0.001
	Group x Age	0.01605	0.01148	1.40	0.16
Ratings of Sentences	(Intercept)	1.26112	0.01863	67.71	<0.001
	Group	-0.13140	0.01281	-10.26	<0.001
	Age	0.07430	0.01275	5.83	<0.001
	Group x Age	0.00254	0.01275	0.20	0.84

[1]Contrast coding: 1=GID, -1=TGD

Table 5. The results of the linear mixed-effects model predicting perception responses in the single-word perception experiment from the acoustic characteristics of the stimuli.

Sound(s)	Measure	Estimate	SE	*t*-value	*p*-value
All	(Intercept)	1.299619	0.046321	28.057	<0.001
Words	Vowel Duration	-0.028493	0.006770	-4.209	<0.001
	F1 Frequency (Bark)	-0.051698	0.011998	-4.309	<0.001
	F2 Frequency (Bark)	0.160325	0.011583	-13.841	<0.001
	F0 (ERB)	-0.035614	0.005791	-6.150	<0.001
Words	(Intercept)	1.18188	0.03367	35.1	<0.001
with /s/	Vowel Duration	-0.04377	0.02788	-1.57	0.12
	F1 Frequency (Bark)	0.11365	0.02196	5.18	<0.001
	F2 Frequency (Bark)	-0.0764	0.02338	-3.27	0.002
	F0 (ERB)	-0.03319	0.01509	-2.2	0.03
	/s/ Duration	0.03061	0.0197	1.55	0.12
	m1	-0.12411	0.02578	-4.81	<0.001
	m2	-0.13579	0.01796	-7.56	<0.001
	m3	0.04893	0.02374	2.06	0.04

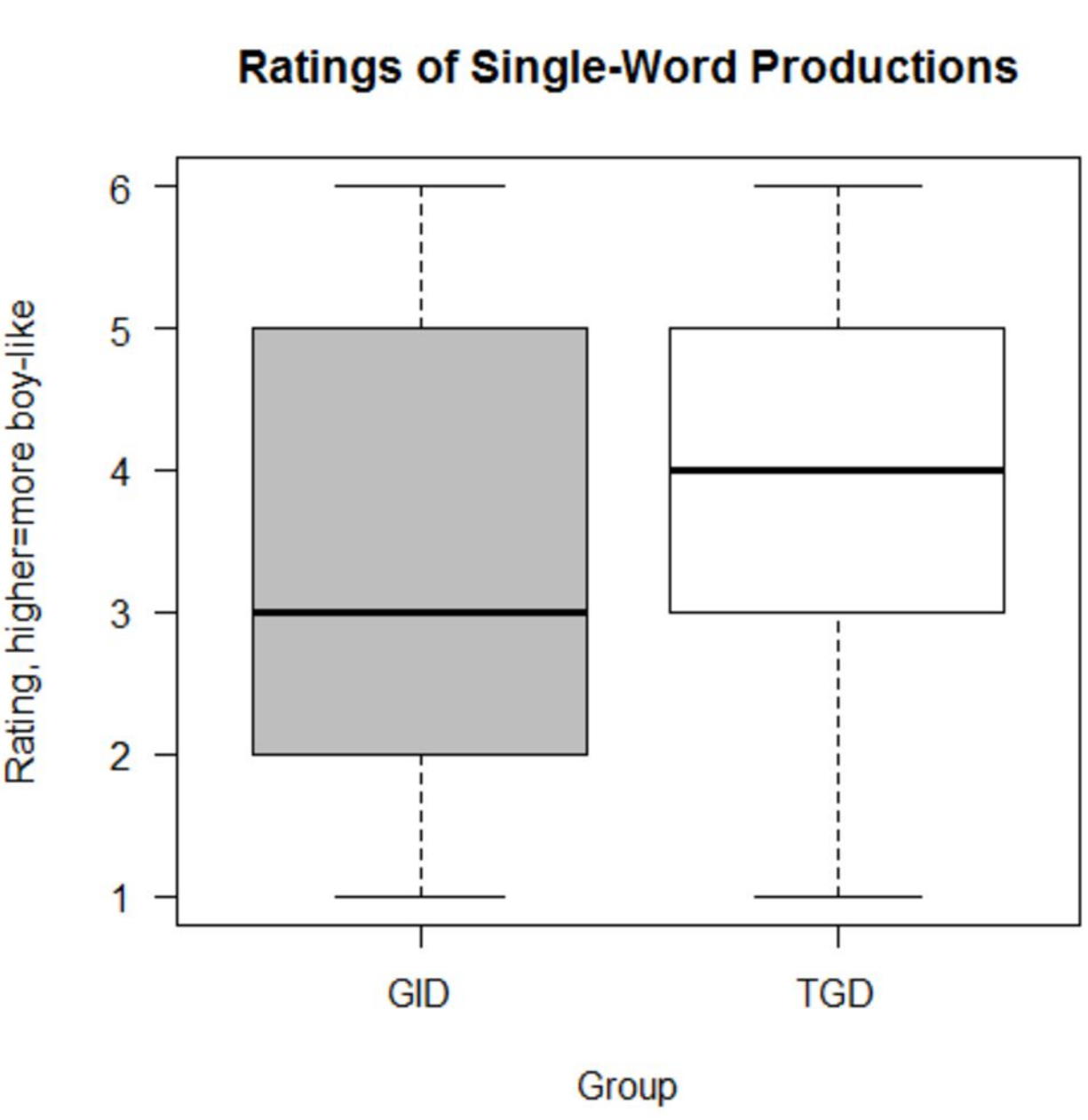
*Table 6.*The results of the linear mixed-effects model predicting perception responses in the sentence perception experiment from the acoustic characteristics of the stimuli.

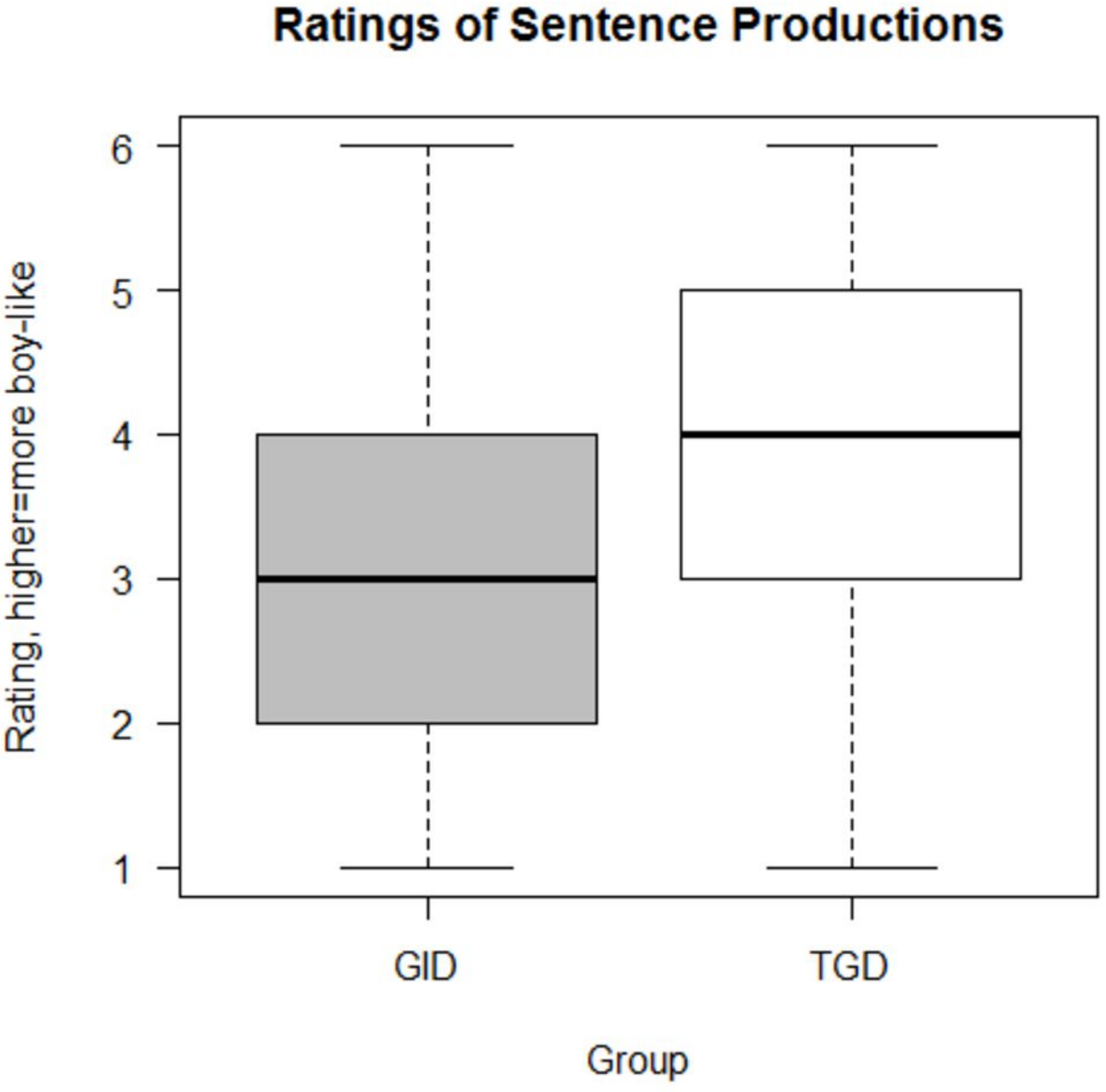
Measure	Estimate	SE	*t*-value	*p*-value
(Intercept)	-0.03768	0.016939	61.54	<0.001
stimulus length in syllables	0.037239	0.009416	-2.22	0.03
stimulus duration in ms	-0.04388	0.005907	3.95	<0.001
median f0	0.015012	0.006212	-7.43	<0.001
f0 interquartile range	-0.03768	0.016939	2.42	0.02

Figure Captions

Figure 1. Gender typicality ratings in the single-word perception experiment for boys with GID and for TGD controls.

Figure 2. Gender typicality ratings in the sentence perception experiment for boys with GID and for TGD controls.


Ratings of Single-Word Productions
Rating, higher=more boy-like
6
5
4
3
2
1
GID
TGD
Group


Ratings of Sentence Productions
Rating, higher=more boy-like
6
5
4
3
2
1
GID
TGD
Group