I use *um* when I'm thinking and *uh* when I'm speechless: The perceived social meanings of *um* and *uh*

Timothy Gadanidis

February 6, 2020

1 Introduction

In media commentary, *um* and *uh* are widely considered to be undesirable, unprofessional, and distracting. A variety of Internet articles counsel readers on purging them from their vocabularies. Writing in Forbes, for example, Rezvani (2014) refers to them as "credibility diminishers" and advises her audience of professionals to "steer [their] speech habits away from "Um" and "Uh" to more surefooted language." Along the same lines, Dlugan (2011), a public speaking blogger, writes that "filler words—including *um* and *uh*—are never written into a speech, and *add nothing when a speaker utters them*" (emphasis added), arguing that they "represent verbal static that has to be filtered out by your audience" and "may be perceived as indicating lack of preparation, lack of knowledge, or lack of passion." In other words, *uh* and *um* are meaningless "empty calories" (McKay & McKay, 2012), which should be eradicated by any means necessary.

Um and *uh* fare a little better in the psycholinguistics literature, where they are ascribed several different roles. For example, Maclay and Osgood (1959) claims that they are floor-taking, -holding and -yielding devices; Levelt (1983) describes them as symptoms of processing problems (e.g., difficulty retrieving a word or planning a sentence); and Clark and Fox Tree (2002) show that they tend to precede pauses, thus analyzing them as signals that pauses are incoming. Note that these analyses treat *um* and *uh* as processing or conversation-managerial phenomena—they are indications that a pause is coming, but they do not convey any other meaning in and of themselves.

While these views are widespread, some authors have argued that there is more to *um* and *uh* than processing difficulties and floor-holding. Tottie (2016), for example, in a corpus-linguistic analysis of the

Santa Barbara Corpus of American English, analyzes them as *planners*, pragmatic markers which speakers use to indicate that they are planning the rest of their utterance. In other words, they don't just indicate a pause, but provide information about the speaker's relationship with the ongoing utterance. Tottie (2017) extends this notion with a corpus-linguistic analysis of *uh* and *um* in journalistic writing, where she shows that the words are used in two major functions: sentence-initially, to indicate the writer's stance on the upcoming sentence; and sentence-medially, to highlight the writer's choice of words:

(1) Tottie (2017):

a. *Um*, what does your wife think about that? (*Redbook* 2010)

b. Obama is more, *um*, seasoned. Barack Obama's ... closely shorn hair appears to be increasingly gray. (*New York Times* 2008)

Tottie (2017) argues that these uses are derived from spoken *uh*s and *um*s: the initial uses parallel the common use of *uh* and *um* before answers to questions in speech, and the medial uses parallel their salient use before a word-search pause. In Tottie's (2017: 21) words, "speakers hesitate to provide answers to questions because they are uncertain about what to say or how to say it, [while] writers merely pretend to hesitate, out of reluctance to say something tactless or hurtful."

Over the last several years, a number of studies have also identified an ongoing change such that *um* is gaining in frequency and the once-dominant *uh* is declining. In synchronic and diachronic of the variable in several corpora English and five other Germanic languages, Wieling et al. (2016) identify a consistent pattern of *um* increasing and *uh* declining across all six languages. The authors found that women led the change, and that where education information was available, more educated speakers led the change as well. Similar work looking specifically at English corpora, such as the British National Corpus (Tottie, 2011) and Philadelphia Neighbourhood Corpus (Fruehwald, 2016) has found similar results.

Fruehwald (2016), Wieling et al. (2016), and (Denis & Gadanidis, 2018) have suggested that this remarkably consistent change may be linked to the potential emergence of a new discourse function for *um*. In Gadanidis (2018), I attempted to identify what this function may be using a study of instant messaging (IM) data produced by Toronto-area young people between 2004–2006 and 2014–2017, the idea being to

filter out "unconscious" uses of *um* and *uh* by examining a written medium. Using two IM corpora, the Tagliamonte Internet Archive (Tagliamonte, 2016; Tagliamonte & Denis, 2008) and a corpus that I built from my own social network, I conducted a variationist study of *um* and *uh* and found the two words were specializing (Kroch, 1994) for message position (*um* being more likely to appear in message-initial position). Qualitative differences in the contexts in which the two words were used were also observed, *um* typically being used in requests (2a) and *uh* disagreements (2b).

(2) a. uhm this is a bit random but if we were to have takoyaki party, would you mind having it at your place?

b. uh hello, you've been trying to change her mind, trick her into liking you back again

This specialization can also be viewed in terms of pragmaticalization (Davis & Gutzmann, 2015). In the first stage of development, *um* and *uh* indicate pauses or planning, which can give rise to a conversational implicature that the speaker is hesitating to say what they are about to say. In the second stage of development, this conversational implicature conventionalizes, indicating hesitancy regardless of whether or not a pause is actually incoming. It is this conventional implicature that allows *um* and *uh* to be used as markers of hesitation, and from there further specialize, as seen in the IM data.

However, the precise nature of these specializations remains somewhat murky. While *um* appears to be used more often in positively-polite and mitigative contexts, and *uh* appears to be used more often in impolite and challenging/disaligning contexts, it's far from clear whether these words are in direct opposition to each other (e.g., *um* is the polite version of *uh* or vice versa) or have different meanings entirely. It's also unclear whether these specialized expressive functions have given rise to social meanings, i.e., language ideologies about the speakers who use them. The issue is further complicated by the apparent change in progress identified by Wieling et al. (2016, among others). Finally, it is not clear whether the functions and meanings of *um* and *uh* in IM are the same as (or similar to) *um* and *uh* in speech.

This study reports the results of two experiments designed to address these gaps in our understanding of *uh* and *um*. This was achieved by making use of a matched-guise design (along the lines of Campbell-Kibler, 2010; Maddeaux & Dinkin, 2017; *inter alia*) which required participants to rate one participant in an IM or

spoken conversation on various scales measuring perceived identity and personality characteristics. In each conversation, the rated IM-er or speaker produced one token of either *um*, *uh*, or neither, depending on the experimental condition. Comparing results from each of these conditions allows for the investigation of the overall research hypothesis that speakers evaluate messages differently based on the presence of *uh* and *um*: since everything else is held constant across conditions, if we see differences between how a person is rated using *um* and how they are rated when not using *um*, those differences can likely be attributed to how the rater perceives *um*.

In Experiment 1, participants were asked to read instant messaging conversations where one speaker could variably use either *um*, *uh*, or neither. They were then asked for quantitative and qualitative feedback about the messages that they had read. Experiment 2 was largely the same as Experiment 1, with the major exception of stimulus medium: the stimuli were auditory rather than textual. Two speakers of Canadian English, one man and one woman, recorded dialogues with the same text as the stimuli of Experiment 1, with necessary changes made to the scripts to make them sound more natural as spoken conversations.

The experiment was designed to test the following hypotheses about how *um* and *uh* are perceived in IM and in speech:

H_1: *um* and *uh* are perceived as hesitant;

H_2: *um* and *uh* are perceived as unintelligent;

H_3: *um* is perceived as feminine and *uh* is perceived as masculine;

H_4: *um* is perceived as polite and *uh* is perceived as impolite.

H_1 and H_2 are derived from salient social commentary (summarized above) about what *um* and *uh* mean, as well as (for H_1) the common linguistic understanding of their function. H_3 is based on the ongoing change in progress, where *um* is more frequently used by women compared to *uh*. H_4 is based on the qualitative findings from Gadanidis (2018) summarized above.

The following section describes the experimental methodology in more detail.

2 Methods

2.1 Participants

Experiment 1 78 L1 English speakers were recruited to participate in the experiment. Participants had the option of receiving either course credit or $5 as compensation.

62 participants reported their gender as female, 15 as male, and 1 as genderqueer. The youngest participant was 18 and the oldest 51, with a median of 21, a mean of 22.5 and a standard deviation of 5.87. Participants were asked to self-report their ethnicity, and reported a wide range of ethnicities, including White, Black, Caribbean, Korean, Chinese, Japanese, Vietnamese, South Asian, and Arab, reflecting the diversity of the population from which they were drawn.

Experiment 2 As with Experiment 1, participants were recruited from the University of Toronto community, and they had the option of receiving either course credit or CA$5 as compensation. As with Experiment 1, reported a wide range of ethnicities, including those mentioned in the previous section, reflecting the diversity of the population from which they were drawn.

To accommodate the new factor, speaker voice, the number of participants was doubled, from 78 to 156 (78 per group). When asked for their gender identity, 88 participants reported that they were female or women, 2 reported that they were non-binary, and 36 reported that they were male or men. The youngest participant reported their age as 16 and the oldest 54, with a median of 20, a mean of 21.43 and a standard deviation of 5.27.

For reference, the age and gender breakdowns by voice group are given in Tables 1 and 2.

voice heard	man/male	woman/female	non-binary
Penguin	22	54	2
Raven	23	54	1
total	45	108	1

Table 1: Gender breakdown by voice

voice heard	mean	median	maximum	minimum	sd
Penguin	20.7	20	48	16	3.76
Raven	22.2	20	54	17	6.39
overall	21.61	20	16	54	5.27

Table 2: Age breakdown by voice

2.2 Materials

Experiment 1 Participants each viewed a total of 16 stimuli, which were inspired by and/or directly modified from messages in my instant messaging corpus (Gadanidis, 2018). Of these, six were critical stimuli, which could either contain *uh*, *um*, or neither. The other ten stimuli were fillers. The stimuli can be viewed in Appendix C.1, or downloaded from the link in Appendix B.

Each participant saw two critical trials containing *uh*, two containing *um*, and two containing *neither*. This was accomplished by assigning each participant to one of three conditions. In condition 1, for example, stimuli 1 and 4 contained *um*, stimuli 2 and 5 contained *uh*, and stimuli 3 and 6 contained neither. The parameters for each condition are given in Table 3.

Condition	*uh* stimuli	*um* stimuli	control stimuli
1	1, 4	2, 5	3, 6
2	2, 5	3, 6	1, 4
3	3, 6	1, 4	2, 5

Table 3: Conditional parameters for critical stimuli

Some filler items were also varied by condition: stimuli 7–10 contained either *lmao* or *lol*, and stimuli 13–16 contained either *eh* or *right*. Stimuli 11 and 12 were invariant. These manipulations are not relevant for the current study and will not be analyzed; they were included only as a potential distractor from the true purpose of the experiment.

Experiment 2 The stimuli for Experiment 2 were spoken conversations using, wherever possible, identical wording to the Experiment 1 IM conversations. In the process of recording, some minor modifications were made to make the conversations sound more natural as spoken conversations, such as the removal of

abbreviations like *rn* 'right now'. Transcripts of the stimuli can be viewed in Appendix C.2, and the audio files can be downloaded from the link in Appendix B.

In an attempt to counterbalance for the potential effect of the speakers' perceived gender, and characteristics of their voice such as vocal quality and pitch, all the stimuli were recorded twice, producing versions with both the man (hereafter "Penguin") and woman (hereafter "Raven") as the rated speaker. Penguin was born in 1989, in Manila, Philippines, and moved to British Columbia in 2001, then Toronto in 2016. Raven was born in 1982, in Liverpool, England, and moved to Southern Ontario in 1984 and Toronto in 2005. The condition was implemented between subjects: each participant always rated either Penguin or Raven. Participants were not informed that the individual they rated was always the same, and were instructed to rate each conversation in isolation. To avoid biasing participants' gender responses, or implying that the individuals across stimuli were the same, participants were instructed to rate "the person who spoke first" or "second" (the order varied across stimuli).

In Experiment 1, some filler stimuli were variable: some stimuli could contain either *lol* or *lmao*, and some could contain either *eh* or *right*. Because *lol* and *lmao* are not typically used in spoken English, they were not included, and the stimuli in question no longer vary. The stimuli containing *eh*/*right* were modified to invariably contain *eh*, for two reasons: (a) the *eh*/*right* variation was intended to serve as a distractor; however, participants in Experiment 1 only commented upon *eh*, not *right*, and none noticed the variation; and (b) making the fillers invariant reduced the amount of recording and splicing necessary when preparing the stimuli.

2.3 Procedure

Experiment 1 The experiment was implemented using jsPsych version 6.0.5 (de Leeuw, 2015).

After reading and signing the informed consent form, participants were seated in front of a computer screen in a quiet room. Participants were first asked to self-report their age, gender, and ethnicity. These were all text boxes, to allow participants to enter whatever they wanted, rather than selecting from a drop-down menu.

After entering their information, participants were given instructions on the screen, as follows:

In this study, you will see some screenshots.

These screenshots are taken from the middle of a conversation which is still ongoing: what you see is not the full conversation, only a part of it.

Your task will be to decide what you think about one of the participants in the conversation. You'll first be asked to rate them on different scales, and then on the next page you will be able to type in any other comments you have. Try to answer the questions at a quick pace, using your gut feeling, without trying to think too hard about them for too long.

When you are ready, press any key to begin the experiment.

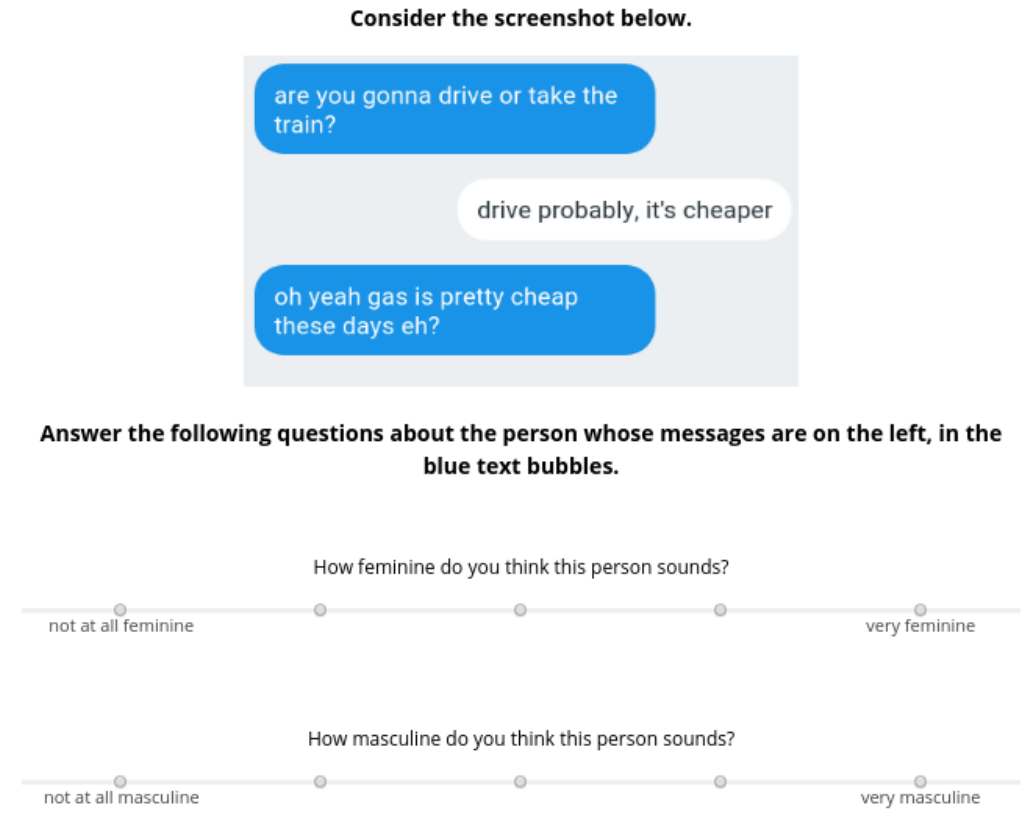

Figure 1: Truncated response display for stimulus 15, *eh* version

After pressing a key, participants proceeded to the main task. Stimuli were presented in a pseudorandom

order, in the display shown in Figure 1. For each stimulus, participants were asked to rate the person whose text bubbles had a blue background on a total of ten five-point Likert scales:

1. "not at all feminine" to "very feminine"
2. "not at all masculine" to "very masculine"
3. "not at all young" to "very young"
4. "not at all queer" to "very queer"
5. "not at all Canadian" to "very Canadian"
6. "not at all intelligent" to "very intelligent"
7. "not at all hesitant" to "very hesitant"
8. "not at all polite" to "very polite"
9. "not at all casual" to "very casual"
10. "not at all friendly" to "very friendly"

After each stimulus, participants were asked (on a separate page) for optional qualitative feedback, with the prompt "Is there anything else you want to say about the screenshot you just saw? You can enter as much or as little text as you like." Participants could either simply click "Continue" to continue without providing a qualitative response, or enter text and then click "Continue".

After all stimuli were viewed and responded to, participants were asked to answer the following pre-debrief questions:

1. What do you think the experiment was about? (required)
2. Did you notice anything interesting about the way language was used in the messages you read? If so, what? (optional)
3. Do you have any general comments about the experiment and your experience doing the experiment? (optional)

Then, participants were debriefed in person by the experimenter. Finally, having learned the purpose of the experiment, they were asked to answer the following two questions, which were designed to determine the extent to which participants had noticed *uh* and *um*:

1. Were you surprised to learn that the experiment was about uh and um? Why or why not? (required)
2. Do you have any final comments about your experience participating in this experiment? (optional)

After these questions were answered, the experiment ended and the data was saved.

Experiment 2 The experimental procedure was the same as for Experiment 1, except that participants were asked two additional questions in the after-debrief questionnaire, #2 and #3 in the following list:

1. Were you surprised to learn that the experiment was about uh and um? Why or why not? (required)
2. What do you think it means when someone uses *uh* or *um*? (required)
3. Do you think *um* and *uh* have different meanings? If so, how are they different? (required)
4. Do you have any final comments about your experience participating in this experiment? (optional)

These additional questions were designed to interrogate participants' ideologies about *um* and *uh* directly, as a qualitative supplement for the quantitative results.

2.4 Pre-registered hypotheses

As outlined in the introduction, of the Likert scales that participants were asked to rate speakers on, five were pre-registered as potentially relevant: hesitation, masculinity, femininity, politeness, and intelligence.

Hesitation was chosen because it has been identified as one of the core meanings of both *uh* and *um* in past work (Gadanidis, 2018; Tottie, 2016, 2017). Intelligence was chosen because it is often implicated in media commentary about how *um* and *uh* make speakers sound. Masculinity and femininity were chosen because the ongoing change in progress is favouring *um* is led by women (Wieling et al., 2016). Politeness was chosen based on my qualitative analysis of *uh* and *um* in IM (Gadanidis, 2018): I identified *um* as more polite than *uh*, based on the contexts in which each variant was used.

Quantitative results for the other scales were analyzed on an exploratory basis. I avoid making firm conclusions about them, given that there was no basis in the literature for their inclusion and they were mainly intended as distractors from the true purpose of the experiment (which seems to have worked as intended, based on participants' qualitative feedback).

2.5 Analysis

2.5.1 Statistical methods

The data were analyzed using R (R Core Team, 2018), a programming language and environment for statistical computing. In particular, the `factanal()` function from the `stats` base package was used for factor analysis. Several packages were also used to extend R's functionality. The `tidyverse` package (Wickham, 2017) was used for data processing and manipulation. The `ggplot2` package (Wickham, Chang, et al., 2008), included in the `tidyverse` package, was used for data visualization. The `ordinal` package (Christensen, 2019) was used for ordinal regression models (using the function `clmm()`), and the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) was used for linear regression models (using the function `lmer()`), along with `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2017), which extends `lme4` to allow signficance testing and p-value computation with `lmer()`. For more details, see Appendix D for the output of R's `sessionInfo()` function, which prints names and version numbers of all loaded packages, as well as R itself.

Following the American Statistical Association's recent recommendations on p values and significance (Wasserstein, Schirm, & Lazar, 2019), for all statistical tests:

- I report effect sizes for fixed effects (β) along with p values
- I report the 95% confidence interval for all estimates, computed using R's `confint()` function
- I report p values as continuous quantities
- I do not dichotomize p values into "significant" and "not significant" categories using bright-line rules such as $p < 0.05$
- When determining the importance of an effect, I take into account effect size, confidence interval, p, and my own domain knowledge, rather than relying solely on the p value

2.5.2 Analytical framework

In interpreting the quantitative and qualitative results, I draw on the concepts of indexical order (Silverstein, 2003) and the indexical field (Eckert, 2008). In this understanding of indexicality, the meanings of a feature or variant are not set or discrete, but are fluid and always open to reinterpretation: once an indexical value is established (nth order), it is available for reinterpretation and reconstrual (n + 1th order); the resulting values are then available for further reinterpretation, and so on (Eckert, 2008: 463). Eckert (2008: 464) conceptualizes these meanings as organized into an indexical field: "a constellation of meanings that are ideologically linked." These abstractions are particularly useful for understanding the meanings of *um* and *uh* because they allow us to capture the observation that their apparently primary or core meaning, hesitation, is fundamentally linked to other potential meanings, such as politeness and face-protection (Gadanidis, 2018)—depending on the context, hesitation can deployed to suggest that the speaker is attempting to be considerate of the other speaker's feelings, or, alternately, to indicate shock, disapproval, or disgust at something the other speaker has said.

Under this framework, although the scales are presented to participants as discrete, (first, rate femininity; then, rate masculinity, and so on) the analysis must consider the ways in which they are connected and interrelated. For example, as we will see shortly, the concepts of "politeness" and "femininity" are inextricably linked due to dominant ideological expectations for women's behaviour in European settler-colonial states like Canada (cf. Lakoff, 1973; Ochs, 1992).

3 Ordinal regression

To determine the extent to which the experimental condition (i.e., whether *um*, *uh* or *neither* was present) affected the results for each Likert scale, I fit ordinal regression models using `clmm()` from the `ordinal` package (Christensen, 2019).

Note that rather than a single intercept, as may be familiar with `lme4` (Bates et al., 2015) models, the ordinal regression models produced by `clmm()` have four threshold coefficients, one for each threshold between values. As I understand it, these values are odds representing the ratio $a : b$, where a is the probability of a response value being below the threshold and b is the probability of a response value being above the

threshold, under the *neither* condition (the reference level) (see moremo, 2018 for helpful discussion). Unlike `glmer()` and `lmer()`, `clmm()` does not compute p-values for these threshold coefficients; accordingly, only z-values are given in the tables. The threshold coefficients are reported in the tables as $x|y$, where x and y are the two values on either side of the threshold. Also unlike `lme4` linear and logistic regression models, in ordinal regression models, the coefficients for fixed effects are *subtracted from*, not added to, the threshold coefficients when interpreting the model. So, for example, in Table 10 on page 18, the −3.09 estimate for the 1|2 threshold coefficient indicates that a response of '1' is predicted to be much less likely than any higher response ('2'–'5'). The estimate for *um*, 0.71, is subtracted from −3.09 to yield an estimate of −3.80, indicating that compared to the *neither* condition, '1' responses are even less likely when the rated message contains *um*.

For each scale, models from Experiment 1 and Experiment 2 are presented in turn (hereafter referred to as E1 and E2 models). For each E1 model, there is one fixed effect: *variant*. There is also a random intercept for *subject*, a random intercept for *stimulus*, and a random by-subject slope for *variant*. No by-subject slope for variant was used for the *queer* model due to lack of convergence when the slope was included.

For each E2 model, there are three fixed effects: *variant*, *voice*, and the interaction between *variant* and *voice*. There are also two random intercepts, one for *subject* and one for *stimulus*, as well as random by-subject slopes for *variant* and *voice*. No by-subject slope for voice was used for the *young* model due to lack of convergence when the slope was included.

For all models, treatment contrast coding was used for the *variant* predictor. This compared each of *uh* and *um* to the reference level, *neither*. The contrast coding matrix is shown in Table 4. For E2 models, the

contrast	neither	*um*	*uh*
uh	0	1	0
um	0	0	1

Table 4: Treatment contrast coding matrix for *variant*.

voice predictor was simple-coded, comparing the *raven* voice to the reference level, *penguin*. The contrast coding matrix is shown in Table 5. Because *variant* was treatment-coded and *voice* was simple-coded, the threshold coefficients are based on each model's predictions at the *neither* level of *variant*, and at the mean

contrast	penguin	raven
raven	-0.5	0.5

Table 5: Simple contrast coding matrix for *voice*.

of both levels of *voice*.

Each set of models is accompanied by a figure showing the raw proportion of responses under each condition: "IM" from Experiment 1, and "Penguin" and "Raven" from Experiment 2. Each figure contains a dotted line at the 50% mark, indicating the median for each bar. In the text accompanying each figure, I describe the patterns that I have identified. Some of the criteria I use to identify patterns include: presence of robust effects in the corresponding model (the main criterion); number of 1–2 or 4–5 ratings (possibly indicating an effect favouring positive or negative responses); number of extreme (1 or 5) ratings (possibly indicating polarizing effects); and number of neutral (3) ratings (possibly indicating less certainty or ambivalence).

3.1 Planned analyses

3.1.1 Hesitant

Figure 2 shows that for Experiment 1, the *um* and *uh* conditions are both rated as much more hesitant than the *neither* condition, (far more 4 or 5 responses, far fewer 1 or 2 responses) with *uh* slightly more hesitant than *um*. The model (Table 6) confirms that both *uh* ($\beta = 1.20, p \approx 0$) and *um* ($\beta = 0.99, p \approx 0$) are predicted to elicit higher *hesitant* ratings than *neither*.

For Experiment 2, the figure shows a similar pattern, with *um* and *uh* being rated as more hesitant than *neither* for both Penguin and Raven. For Penguin, there is also a similar pattern as the Experiment 1 data where *uh* is rated slightly more hesitant than *um*, but this pattern is absent for Raven. The model is shown in Table 7. Both *um* ($\beta = 1.35, p \approx 0$) and *uh* ($\beta = 1.34, p \approx 0$) are predicted to elicit higher *hesitant* ratings. There is also an interaction between voice and variant: *uh* is predicted to be rated as less hesitant for Raven than for Penguin ($\beta = -0.83, p = 0.01$); there is also a similar but weaker effect for *um* ($\beta = -0.56, p = 0.09$).

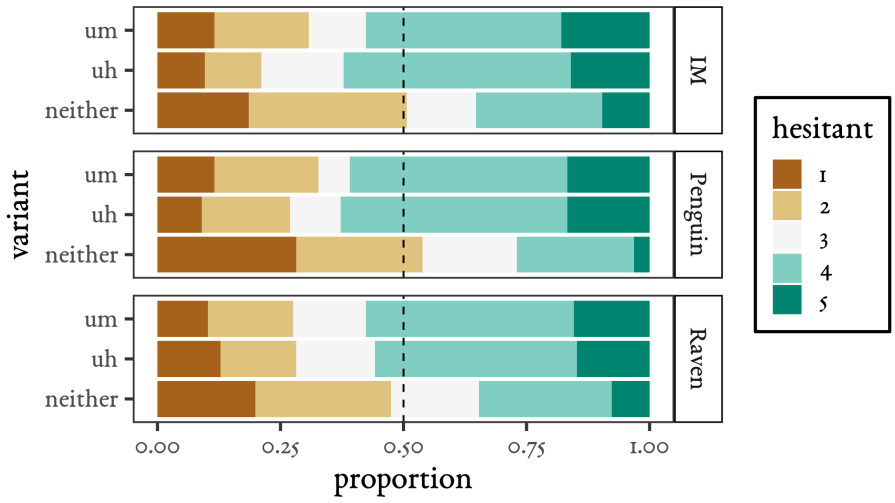

Figure 2: Proportion of *hesitant* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-1.72	-2.51	-0.94	0.40	-4.31	0.00
2|3	-0.19	-0.95	0.56	0.39	-0.50	0.62
3|4	0.61	-0.15	1.37	0.39	1.57	0.12
4|5	3.03	2.20	3.86	0.42	7.16	0.00
variant = uh	1.20	0.75	1.66	0.23	5.22	0.00
variant = um	0.99	0.53	1.45	0.23	4.24	0.00

Table 6: E1 model for the 'hesitant' scale.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-1.43	-2.10	-0.76	0.34	-4.19	0.00
2|3	0.07	-0.58	0.73	0.34	0.22	0.83
3|4	0.88	0.22	1.54	0.34	2.60	0.01
4|5	3.51	2.80	4.23	0.36	9.62	0.00
variant = uh	1.35	1.00	1.70	0.18	7.59	0.00
variant = um	1.34	1.00	1.68	0.17	7.74	0.00
voice = raven	0.55	0.04	1.06	0.26	2.10	0.04
uh x raven	-0.83	-1.50	-0.16	0.34	-2.42	0.02
um x raven	-0.56	-1.21	0.08	0.33	-1.71	0.09

Table 7: E2 model for the 'hesitant' scale.

3.1.2 Intelligent

Figure 3 shows that for Experiment 1, *um* and *uh* received more 1–2 ratings and fewer 4–5 ratings than *neither*. This is also shown in the model (Table 8), where both *uh* ($\beta = -0.75, p \approx 0$) and *um* ($\beta = -0.61, p = 0.01$) are both predicted to elicit lower *intelligence* ratings than *neither*.

The figure for Experiment 2, in contrast, shows less clear by-variant differences. The ratings for Penguin are virtually identical across all three variants, and while Raven receives more 4 ratings with *um* than with *uh* and *neither*, she also receives less 5 ratings and more 1–2 ratings. However, it is clear that Raven is rated as less intelligent than Penguin; this is borne out in the model in Table 9 ($\beta = -1.05, p \approx 0$).

It should also be noted that more participants were much less decisive about intelligence than, for example, hesitancy, where the median rating for both *um* and *uh* across all conditions was 4, and the *neither* condition's median was 2 (or 3 for Raven). Here, the median intelligence rating for all conditions is 3 by a long shot.

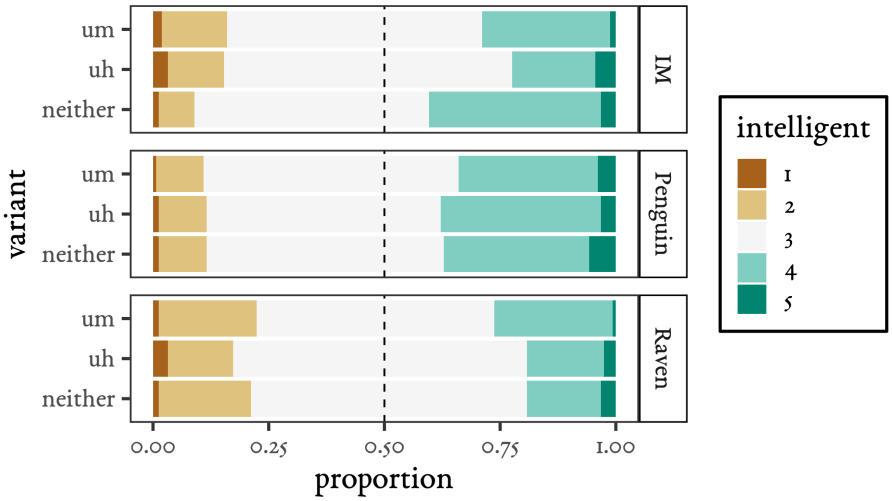

Figure 3: Proportion of *intelligent* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-4.80	-5.63	-3.98	0.42	-11.41	0.00
2|3	-2.72	-3.29	-2.14	0.29	-9.27	0.00
3|4	0.55	0.05	1.06	0.26	2.14	0.03
4|5	3.69	2.94	4.44	0.38	9.65	0.00
variant = uh	-0.75	-1.20	-0.30	0.23	-3.27	0.00
variant = um	-0.61	-1.06	-0.15	0.23	-2.63	0.01

Table 8: E1 model for the 'intelligent' scale.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-5.00	-5.72	-4.27	0.37	-13.53	0.00
2|3	-2.18	-2.67	-1.69	0.25	-8.70	0.00
3|4	1.17	0.70	1.64	0.24	4.87	0.00
4|5	4.40	3.76	5.04	0.33	13.42	0.00
variant = uh	0.04	-0.28	0.36	0.16	0.23	0.82
variant = um	0.03	-0.28	0.35	0.16	0.21	0.83
voice = raven	-1.05	-1.66	-0.43	0.31	-3.34	0.00
uh x raven	0.10	-0.54	0.75	0.33	0.31	0.75
um x raven	0.27	-0.36	0.91	0.33	0.84	0.40

Table 9: E2 model for the 'intelligent' scale.

3.1.3 Feminine

Figure 4 shows the proportion of feminine responses in both experiments. In Experiment 1, *um* is rated more feminine (i.e., more 4 responses and less 2 responses) than *uh* and *neither*, which have roughly the same proportions. This is borne out in the model for Experiment 1 (Table 10), where *um* is predicted to elicit higher values on the feminine scale ($\beta = 0.72, p \approx 0$).

In Experiment 2, however, variant choice has a comparatively small effect. Which speaker was heard appears to largely determine listeners' ratings here, with Penguin being given a 1 or 2 more than 50% of the time, and Raven being given a 4 or 5 well over 80% of the time. This too is borne out in the model (Table 11), which predicts Raven to be rated much more feminine than Penguin ($\beta = 5.11, p \approx 0$). For Raven specifically, *um* actually received the *least* feminine responses.

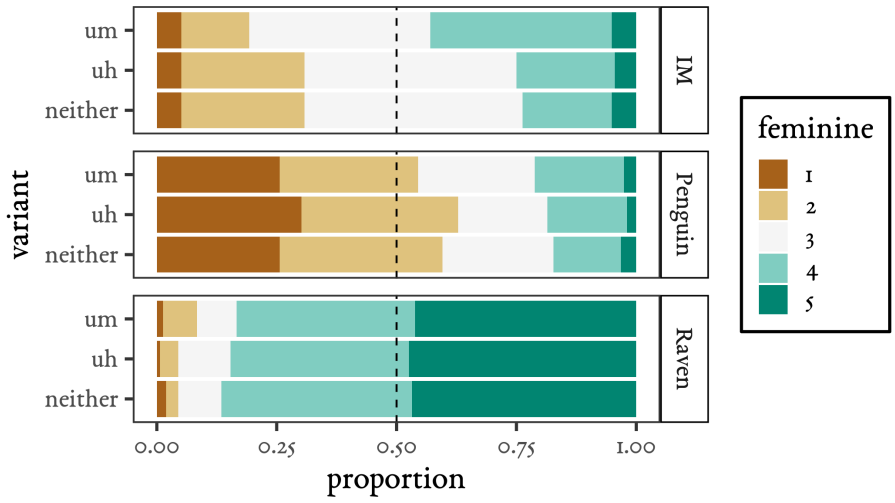

Figure 4: Proportion of *feminine* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-2.94	-3.59	-2.29	0.33	-8.81	0.00
2|3	-0.86	-1.39	-0.34	0.27	-3.21	0.00
3|4	1.20	0.67	1.74	0.27	4.42	0.00
4|5	3.49	2.81	4.18	0.35	9.96	0.00
variant = uh	0.02	-0.41	0.45	0.22	0.09	0.93
variant = um	0.72	0.27	1.17	0.23	3.16	0.00

Table 10: E1 model for the 'feminine' scale.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-4.11	-4.67	-3.56	0.28	-14.44	0.00
2|3	-1.89	-2.36	-1.41	0.24	-7.81	0.00
3|4	-0.29	-0.75	0.16	0.23	-1.26	0.21
4|5	2.59	2.09	3.08	0.25	10.27	0.00
variant = uh	-0.16	-0.53	0.20	0.19	-0.88	0.38
variant = um	0.06	-0.27	0.39	0.17	0.34	0.73
voice = raven	5.11	4.24	5.98	0.45	11.48	0.00
uh x raven	0.06	-0.68	0.80	0.38	0.17	0.87
um x raven	-0.25	-0.94	0.44	0.35	-0.71	0.48

Table 11: E2 model for the 'feminine' scale.

3.1.4 Masculine

The raw results for masculinity are presented in Figure 5. For Experiment 1, *um* received the least masculine ratings (more 2 ratings, less 4 ratings), and *uh* received the most. The effect in the model (Table 12) for *um* compared to the *neither* condition is quite robust ($\beta = -0.72, p \approx 0$). The trend for *uh* visible in the figure also appears in the model ($\beta = 0.32, p = 0.15$), though it should be noted that the 95% confidence interval overlaps 0 and p is comparatively high.

In Experiment 2, as with femininity, ratings for masculinity appear largely to be determined by speaker voice. Penguin is rated as 4 or 5 around 50% of the time, whereas Raven is rated as 1 or 2 around 85% of the time. For Penguin, there is a visible difference between variants, however, where *um* receives the lowest responses and *neither* receives the highest responses. In line with the figure, the model (Table 13) predicts Raven to have far lower ratings than Penguin for masculinity ($\beta = -5.03, p \approx 0$). There is also a main effect for *um* such that compared to the *neither* condition, the model predicts lower ratings for *um* ($\beta = -0.46, p = 0.01$).

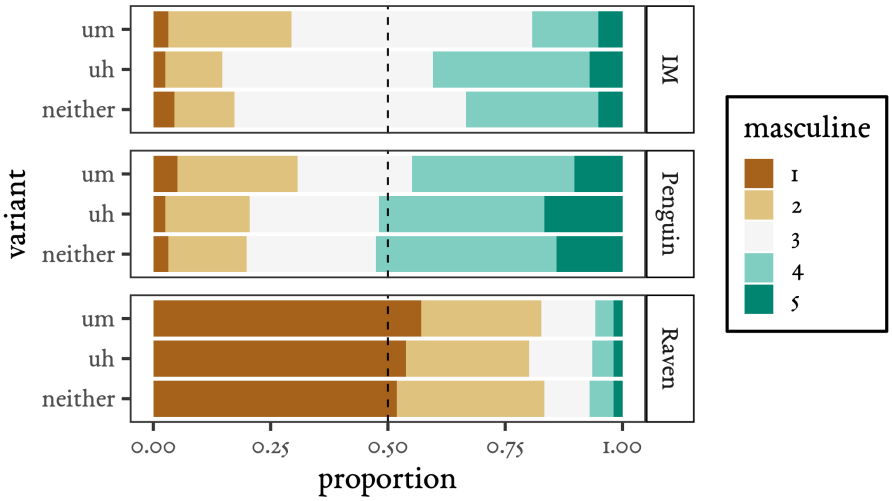

Figure 5: Proportion of *masculine* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-3.97	-4.79	-3.15	0.42	-9.48	0.00
2|3	-1.76	-2.37	-1.16	0.31	-5.70	0.00
3|4	0.83	0.26	1.40	0.29	2.84	0.00
4|5	3.12	2.41	3.84	0.36	8.57	0.00
variant = uh	0.32	-0.12	0.77	0.23	1.44	0.15
variant = um	-0.72	-1.17	-0.26	0.23	-3.07	0.00

Table 12: E1 model for the 'masculine' scale.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-2.11	-2.61	-1.60	0.26	-8.22	0.00
2|3	0.46	-0.03	0.95	0.25	1.85	0.06
3|4	2.31	1.80	2.81	0.26	8.94	0.00
4|5	5.15	4.53	5.78	0.32	16.21	0.00
variant = uh	0.04	-0.31	0.39	0.18	0.23	0.82
variant = um	-0.46	-0.82	-0.10	0.18	-2.53	0.01
voice = raven	-5.03	-5.97	-4.08	0.48	-10.42	0.00
uh x raven	-0.11	-0.83	0.61	0.37	-0.31	0.76
um x raven	0.19	-0.54	0.91	0.37	0.50	0.61

Table 13: E2 model for the 'masculine' scale.

3.1.5 Polite

Figure 6 shows that in Experiment 1, *uh* and *um* received less 4–5 responses and more 1–2 responses than *neither*. This is especially the case for *uh*. In the model (Table 14), *uh* is predicted to elicit lower *polite* ratings ($\beta = -0.57, p = 0.03$) compared to the *neither* condition. There is an effect for *um* in the same direction, but it is weaker and less robust ($\beta = -0.37, p = 0.13$).

For Experiment 2, for both Raven and Penguin, *um* actually received *more* 4–5 and less 1–2 responses than *neither*. For Penguin, the same is true for *uh*, although for Raven, *uh* received less 4–5 and more 1–2 responses than both *um* and *neither*. Raven also receives overall lower politeness ratings than Penguin does. In the model (Table 15), *um* is predicted to elicit higher polite ratings ($\beta = 0.29, p = 0.06$), although it should be noted that the 95% confidence interval slightly overlaps 0. Raven is also predicted to receive lower ratings than Penguin ($\beta = -1.22, p \approx 0$).

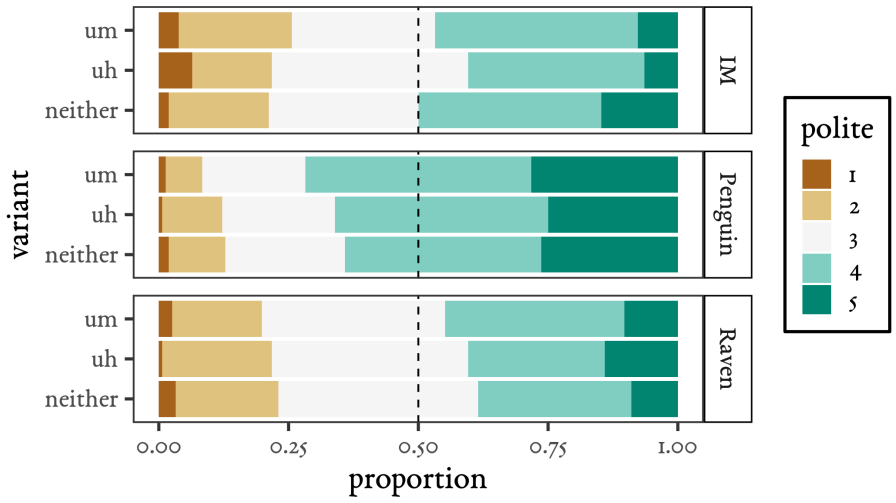

Figure 6: Proportion of *polite* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-4.12	-4.97	-3.27	0.43	-9.54	0.00
2|3	-1.89	-2.56	-1.21	0.35	-5.47	0.00
3|4	-0.11	-0.75	0.52	0.32	-0.35	0.72
4|5	2.48	1.77	3.20	0.36	6.81	0.00
variant = uh	-0.57	-1.06	-0.07	0.25	-2.24	0.03
variant = um	-0.37	-0.86	0.11	0.25	-1.51	0.13

Table 14: E1 model for the 'polite' scale.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-4.60	-5.23	-3.97	0.32	-14.24	0.00
2|3	-1.96	-2.37	-1.55	0.21	-9.37	0.00
3|4	-0.11	-0.50	0.27	0.20	-0.58	0.56
4|5	1.98	1.57	2.39	0.21	9.49	0.00
variant = uh	0.11	-0.19	0.41	0.15	0.71	0.48
variant = um	0.29	-0.01	0.59	0.15	1.88	0.06
voice = raven	-1.22	-1.73	-0.72	0.26	-4.73	0.00
uh x raven	0.17	-0.43	0.78	0.31	0.57	0.57
um x raven	-0.02	-0.62	0.57	0.30	-0.08	0.94

Table 15: E2 model for the 'polite' scale.

3.2 Exploratory analyses

3.2.1 Queer

Figure 7 shows the raw ratings for the *queer* scale across both experiments. In Experiment 1, queer ratings above 3 were extremely uncommon, at around 5–10%. If there is any difference between the variants, it is very slim; while *um* has slightly more 4–5 responses than *uh* and *neither*, and *uh* has slightly less 1–2 responses, Table 16 shows that the effect are not very robust. The estimate for *um* vs. *neither* ($\beta = 0.26$, $p = 0.34$) has an extremely wide confidence interval and high p value, and the estimate for *uh* is similar ($\beta = 0.43, 0.12$), though slightly stronger.

In contrast, for Experiment 2, the figure suggests that *um* has less 1–2 responses and more 4–5 responses than either of the other variants, for both Penguin and Raven. This is borne out in the model (Table 17), where *um* is predicted to elicit higher *queer* ratings than the *neither* condition ($\beta = 0.32$, $p = 0.07$).

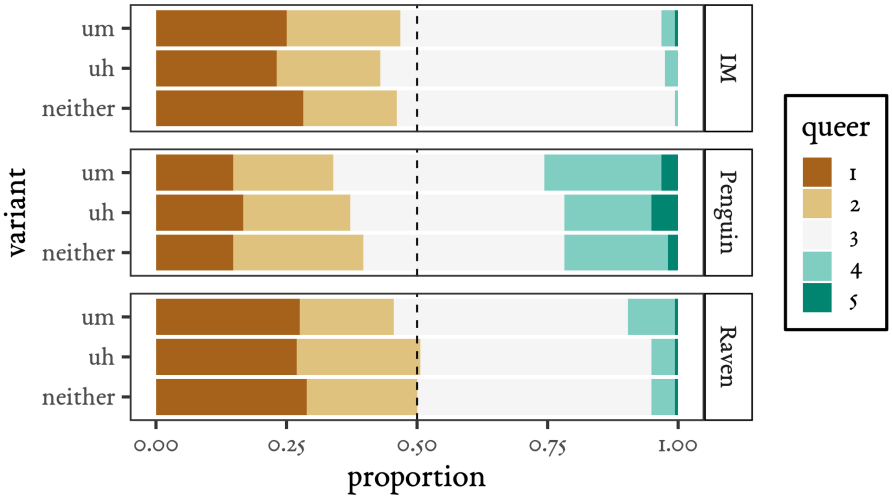

Figure 7: Proportion of *queer* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-2.03	-2.85	-1.20	0.42	-4.82	0.00
2|3	0.06	-0.74	0.85	0.40	0.14	0.89
3|4	6.40	5.26	7.53	0.58	11.06	0.00
4|5	8.74	6.56	10.92	1.11	7.85	0.00
variant = uh	0.43	-0.12	0.97	0.28	1.54	0.12
variant = um	0.26	-0.28	0.80	0.28	0.96	0.34

Table 16: E1 model for the 'queer' scale.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-2.18	-2.68	-1.68	0.25	-8.60	0.00
2|3	-0.16	-0.61	0.30	0.23	-0.67	0.50
3|4	3.20	2.67	3.73	0.27	11.88	0.00
4|5	5.98	5.21	6.76	0.40	15.15	0.00
variant = uh	0.06	-0.29	0.41	0.18	0.33	0.74
variant = um	0.32	-0.03	0.66	0.17	1.81	0.07
voice = raven	-1.28	-2.09	-0.47	0.41	-3.09	0.00
uh x raven	-0.15	-0.85	0.55	0.36	-0.42	0.68
um x raven	-0.16	-0.84	0.52	0.35	-0.46	0.64

Table 17: E2 model for the 'queer' scale.

3.2.2 Young

Figure 8 shows the raw ratings for the *young* scale across both experiments. In Experiment 1, both *um* and *uh* received more 4–5 ratings and less 1–2 ratings than the *neither* condition. These differences also appear in the model (Table 18), with both *uh* ($\beta = 0.62$, $p = 0.01$) and *um* ($\beta = 0.50$, $p = 0.02$) being predicted to elicit higher *young* ratings.

For Experiment 2, the results are less clear. Some minor differences between variants appear to be present in the figure, but the direction is unclear, with *um* being rated the youngest for Penguin, but the oldest for Raven. There are no robust main effects for *um* or *uh* compared to *neither* in the model (Table 19), nor is there a robust main effect for voice. There is an moderately-sized interaction between variant and voice, such that both *um* trends toward higher *young* ratings with Raven than with Penguin.

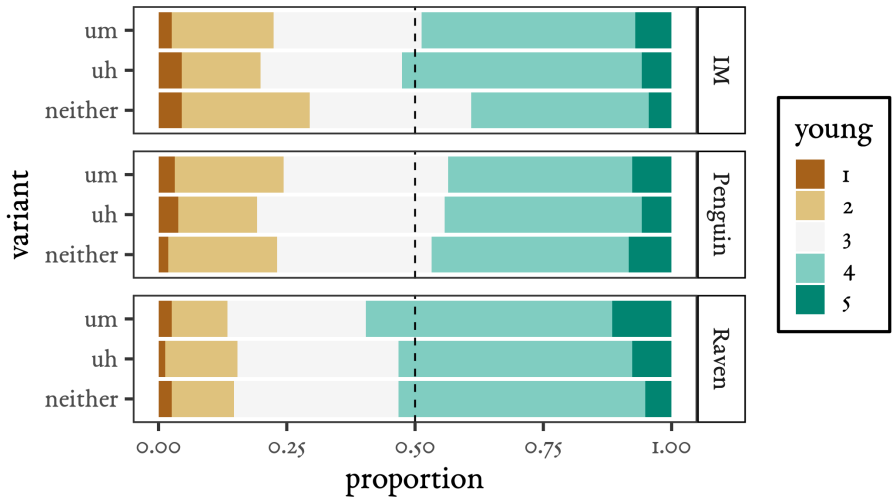

Figure 8: Proportion of *young* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-3.35	-4.16	-2.54	0.41	-8.11	0.00
2|3	-1.08	-1.76	-0.40	0.35	-3.11	0.00
3|4	0.53	-0.15	1.20	0.34	1.53	0.13
4|5	3.70	2.90	4.51	0.41	9.01	0.00
variant = uh	0.62	0.17	1.06	0.23	2.72	0.01
variant = um	0.50	0.06	0.93	0.22	2.25	0.02

Table 18: E1 model for the 'young' scale.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-4.62	-5.35	-3.89	0.37	-12.49	0.00
2|3	-1.99	-2.58	-1.40	0.30	-6.62	0.00
3|4	0.17	-0.40	0.74	0.29	0.59	0.55
4|5	3.75	3.10	4.41	0.33	11.30	0.00
variant = uh	-0.06	-0.50	0.38	0.22	-0.25	0.80
variant = um	-0.12	-0.56	0.33	0.23	-0.52	0.60
voice = raven	0.31	-0.35	0.96	0.33	0.92	0.36
uh x raven	0.17	-0.45	0.79	0.32	0.55	0.58
um x raven	0.54	-0.08	1.17	0.32	1.70	0.09

Table 19: E2 model for the 'young' scale.

3.2.3 Canadian

Figure 9 shows that in Experiment 1, while each variant receives roughly similar amounts of 4–5 ratings for Canadianness, *um* receives the most 1–2 responses, followed by *uh* and then *neither*. However, the model in Table 20 indicates that these differences are not robust.

The results in the figure for Experiment 2 are largely unclear, with the only apparent differences being that *uh* has very slightly more 4–5 responses than the other two variants, and that Penguin is rated as somewhat more Canadian than Raven. The model in Table 21 bears this out, with a robust effect for voice such that Raven is predicted to be rated less Canadian than Penguin ($\beta = -0.82$, $p = 0.02$) but no robust effects for variant.

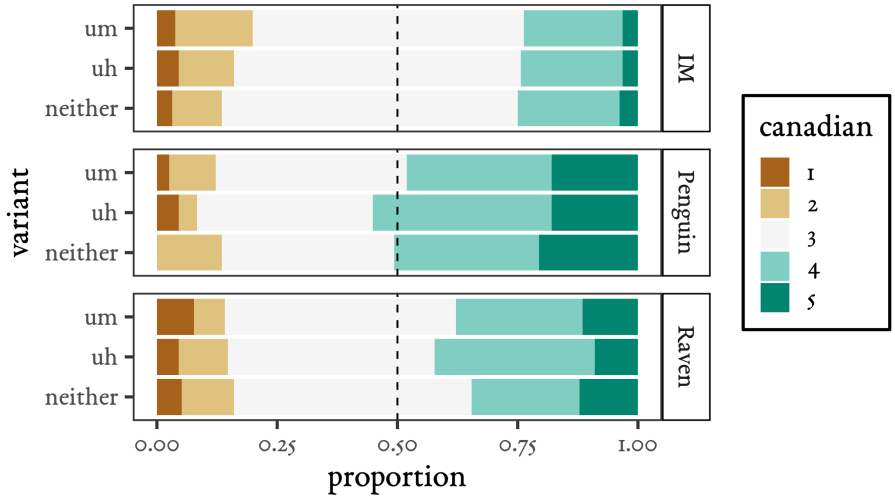

Figure 9: Proportion of *Canadian* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-4.55	-5.37	-3.74	0.42	-10.94	0.00
2|3	-2.42	-3.00	-1.84	0.30	-8.18	0.00
3|4	1.39	0.86	1.91	0.27	5.19	0.00
4|5	4.21	3.40	5.02	0.41	10.20	0.00
variant = uh	-0.20	-0.67	0.27	0.24	-0.83	0.41
variant = um	-0.31	-0.79	0.17	0.25	-1.26	0.21

Table 20: E1 model for the 'Canadian' scale.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-4.43	-5.04	-3.82	0.31	-14.20	0.00
2|3	-2.67	-3.17	-2.18	0.25	-10.63	0.00
3|4	0.36	-0.09	0.81	0.23	1.55	0.12
4|5	2.67	2.18	3.17	0.25	10.58	0.00
variant = uh	0.21	-0.10	0.52	0.16	1.35	0.18
variant = um	0.02	-0.29	0.33	0.16	0.10	0.92
voice = raven	-0.82	-1.53	-0.10	0.36	-2.24	0.02
uh x raven	0.03	-0.59	0.65	0.32	0.08	0.93
um x raven	0.18	-0.44	0.80	0.32	0.56	0.57

Table 21: E2 model for the 'Canadian' scale.

3.2.4 Casual

For Experiment 1, Figure 10 shows that the proportion of *casual* responses is largely the same across variants, although *uh* does receive slightly more casual ratings than *neither* (and to a lesser extent, *um*). The model in Table 22 indicates that neither *um* nor *uh* has a robust effect.

For Experiment 2, the figure indicates that both speakers received *casual* ratings of 4–5 the majority of the time. For Penguin, there is a straightforward pattern such that *um* receives the least casual ratings and *neither* receives the most. For Raven, while the amount of 4–5 responses is relatively stable across variants, the distribution of 4 vs. 5 responses is stratified by variant: she receives the least 5 responses with *um* and the most with *neither*. She also receives more 1–2 responses with *uh* than with *um* or *neither*. The model in Table 23 indicates that both *um* ($\beta = -0.46$, $p = 0.01$) and *uh* ($\beta = -0.18$, $p = 0.30$) are predicted to elicit lower *casual* responses, with the effect for *um* being much stronger and more robust than the effect for *uh*.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-4.06	-4.93	-3.19	0.44	-9.15	0.00
2|3	-1.72	-2.41	-1.02	0.36	-4.83	0.00
3|4	-0.24	-0.92	0.43	0.34	-0.70	0.48
4|5	2.34	1.62	3.05	0.37	6.38	0.00
variant = uh	0.24	-0.18	0.66	0.21	1.11	0.27
variant = um	0.12	-0.31	0.54	0.22	0.54	0.59

Table 22: E1 model for the 'casual' scale.

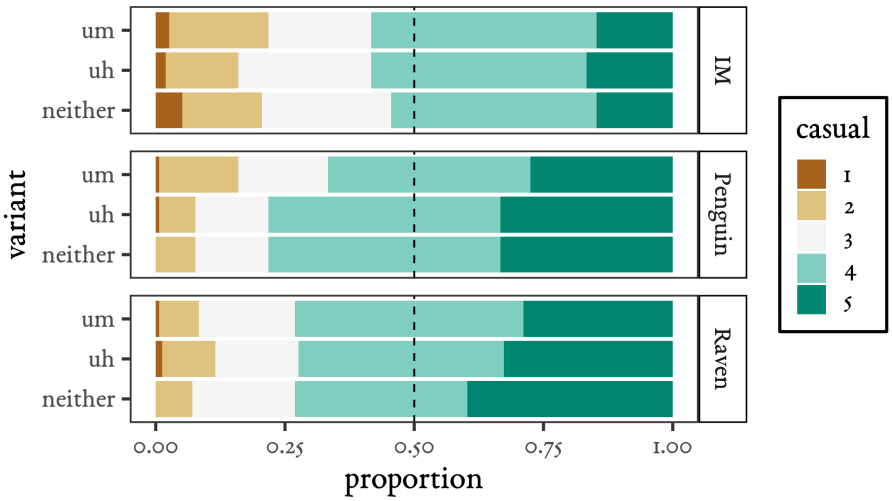

Figure 10: Proportion of *casual* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	*z*-score	*p*-value
1|2	-6.46	-7.56	-5.36	0.56	-11.54	0.00
2|3	-3.19	-3.83	-2.54	0.33	-9.64	0.00
3|4	-1.65	-2.25	-1.05	0.31	-5.35	0.00
4|5	0.79	0.20	1.38	0.30	2.63	0.01
variant = uh	-0.18	-0.53	0.17	0.18	-1.03	0.30
variant = um	-0.46	-0.81	-0.11	0.18	-2.60	0.01
voice = raven	0.05	-0.53	0.64	0.30	0.17	0.86
uh x raven	-0.31	-1.01	0.39	0.36	-0.88	0.38
um x raven	0.24	-0.45	0.93	0.35	0.69	0.49

Table 23: E2 model for the 'casual' scale.

3.2.5 Friendly

Figure 11 shows that for Experiment 1, *um* and *uh* receive lower *friendly* ratings than *neither*, with *uh* receiving slightly more *friendly* ratings than *um*. The model in Table 24 predicts *um* to elicit lower *friendly* ratings than *neither* ($\beta = -0.51$, $p = 0.02$). *Uh* is also predicted to elicit lower ratings than *neither*, but the effect is weaker and less robust ($\beta = -0.34$, $p = 0.12$).

For Experiment 2, the figure shows that Penguin receives higher *friendly* ratings than Raven overall. There are also some slight differences between variants in a similar direction to the patterns from Experiment 1 (i.e., *um* and *uh* are rated as less friendly than *neither*). The model in Table 25 predicts Raven to elicit lower ratings than Penguin ($\beta = -1.04$, $p \approx 0$). *Um* ($\beta = -0.21$, $p = 0.17$) and *uh* ($\beta = -0.27$, $p = 0.08$) are predicted to have lower ratings than the *neither* condition, with the *uh* effect being stronger and comparatively more robust.

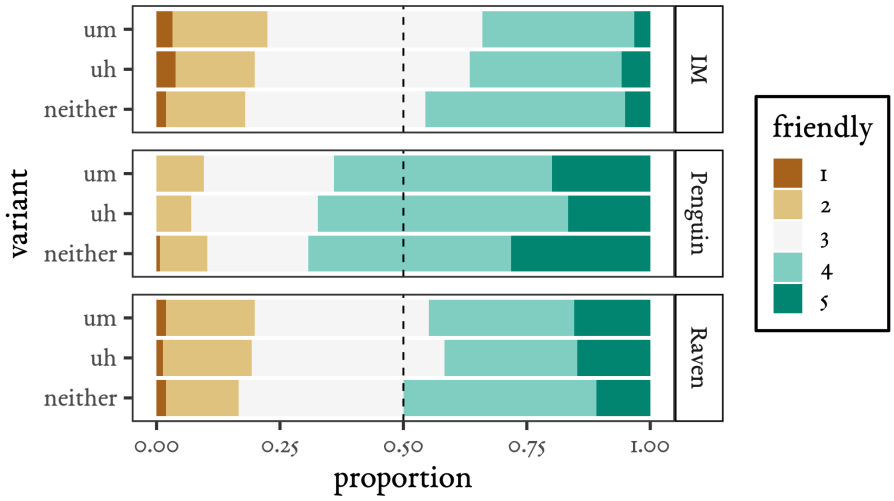

Figure 11: Proportion of *friendly* responses across both experiments.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-4.61	-5.54	-3.68	0.48	-9.69	0.00
2|3	-2.12	-2.87	-1.37	0.38	-5.54	0.00
3|4	0.34	-0.38	1.06	0.37	0.93	0.35
4|5	3.50	2.65	4.34	0.43	8.08	0.00
variant = uh	-0.34	-0.78	0.09	0.22	-1.55	0.12
variant = um	-0.51	-0.94	-0.08	0.22	-2.31	0.02

Table 24: E1 model for the 'friendly' scale.

term	estimate	conf. (low)	conf. (high)	std. error	z-score	p-value
1|2	-5.62	-6.43	-4.82	0.41	-13.71	0.00
2|3	-2.54	-3.01	-2.08	0.24	-10.76	0.00
3|4	-0.52	-0.94	-0.09	0.22	-2.37	0.02
4|5	1.85	1.41	2.30	0.23	8.16	0.00
variant = uh	-0.27	-0.57	0.03	0.15	-1.73	0.08
variant = um	-0.21	-0.51	0.09	0.15	-1.36	0.17
voice = raven	-1.04	-1.60	-0.48	0.29	-3.61	0.00
uh x raven	0.06	-0.54	0.66	0.31	0.20	0.84
um x raven	0.20	-0.40	0.81	0.31	0.67	0.50

Table 25: E2 model for the 'friendly' scale.

3.3 Summary and discussion

3.3.1 Pre-registered hypotheses

Table 26 summarizes the results of the scales for which there were pre-registered hypotheses, which are disussed in more detail in the following short sections.

	hypothesis	E1	E2
H_1:	*um* and *uh* hesitant	✓	✓
H_2:	*um* and *uh* unintelligent	✓	✗
H_3a:	*um* feminine	✓	✓
H_3b:	*uh* masculine	✗	✗
H_4a:	*um* polite	✗	✓
H_4b:	*uh* impolite	✓	✗

Table 26: Summary of pre-registered hypotheses.

Hesitance In line with the initial hypothesis, *um* and *uh* are rated as more hesitant than *neither* in all cases. This reflects the common perception, both among laypeople and linguists, that these words are used to indicate hesitation (such as to plan the following utterance, decide what to say, and so on). In both models, the effect size for *uh* is larger than that of *um*, though only slightly.

Intelligence The initial hypothesis, that *um* and *uh* are rated as less intelligent than the *neither* condition, was confirmed in IM, but not in speech. Since few participants mentioned intelligence in their qualitative responses, this finding is difficult to interpret. I can think of two (not necessarily mutually exclusive) reasons for the difference across registers. First, *um* and *uh* may be more salient in IM than in speech, since they are commonplace in speech but much less frequent in writing, even online (Gadanidis, 2018; Wieling et al., 2016). Second, the use of *um* and *uh* may be a less important cue for (perceived) intelligence than qualities of the voice, like pitch or creak, leading *um* and *uh* to be given less weight when vocal cues are available.

Gender The results of the masculine and feminine regression analyses suggest that listeners attribute femininity and/or un-masculinity to *um*, in line with the initial hypothesis. In IM, listeners are predicted to rate *um* as more feminine and less masculine than *neither*, and in speech, while there is no effect for femininity, *um* is rated as less masculine than *neither*. Given the attested change in progress toward *um*, which is led by women, it may be the case that listeners are picking up on this statistical pattern and using that to inform their responses (not necessarily consciously). The lack of a femininity effect in speech may be partially due to ceiling effects: Raven is rated as feminine or very feminine in almost all cases, and Penguin is rated as not very feminine or not at all feminine the majority of the time. This highlights the value of testing perceptual evaluation of gender using textual stimuli as well as audio stimuli: it may be the case that when speakers have access to audio cues to speaker gender, such as pitch, they assign lower weight to comparatively less-reliable cues, such as *um* use.

On the other hand, there is little to no gender association for *uh*: contrary to the initial hypothesis, there are no robust effects indicating that *uh* is masculine or un-feminine. This suggests that rather than falling on opposite positions of a gender continuum, the difference between *um* and *uh* is whether or not they are associated with gender at all. This is reminiscent of Campbell-Kibler's (2010) finding that for variable (ing), [-ɪn] and [-ɪŋ] are not opposites of one another, but have different social meanings entirely. As with

(ing), the locus of perceived gender is on the variant *um* rather than on the variable itself.

Politeness (and casualness) The initial hypothesis for politeness was that *um* would be rated as polite and *uh* as impolite. This was not confirmed: *um* and *uh* are both rated as impolite in IM (as well as more casual, which was not predicted but may be related), and in speech, *um* is rated as polite. The results from IM may be a function of perceived unconservativeness, with the less conservative forms being rated as less polite (and more casual). However, it is intriguing that rather than simply disappearing in speech, *um* is actually perceived as more polite and less casual than *neither*. This suggests that *um* may be linked to politeness and/or formality, which I will return to in the following section on qualitative responses.

3.3.2 Other findings

Raven's lower ratings On a number of scales, participants rate Raven "worse": she is rated as less intelligent, less polite, and less friendly. This is in line with previous work (e.g., Andrews, 2003) indicating that women tend to be rated lower on personal characteristics in this type of experiment.

Queerness The finding that *um* is rated as more queer in the *um* condition than in the *neither* condition (in speech) is surprising: queerness was originally included as a filler/distractor scale, and there were no hypotheses about queerness and *um* prior to running the experiment. However, it is interesting, and potentially suggestive, that this effect surfaces only in speech, rather than in IM. This is because gender and queerness are indexically linked: gay men are often construed as effeminate or less masculine than their straight counterparts. For example, Levon (2006) found in a perception study that ratings of "straight/-gay" and ratings of "effeminate/masculine" were negatively correlated, indicating that listeners associated straightness with masculinity and gayness with effeminacy. The queerness effect for *um* in speech may thus be linked to the same characteristics of *um* that lead listeners to rate it as less masculine (and, in IM, more feminine). However, it should be noted that much more work would be required to validate this speculation.

Youth In IM, *um* and *uh* are rated as more young than *neither*. However, there are no apparent effects for youth in speech. These differences may be interpretable as register differences: In IM, using discourse

markers like *um* and *uh* may connote youth because older individuals would be expected to use more formal language when communicating online. The lack of the youth effect in speech is then unsurprising, because speakers of all ages use *um* and *uh* in speech. Ceiling effects and cue weighting may play a role here as well: with more audio information available to the listener, the use of *um* or *uh* may be given less weight.

4 Qualitative responses

As described in §2.3, in addition to their answers using the Likert scales, participants had the option to provide qualitative responses to each stimulus. They were also asked for qualitative feedback before and after being debriefed. The opportunities for qualitative feedback were the same across both experiments, with the exception of Experiment 2's post-debriefing questionnaire, which added two questions, one about the meaning of *um* and *uh* and one about whether or not there was any difference between the two.

This section summarizes these qualitative comments, first for Experiment 1 and then for Experiment 2. It should be taken into consideration that these responses were given after some stimuli (all, in the case of the post-debrief questionnaire), and the scales, had already been seen. This means that when providing qualitative responses, speakers might be more likely to describe something as "hesitant", "feminine", and so on, rather than using descriptors that they had not been primed with. This is unfortunate, but unavoidable, since asking participants these questions prior to the experiment would have informed them of the experiment's purpose.

4.1 Experiment 1

General remarks While some participants made off-hand remarks about *uh* and *um* in the pre-debrief comment section, no participant identified *uh* and *um* as the focus of the experiment. Some common guesses related to language and gender (3a), linguistic profiling (3b), and perception of language in IM generally (3c).

(3) a. detecting whether we can perceive gender roles in text language (pre-debrief comment)

b. How we profile people based on how they use language (pre-debrief comment)

c. Understanding how we respond to messages given to us. Our perception of different text patterns (pre-debrief comment)

After being told that the focus was on *uh* and *um*, some participants said that they were surprised:

(4) a. Yes. I noticed those words but since I use them a lot when I text myself, it didn't come to mind that the experiment was about uh and um; and so they just flew by my head. (post-debrief comment)
b. yes because i was expecting it to be about gender (post-debrief comment)
c. Yes i was. It seemed to play a minor role in texting. I didnt see it as out of the ordinary but now that i notice it it does seem like an awkward way of texting (post-debrief comment)

Others were less so:

(5) a. Not super surprised. I study linguistics. (post-debrief comment)
b. No, because the words appeared enough throughout the experiment that it seemed to be one of the main focuses if not the main focus (post-debrief comment)
c. I'm not fully surprised, as that was something I noticed while answering the previous questions. They acted as flags when trying to identify things about the person. (post-debrief comment)

Some participants made prescriptive remarks about the language in the messages, such as the following:

(6) Typical conversation between two females that have know grasp of the English language. (stimulus 2, *um*)

However, prescriptive comments like this were rare, and many participants commented that the messages were similar to the way that they texted/IMed, making them relatable:

(7) a. these text messages looked a lot like me and my friends so it was a very easy and relatable experiment (pre-debrief comment)

b. Honestly I felt like I could have written 95% of those texts (pre-debrief comment)

c. a lot of these screenshots remind me of conversations I had with my friends (pre-debrief comment)

Um* and *uh In the comments that explicitly mention *uh* and *um*, by far the most common meaning attributed to them is hesitation. Some of the many examples are shown in (8):

(8) a. i interpreted hesitance with the "um" put in the second message, as a word of advice while also displaying possible reluctance to outright respond with concern for their friend (Stimulus 2, *um*)

b. seemed hesitant to give advice to the other person because they used "uh," but still worded their words nicely in order to not offend the other person. (Stimulus 2, *uh*)

c. The usage of "Um" made it seem like the subject is hesitant and worded their words very carefully in order to not offend the other person. (Stimulus 4, *um*)

d. One thing is that people include "uh" and "ums" to add hesitance and to try and soften there words. (pre-debrief comment)

e. While the IMs with uh and um stood out (and made me think of hesitation), I didn't think the study was specifically about them. (post-debrief comment)

f. I noticed when texts began with "uh" or "um" the message began to have a more hesitant and perhaps impolite tone to them that they might not have otherwise. (pre-debrief comment)

Stimulus 3, which involved the speaker confessing to their interlocutor that they had broken their mug while doing the dishes, was a site of particularly interesting commentary about hesitation. Some speakers, as in (9), found this stimulus strange in the *neither* condition:

(9) a. The question that sticks out to me for this message is the one on hesitation. As the person is delivering bad news, I'd imagine more texts, and more hesitation rather than just saying "I broke your mug." I'd imagine more apologizing before what happened was even mentioned. (Stimulus 3, *neither*)

b. no real apology for breaking the mug (Stimulus 3, *neither*)

Notice that the participant in (9a) explicitly notes that more hesitation would be warranted. This can be contrasted with the following comments, from the same stimuli but with *um*/*uh* included:

(10) a. sounds like the person was hesitant to admit that they had done wrong (Stimulus 3, *um*)
b. the "uh" makes it seem hesitant (Stimulus 3, *uh*)

It should be noted, though, that *um* was not always considered sufficient to indicate contrition:

(11) doesnt sound like theyre sorry, just sounds like theyre reporting what happened and they dont really care about the consequences (Stimulus 3, *um*)

It was rare for participants to explicitly contrast *uh* and *um*; I only identified one example of this, shown in (12):

(12) Now that it's been pointed out I find that 'um' is a lot more impolite (to use the experiment's terms) and 'uh' as more hesitant (although I can't remember if my answers show that or not). (post-debrief comment)

Interestingly, this statement does not match the trend established in the ordinal regression, where only *uh* was negatively correlated with politeness, and had a larger effect size than *um*.

Another interesting case is the following set of comments, all from the same participant:

(13) a. sounds like the person was hesitant to admit that they had done wrong (Stimulus 3, *um*)
b. It sounds like the blue dialogue person may have self esteem issues, very submissive (Stimulus 6, *um*)
c. Sounds like the person has confidence is assured of themselves (Stimulus 4, *uh*)
d. sounds like he or she is very concerned for their friend's glasses whereabouts (Stimulus 1, *uh*)

Notice that when *um* is present, the speaker is described as hesitant and submissive, but when *uh* is present,

these descriptors are not used—and in fact, in (13c), the speaker is described as confident and sure of themselves.

The following comments, again all from one (different) participant, are similarly interesting:

(14) a. the person initiated the conversation first, but it doesn't seem like the person is good at creating conversations. his answers are blunt and have no substance to them, I almost do not see the point of the person carrying out this conversation. (Stimulus 5, *uh*)

b. The usage of "Um" made it seem like the subject is hesitant and worded their words very carefully in order to not offend the other person.

Notice that while the participant does not specifically mention *uh* in (14a), they describe the *uh*-user as blunt and assume that they use he/him pronouns. In contrast, in (14b), *um* is explicitly identified and described as indicating hesitancy.

These patterns, while not substantial enough to draw major conclusions from, are suggestive of the social meanings that *uh* and *um* appear to hold for these participants.

4.2 Experiment 2

By-stimulus responses Only three participants mentioned *um* or *uh* in their optional by-stimulus responses, all three of which were with Penguin, and which are presented below:

(15) a. "um" and "I don't think so" are hedging, indicates reservation (Penguin, Stimulus 1, *um*)

b. pause after the "uh" with the second person felt very unnatural (Penguin, Stimulus 1, *uh*)

c. The way the second person said "uh" was very abrupt and plosive (kinda strange to me). I don't think the speaker is from America or Southern Ontario. (Penguin, Stimulus 1, *uh*)

Only in (15a) is any explicit function or meaning attributed to *um* (hedging/reservation). (15b) and (15c) are comments specifically about the naturalness of one stimulus, the *uh* version of Stimulus 1. In this stimulus, the vowel quality of the *uh* production is somewhat different than the others, which may be what these listeners are picking up on.

Not much can be gleaned from these responses about *um* or *uh*, but it is interesting that participants noticed them (or found them remarkable enough to enter a qualitative comment) in speech far less than in IM. This suggests that *um* and *uh* may be more salient in IM than in speech.

What does it mean when someone uses *uh* or *um*? The vast majority of participants indicated that *um* and *uh* indicate hesitation, nervousness, uncertainty, or speech planning. A small selection of these are shown in (16).

(16) a. I think it means that the speaker needs more time to collect their thoughts, or it can be that they are hesitant to say or share something.
b. The person is uncomfortable, hesitant or anxious
c. Either thinking or hesitate to say something
d. I think that it means that there is hesitation or uncertainty. Also i think that they allow people longer to think about what they want to say about something

Five participants implicated politeness or "softening"; their responses are shown in (17).

(17) a. I feel like is has many meanings. Mainly, the speaker doesn't want to sound to harsh (a certain tone of "uh" or "um" could perhaps be used to deliver bad news in a more polite manner), they want to sound sassy, they want to sound casual, or they are unsure of what to say/are nervous.
b. It means they're more hesistant to come off as impolite usually. They are trying proabbly subconsciously to tell the other person that they would rather not bring up the topic rather than jump right in.
c. It means they are a little hesitant or they're trying to think of ways to be more polite and considerate. Often used when you don't know what to say right away as it gives you some extra time to think
d. I think it is a sign of hesitation, uncertainty or to ease the tension to seem polite
e. I know they are hedging words; they "soften" what is going to be said next and express hesi-

tation or politeness.

Some responses were more idiosyncratic, implicating free variation (18a), self-confidence or honesty (18b), or even non-Canadianness or non-heterosexuality (18c) (these were likely due to the scales the participant had just been exposed to).

(18) a. They are trying to gather their thoughts, simply just put them in there for no reason, maybe trying to approach a subject more cautiously/with more hesitance

b. That they're unsure about an answer and trying to think of one. I think the use of these also relate to a person's self-confidence (don't have faith in their own abilities) and honesty (trying to think of an excuse).

c. I think it has to do with if an individual is unsure or they are more casual when they speak or it is an indicator that they are more introverted

Do *um* and *uh* have different meanings? When asked whether there was any difference in meaning between *um* and *uh*, 41 participants indicated that there was no difference, and/or that they were in free variation, as in (19).

(19) a. I don't think they have different meanings more so one is preferred over the other based on that persons tendencies

b. No, I think that they both serve the same purpose and therefore have the same definition.

c. I feel like they can be used interchangeably as they both can equate to hesitation

77 participants indicated that there was a difference, and 8 participants were unsure or did not provide enough detail for the researcher to determine what their opinion was.

***um* as thoughtful or deliberate** Of the 77 participants who indicated that there was a difference, a theme emerged where *um* was described as more thoughtful, deliberate, or socially-motivated than *uh*, which was described as more vacuous, less deliberate, or less thoughtful. This broad theme—*um* more thoughtful or intentional and/or *uh* less thoughtful or intentional—was echoed by 35 participants, a selection of whose

responses is below in (20).

(20) a. I think that 'um' is often used more when someone is consciously thinking about something and is willing to let other people know that they are hesitating or contemplating something, whereas 'uh' is more of a reflex

b. um seems more nervous than uh - uh seems like they lost their train of thought

c. yes. um can be an indication of someone taking some time to think as they have a possible answer, but uh can be a sign that the person doesn't really have an answer.

d. I think um is used more for actual thinking/hesitance. Uh is more of a conversation filler

e. I think "uh" is more vacuous and is said in other places (e.g. forgetting the next word in a sentence) whereas "um" has more meaning as a hedging word.

f. um i feel is more for thinking of an answer and uh is more for confusion or scrambling

g. I think of "um" as meaning "I'm thinking about something", and "uh" as more of a filler word in a sentence.

h. Although they can probably be used interchangeably, uh seems to show hesitation before committing to saying something whereas um might signify that a person is thinking about how to phrase their thoughts better.

i. I typically use um when I'm thinking and uh when I'm speechless

j. Not really. they are different because umm makes it seem more like a person is thinking while uh makes a person seem less sure

k. I think that um can mean that a person is considering something while uh is perhaps more of an involuntary hesitation.

Four participants had the opposite view:

(21) a. Um is more nervous/hesitant, uh is more a placeholder when searching for something to say

b. To me, 'um' sounds more hesitant whereas 'uh' sounds like the person is trying to think of how to elaborate on their previous point. 'Um' makes me feel like the person is really nervous.

c. "Uh" sounds like you're thinking or processing information you've just learned, while "um"

tends to sound like you're trying to find words to say and/or making up speech without thinking it through

d. Yes. Umm.. is like prolonging a decision and uh is for thinking purposefully

The remainder of the participants did not fit into either category. While the view of *um* as more intentional or thoughtful is clearly not universal, the number of responses that mention it is remarkable—and the comparatively lower number of responses identifying *uh* as the thoughtful or deliberate variant suggests that this is unlikely to be a coincidence.

Gender Four participants made explicit reference to gender when discussing the difference between *um* and *uh*. As shown in (22), all four indicated that *um* was more feminine and/or *uh* was more masculine, and never the reverse.

(22) a. before the experiment i did not think about it but now i thing um is more feminine

b. I think they have the same meaning, but to me um seems more feminine and uh seems more masculine. Perhaps this is due to um being said the mouth closed, and uh being said with the mouth open.

c. I tend to subconsciously correlate um with femininity and uh with masculinity. It's obviously not always true but it feels more natural

d. uh might be more masculine

It is *prima facie* unclear why (22b) associated closed-mouthedness with femininity and open-mouthedness with masculinity. However, as we will see below, this may be related to politeness (potentially mediated via a process of indirect indexicality, per Ochs 1992).

Politeness and formality Six participants made reference to formality and/or politeness. In five cases, *um* is judged as more polite or formal, and/or *uh* is judged as more impolite or casual.

(23) a. "um" may be a bit more polite possibly due to the fact that you have your mouth closed during this phrase while with "uh", your mouth is wide open and could be rude to someone

talking to you if you held it for too long

b. in passing conversation, i don't think they have different meanings. however, when they're drawn out in pause, i find that "um" sounds more polite and formal compared to "uh". in cases in say, a presentation, saying "um" for some reason makes me think that they have it more together than someone who says "uh"

c. 'Um' holds a heavier weight when hesitating. Uh could be more casual, easier to transition. Um alludes to needing more time to process

d. I think "uh" is a little more casual

e. To me, um sounds more like a word to use when youre thinking about what to say next and uh could be portrayed as more hesitant or sometimes rude

In only one case is *uh* described as less casual than *um*:

(24) 'Uh' might sound slightly less casual, and give off an impression of less intelligence, than 'um'

Intelligence Three participants made reference to intelligence. In all three cases, *uh* is judged less intelligent than *um*:

(25) a. Uh seems to be the less intelligent form. Um seems more relatable, and even kind of cute sometimes

b. 'Uh' might sound slightly less casual, and give off an impression of less intelligence, than 'um'

c. I feel like they essentially mean the same thing, but "uh" seems associated with less intelligence while "um" is more "I'm thinking about it, considering it." Not that this is true, but that's like my knee-jerk interpretation.

4.3 Summary and discussion

The thoughtfulness of *um* The most striking finding from the qualitative results is that, among the 77 participants who said that *um* and *uh* were different, 35 (almost half) indicated that *um* was more thoughtful, deliberate, and/or intentional than *uh*, which was described as "vacuous", "a reflex", and "involuntary".

(*Uh* is also described as less intelligent than *um*, which may be related.) That so many participants made comments along these lines is remarkable, especially given that, unlike some of the other characteristics associated with *um* and *uh* in the previous section, this characterization does not seem likely to have been cued by participants' exposure to the experimental scales: none of the scales in the experiment made reference to deliberateness, thoughtfulness, vacuity, involuntariness, or similar characteristics. On the other hand, this lack of an overt link between this pattern and the scales makes it difficult to connect to the quantitative results. I return to this issue in the general discussion.

Gender and politeness Gender and politeness were both mentioned much less often than thoughtfulness and deliberateness in the qualitative responses. However, evidence from when they *were* mentioned suggests that *um* is associated with femininity and politeness. The pair of responses that mention the mouth being opened or closed, reproduced as (26), are also tentative evidence for a link between femininity and politeness with *um*:

(26) a. "um" may be a bit more polite possibly due to the fact that you have your mouth closed during this phrase while with "uh", your mouth is wide open and you could be rude to someone talking to you if you held it for too long

b. I think they have the same meaning, but to me um seems more feminine and uh seems more masculine. Perhaps this is due to um being said the mouth closed, and uh being said with the mouth open.

It is not particularly surprising that the variant linked with politeness would also be linked with femininity, given the close ideological link between those two traits in the North American context. (Lakoff 1973: 56 describes it as a "general fact" that "women's speech sounds much more 'polite' than men's.") If the polite meaning is in fact related to the physiological characteristics of *um*, then the gender meaning may have accrued later, through a process of indirect indexicality (Ochs, 1992): if *um* is associated with politeness, and politeness is ideologically associated with femininity, then *um* can become associated with femininity indirectly.

5 Discussion and conclusions

The goal of this study was to identify the social meanings of *uh* and *um*, both in instant messaging and in speech. To that end, two experiments were conducted in which participants viewed instant messages or listened to conversations containing either *um*, *uh* or neither, and provided both quantitative and qualitative feedback about the individuals whose messages they read or voices they heard.

The quantitative results indicate that readers and listeners evaluate writers or speakers differently depending on whether their messages or utterances contain *um*, *uh*, or neither.

The various characteristics in this study fall into two overall categories. One category consists of characteristics that are shared between both *um* and *uh* (though potentially to different degrees), such as youth, casualness and impoliteness (in IM) and un-casualness (in speech), and of course hesitancy (in both speech and IM). The other category consists of characteristics associated only or largely with *um*, such as femininity and un-masculinity, politeness, and, in speech, queerness. There is no third category for *uh*: in all of the quantitative results, in all cases where *uh* is distinguished from *neither*, *um* is as well—in the same direction. And although *uh* is described as masculine by three participants, only one does so without explicitly contrasting it with *um*. (The dominance of the binary in Western gender ideologies, as well as the presence of both "masculine" and "feminine" in the scales presented to participants, makes it relatively unsurprising that participants might assign masculinity to *uh* when explicitly asked to contrast it with *um*.)

In other words, these data suggest that relative to *uh*, the variant *um* is the primary carrier of social meaning (see also Campbell-Kibler, 2010). Where *uh* carries social meaning, it appears to do so as part of the larger *uh*/*um* variable, whereas *um* carries social meanings of its own—in particular, gender.

Additional support for this argument comes from the trend in the qualitative data that *um* is judged as more thoughtful or deliberate than *uh*, which is described as "vacuous", "involuntary", a "filler word" and "a reflex". In a way, these listeners are ascribing psycholinguistic meaning to *uh* and sociolinguistic meaning to *um*. Words like "vacuous" and "involuntary" are part of one common psycholinguistic understanding of what filled pauses are: symptoms that are produced automatically and mean nothing more than processing difficulty (e.g., Levelt, 1983).

Under this analysis, both *um* and *uh* can be taken to have a first-order index of "hesitation", but *um* can

be taken to index a specific *type* of hesitation—one associated with femininity, un-masculinity, politeness, and thoughtfulness. In a previous qualitative analysis of IM corpus data, I have argued that *um* can function as a mitigator of face challenges (Gadanidis, 2018). For example, (27) is taken from a IM conversation between a cohabiting and romantically-involved man and woman. A uses *um* (which she spells ⟨uhm⟩), along with a host of other mitigative or hedging markers and syntactic features, to soften criticism of her partner, B, who mistakenly left the rice cooker on.

(27) A: **Uhm**, the rice cooker is super hot cuz it was still in keep warm mode o-o
B: Holy fuck sorry
A: It's okay, let's just be careful next time o.o

Rather than direct admonishment or criticism such as "you left the rice cooker on", "be careful next time", A uses the passive "it was still in keep warm mode" and the cohortative "let's just be careful next time". This is presumably intended to protect B's face (even though B's apologetic reaction indicates that he understands that he is being criticized). *Um*'s association with femininity and politeness links it to this kind of face-protecting hesitation, which Tottie (2017: 20) describes[1] as having "a strong connotation of reluctance or unwillingness, reluctance to be tactless, to hurt or insult someone." While there is no reason to believe that *uh* cannot be used in this function as well (in fact, in Gadanidis 2018, I found evidence that it can), the data from this study suggest that *um* is most associated with this function.

This is not to say that *um* always fulfills this function in interaction. However, frequent use of the variant in this context can lead the variant to become associated not just with polite hesitation, but with people or personae who are ideologically-expected to hesitate politely (Raunomiaa 2003, cited in Bucholtz and Hall 2005, refers to this as *stance accretion*). As described above, through indirect indexicality (Ochs, 1992), this category of people likely includes women, who are ideologically-expected to speak politely (Lakoff, 1973). (It is likely impossible to determine whether the feminine chicken or polite egg came first.) The finding in some corpora that more educated speakers are more likely to use *um* (Wieling et al., 2016) may also be linked to ideologies of politeness and properness.

[1]Note that Tottie does *not* identify this type of hesitation with *um*. Rather, she identifies it with hesitation in writing. I include the quotation here as an apt description of the type of hesitation I am referring to, not to imply that Tottie agrees with my analysis.

Under this analysis, *uh* can be construed as a more canonical or typical filled pause (at least in perception). This may help explain why, in the ordinal regression models for two scales commonly associated with filled pauses more generally—more hesitant, less intelligent—*uh* and *um* trend in the same direction, but *uh* has the larger effect size.

It goes without saying that more work is needed to confirm whether the analysis I am sketching out here is on the right track. For example, a storyboard elicitation method such as the one used by Wiltschko, Denis, and D'Arcy (2018) could help determine the degree to which *um* and *uh* are felicitous in different types of discourse. The prediction would be that in discourse such as (27), *um* would be more felicitous than *uh*. Various other experiment designs, such as self-paced reading or eliciting naturalness judgements, could also be promising.

However, there is at least one independent piece of evidence supporting the present analysis. In a study comparing the use of *uh* and *um* in children with autism spectrum disorders (ASD), children with specific language impairment (SLI), and typically developing children (TD), Gorman et al. (2016) found that children with ASD were less likely to use *um* than TD children, while children with SLI used *um* at similar rates to their TD peers. Because "social-communicative deficits are a defining feature of ASD," the authors interpret their findings as "evidence for the essentially social function of fillers" (Gorman et al., 2016: 862). These results are predicted under an analysis where *um* plays a more social role than *uh*.[2]

This work also highlights the value of computer-mediated communication in perceptual evaluation studies. While potential register effects must be taken into account, using IM data allows for less-salient features, like *um* and *uh*, to be presented in a written format, where they are potentially more noticeable. The written format also eliminates complications associated with speakers' voices, allowing traits like perceived gender to be measured without the influence of prosodic factors like pitch. This method could potentially be productively applied to a great deal of different morphosyntactic or discourse-pragmatic variables, to investigate how readers perceive writer gender when different variants are used, when only text information is available.

[2]Of course, there are other potential explanations for the difference. For example, the results would also be predicted if children with ASD are less likely to pick up on ongoing changes in progress, and thus do not increase their use of *um*, the incoming variant in the ongoing change. More work would be necessary to say anything definitive about the link between *um* and ASD; until then, this is only suggestive evidence. A qualitative analysis of the utterances used in Gorman et al. (2016) would also be helpful in interpreting their results.

A number of recent studies (e.g., Denis & Gadanidis, 2018; Fruehwald, 2016; Wieling et al., 2016) have speculated that the documented change in progress from *uh* to *um* could be linked to a new discourse function, leading to its rise in speech. This study provides some evidence for what that discourse function might be, and how it might differ from *uh*'s: while both variants indicate hesitation, the type of hesitation they index differs, with *um* indexing a more polite or face-protecting hesitation than *uh*. While more work, with more voices, is necessary before coming to firm, generalizable conclusions about *um*, these results are a promising place to begin.

References

Andrews, D. R. (2003). Gender effects in a russian and american matched-guise study: A sociolinguistic comparison. *Russian Linguistics*, *27*(3), 287–311.

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

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, *7*, 585–614.

Campbell-Kibler, K. (2010). The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, *22*(3), 423–441.

Christensen, R. H. B. (2019). ordinal—regression models for ordinal data. R package version 2019.3-9. Retrieved from http://www.cran.r-project.org/package=ordinal/

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Davis, C., & Gutzmann, D. (2015). Use-conditional meaning and the semantics of pragmaticalization. In *Proceedings of Sinn und Bedeutung* (Vol. 19, pp. 197–213).

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12.

Denis, D., & Gadanidis, T. (2018). Before the rise of *um*, Paper presented at DiPVaC4, Helsinki, Finland.

Dlugan, A. (2011). How to Stop Saying Um, Uh, and Other Filler Words. *Six Minutes*. Retrieved from http://sixminutes.dlugan.com/stop-um-uh-filler-words/

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 6.

Gadanidis, T. (2018). *Um*, about that, *uh*, variable. M.A. forum paper, University of Toronto.

Gorman, K., Olson, L., Hill, A. P., Lunsford, R., Heeman, P. A., & van Santen, J. P. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, *9*(8), 854–865.

Kroch, A. (1994). Morphosyntactic variation. In *Proceedings of the Thirtieth Annual Meeting of the Chicago Linguistics Society* (Vol. 2, pp. 180–201).

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. doi:10.18637/jss.v082.i13

Lakoff, R. (1973). Language and woman's place. *Language in society*, *2*(1), 45–79.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Levon, E. (2006). Hearing "gay": Prosody, interpretation, and the affective judgments of men's speech. *American speech*, *81*(1), 56–78.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19–44.

Maddeaux, R., & Dinkin, A. (2017). Is like like like?: Evaluating the same variant across multiple variables. *Linguistics Vanguard*, *3*(1).

McKay, B., & McKay, K. (2012). Becoming Well-Spoken: How to Minimize Your Uh's and Um's. *Art of Manliness*. Retrieved from https://www.artofmanliness.com/articles/becoming-well-spoken-how-to-minimize-your-uhs-and-ums/

moremo. (2018). Interpretation of ordinal logistic regression. Cross Validated. Retrieved from https://stats.stackexchange.com/q/323621

Ochs, E. (1992). Indexing gender. In A. Duranti (Ed.), *Rethinking context: Language as an interactive phenomenon*.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rezvani, S. (2014). Four Ways to Stop Saying "Um" And Other Filler Words. *Forbes*. Retrieved from https://www.forbes.com/sites/work-in-progress/2014/12/17/four-ways-to-stop-saying-um-and-other-filler-words/

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & communication*, *23*(3-4), 193–229.

Tagliamonte, S. A. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, *45*(1), 1–32.

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, *83*(1), 3–34.

Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173–197.

Tottie, G. (2016). Planning what to say: Uh and um among the pragmatic markers. In G. Kaltenböck, E. Keizer, & A. Lohmann (Eds.), *Outside the clause: Form and function of extra-clausal constituents* (pp. 97–122). John Benjamins Publishing Company.

Tottie, G. (2017). From pause to word: *uh*, *um* and *er* in written American English. *English Language & Linguistics*, 1–26.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, *73*(sup1), 1–19. doi:10.1080/00031305.2019.1583913. eprint: https://doi.org/10.1080/00031305.2019.1583913

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1. Retrieved from https://CRAN.R-project.org/package%20=%20tidyverse

Wickham, H., Chang, W. et al. (2008). Ggplot2: An implementation of the grammar of graphics. *R package version 0.7*. Retrieved from http://CRAN.R-project.org/package%20=%20ggplot2

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*(2), 199–234.

Wiltschko, M., Denis, D., & D'Arcy, A. (2018). Deconstructing variation in pragmatic function: A transdisciplinary case study. *Language in Society*, *47*(4), 569–599.

A Dataset and analyses

The R code and datasets for both Experiment 1 and Experiment 2 (in `.rds` and `.csv` formats) are available at https://github.com/gadanidis/umanalysis.

B Experiment code

The JavaScript code for Experiment 1, along with associated image and HTML files, is available at https://github.com/gadanidis/ratemessages.

The JavaScript code for Experiment 2, along with associated audio and HTML files, is available at https://github.com/gadanidis/rateconversations.

C Experimental stimuli

C.1 Experiment 1 stimuli

This section provides images of the stimuli used in Experiment 1, for easier reference than GitHub link above. Where stimuli varied across conditions, only one version is shown.

C.1.1 Critical stimuli

All of the critical stimuli could contain either *um*, *uh*, or neither, always in the same place (and with the same capitalization, where applicable). In this appendix, only the versions with *uh* are shown.

Figure 12: Stimulus 1, *uh* version

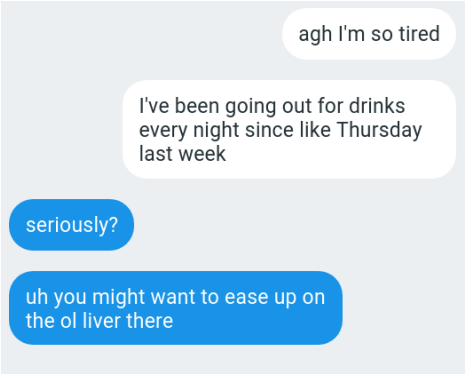

Figure 13: Stimulus 2, *uh* version

Figure 14: Stimulus 3, *uh* version

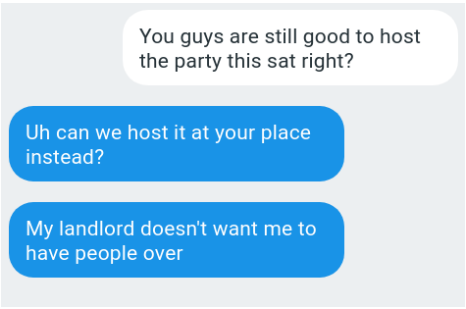

Figure 15: Stimulus 4, *uh* version

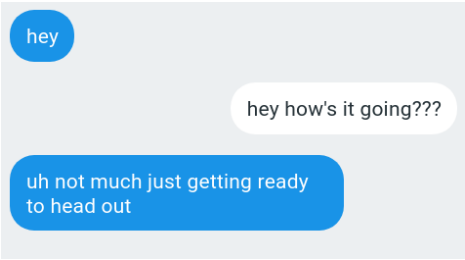

Figure 16: Stimulus 5, *uh* version

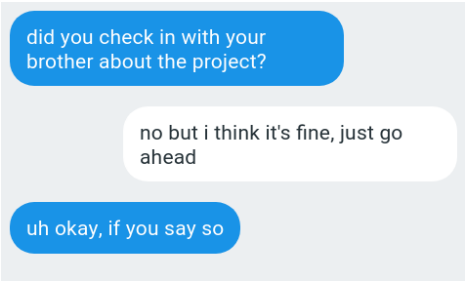

Figure 17: Stimulus 6, *uh* version

C.1.2 Filler stimuli

Some of the filler stimuli contained variation. Stimuli 7–10 contained either *lol* or *lmao*; the *lmao* versions are shown here. Stimuli 11 and 12 were invariant. Stimuli 13–16 contained either *eh* or *right*; the *eh* versions

are shown here.

Figure 18: Stimulus 7, *lmao* version

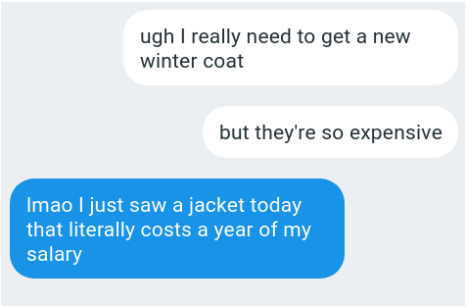

Figure 19: Stimulus 8, *lmao* version

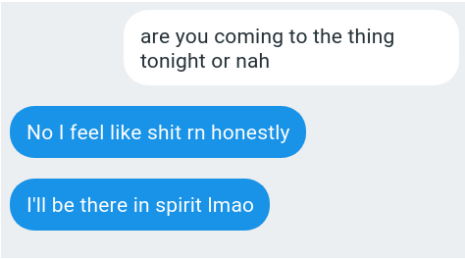

Figure 20: Stimulus 9, *lmao* version

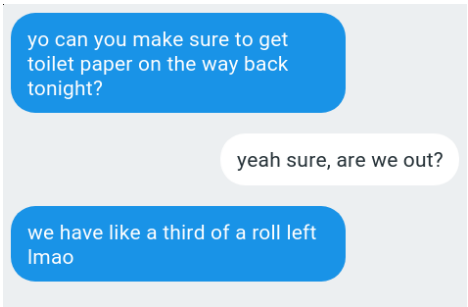

Figure 21: Stimulus 10, *lmao* version

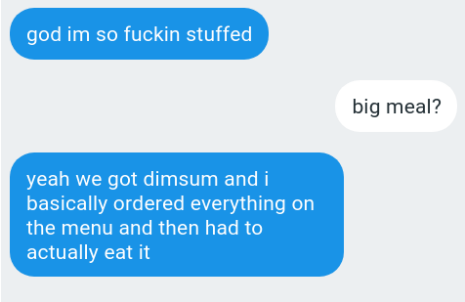

Figure 22: Stimulus 11 (invariant)

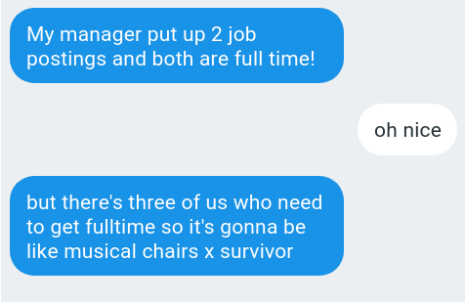

Figure 23: Stimulus 12 (invariant)

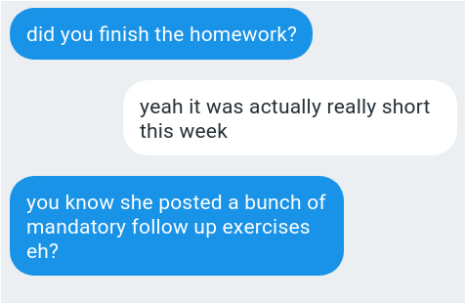

Figure 24: Stimulus 13, *eh* version

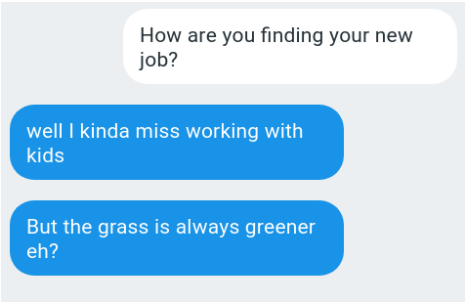

Figure 25: Stimulus 14, *eh* version

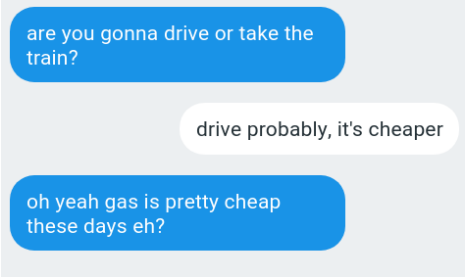

Figure 26: Stimulus 15, *eh* version

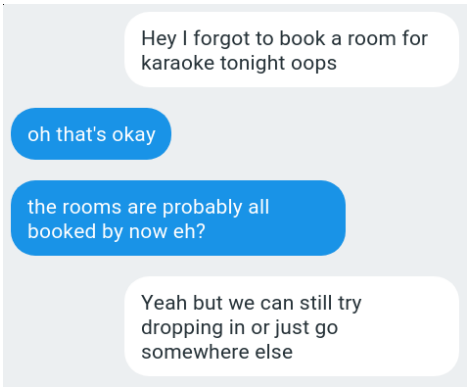

Figure 27: Stimulus 16, *eh* version

C.2 Experiment 2 stimuli

This section provides transcripts of the stimuli used in Experiment 2, for easier reference than the GitHub link above. In the critical stimuli, the point where *um*, *uh*, or neither could appear is marked with {UHUM}. The speaker who is rated is designated R, and the other speaker is designated O.

C.2.1 Critical stimuli

Stimulus 1

O Hey, have you seen my glasses? I know I left them somewhere around here, but I can't find them anywhere.

R {UHUM}, I don't think so. Where did you last see them?

Stimulus 2

O Ugh, I'm so tired. I've been going out for drinks every night since like Thursday last week.

R Seriously? {UHUM}, you might want to ease up on the ol' liver there.

Stimulus 3

R Hey, I have some bad news.

O Oh no, what happened?

R {UHUM}, I broke your mug while I was doing the dishes.

Stimulus 4

O You guys are still good to host the party this Saturday right?

R {UHUM}, can we host it at your place instead? My landlord doesn't want me to have people over.

Stimulus 5

R Oh hey!

O Hey, how's it going?

R {UHUM}, not bad, just getting ready to head out.

Stimulus 6

R Did you check in with your brother about the project?

O No, but I think it's okay if you just go ahead.

R {UHUM}, okay, if you say so.

C.2.2 Filler stimuli

Stimulus 7

O Hey, how's it going?

R Well, I bumped into my ex for the first time in like three weeks.

O Did you stop and chat?

R Yeah, it was super awkward.

Stimulus 8

O Ugh, I really need to get a new winter coat, but they're so expensive.

R Yeah, I just saw a jacket today that literally cost a year of my salary.

Stimulus 9

O Did you decide if you're coming to the thing tonight?

R No, I feel like shit honestly. I'm probably just going straight home after this.

Stimulus 10

R Hey, can you make sure to get some toilet paper on your way home tonight?

O Yeah sure. Are we almost out?

R Yeah, we have like a third of a roll left.

Stimulus 11

R God, I'm so fucking full.

O Big lunch?

R Yeah, I got dim sum with my friend and I basically ordered everything on the menu and then had to actually eat it.

Stimulus 12

R My manager put up two job postings and both are for full time.

O Oh nice!

R Yeah, but there's three people who are trying to move into full time, so it's going to be like musical chairs.

Stimulus 13

R Did you start the homework for this week yet?

O Yeah, it was actually really short this week.

R You know she posted a bunch of mandatory follow-up exercises, eh?

Stimulus 14

O How are you finding the new job?

R Well, I kind of miss working with kids, but the grass is always greener, eh?

Stimulus 15

R Are you guys going to drive, or take the train?

O Drive probably, it's cheaper.

R God I know eh? VIA Rail is so expensive.

Stimulus 16

R Hey I forgot to call and book the room for karaoke tonight. The rooms are probably all taken by now, eh?

O Yeah, but that's fine. We can try dropping in or just go somewhere else.

D R session info

The following is the output of the `sessionInfo()` command in R on my machine after running the script that was used for data analysis. The output provides information about my version of R, my operating system, and attached (activated) and loaded packages.

```
> sessionInfo()

R version 3.6.1 (2019-07-05)

Platform: x86_64-pc-linux-gnu (64-bit)

Running under: Manjaro Linux

BLAS:   /usr/lib/libblas.so.3.8.0

LAPACK: /usr/lib/liblapack.so.3.8.0


locale:

 [1] LC_CTYPE=en_CA.UTF-8       LC_NUMERIC=C

 [3] LC_TIME=en_CA.UTF-8        LC_COLLATE=en_CA.UTF-8

 [5] LC_MONETARY=en_CA.UTF-8    LC_MESSAGES=en_CA.UTF-8

 [7] LC_PAPER=en_CA.UTF-8       LC_NAME=C


attached base packages:

[1] stats     graphics  grDevices utils     datasets  methods   base


other attached packages:

 [1] xtable_1.8-4       cowplot_1.0.0      RColorBrewer_1.1-2

 [4] broom_0.5.2        ordinal_2019.4-25  jsonlite_1.6

 [7] RJSONIO_1.3-1.2    forcats_0.4.0      stringr_1.4.0

[10] dplyr_0.8.3        purrr_0.3.2        readr_1.3.1

[13] tidyr_1.0.0        tibble_2.1.3       ggplot2_3.2.1

[16] tidyverse_1.2.1    plyr_1.8.4         nvimcom_0.9-83
```

```
loaded via a namespace (and not attached):
 [1] tidyselect_0.2.5    reshape2_1.4.3      haven_2.1.1
 [4] lattice_0.20-38     tcltk_3.6.1         testthat_2.2.1
 [7] colorspace_1.4-1    vctrs_0.2.0         generics_0.0.2
[10] utf8_1.1.4          rlang_0.4.0         pillar_1.4.2
[13] glue_1.3.1          withr_2.1.2         modelr_0.1.5
[16] readxl_1.3.1        lifecycle_0.1.0     munsell_0.5.0
[19] gtable_0.3.0        cellranger_1.1.0    rvest_0.3.4
[22] labeling_0.3        fansi_0.4.0         Rcpp_1.0.2
[25] scales_1.0.0        backports_1.1.4     desc_1.2.0
[28] pkgload_1.0.2       hms_0.5.1           digest_0.6.20
[31] stringi_1.4.3       rprojroot_1.3-2     numDeriv_2016.8-1.1
[34] grid_3.6.1          cli_1.1.0           tools_3.6.1
[37] magrittr_1.5        lazyeval_0.2.2      ucminf_1.1-4
[40] crayon_1.3.4        pkgconfig_2.0.2     zeallot_0.1.0
[43] MASS_7.3-51.4       ellipsis_0.3.0      Matrix_1.2-17
[46] xml2_1.2.2          lubridate_1.7.4     assertthat_0.2.1
[49] httr_1.4.1          rstudioapi_0.10     R6_2.4.0
[52] nlme_3.1-140        compiler_3.6.1
```