

Chapter 5

The Social Evaluation of Abstract Phonological Structure

Given the robust evidence that speakers are producing variation between the abstract parameter of *PHL* and the abstract parameter of *NAS* demonstrated in Chapter 4 and the community-wide social patterning of this change outlined in Chapter 2, it follows that the allophonic restructuring of /æ/ in Philadelphia may also attract social evaluation. While social evaluation and the social motivation for sound change have been at the heart of sociolinguistic inquiry since Labov (1963), the ability of abstract phonological structure to be the target of social evaluation has been contested (see, e.g., Labov, 1993; Eckert and Labov, 2017). Because the allophonic restructuring of /æ/ in Philadelphia is a socially stratified abstract phonological change, it provides an important opportunity to test the social evaluation of phonological structure.

In this chapter, I present two experiments conducted prior to the projects reported in Chapters 2 and 4, which were designed to test the social evaluation of abstract structure. In §5.2, I present a matched-guise experiment designed to test the overall implicit social evaluation of *PHL* and *NAS*, finding that Philadelphian participants can in fact identify *PHL* and *NAS* along a scale of *accentedness*. In §5.3, this is followed by a magnitude estimation task which obtains participants' explicit evaluation of the six primary conditioning factors between *PHL* and *NAS*. I find that participants produce surprisingly systematic evaluations of these allophonic systems, with younger speakers

rating NAS highly and tense PHL tokens poorly and older speakers evaluating the conditioning factors rather than their phonetic realizations. These findings suggest that abstract phonological structure is targeted for social evaluation in this change from PHL to NAS in Philadelphia English.

5.1 The Unobservability of Structure

Speakers' ability to identify and furthermore evaluate structural variables such as a phonological rule is not well established in the literature. Labov (1993) argues that linguistic structure is unobservable, and that it is instead the phonetic output that is subject to social evaluation by listeners. This is not conceived of as *purely* phonetic output, but rather as the phonetic implementation of a surface phonological form, as in the tense production of an /æ/ allophone. Eckert and Labov (2017) point out, for example, that a production of [e:⁹] is not negatively evaluated when it appears in the word *yeah*, but is negatively evaluated as the phonetic output of the tense PHL rule. Additional work (Campbell-Kibler, 2007; Dinkin, 2015) carries this argument further with evidence that listeners attach social meaning to a variant itself (such as the use of “like” across the different variables of quotatives and discourse markers), regardless of the structural composition of the variable.

Eckert and Labov (2017) make the question of the evaluation of phonological structure explicit: “what kinds of phonological structures take on social meaning?” Eckert and Labov (2017) argue that while phonological variables are well suited for relaying social meaning, given that phonological variation rarely has referential meaning and is therefore maximally available for indexical meaning, the abstract structures governing relations between phonological entities is not well suited for this task. They go on to examine the case of phonological mergers, which occasionally attract social meaning, as in the case of the PIN-PEN merger in Northern California which is associated in production with an ‘outdoorsy’ lifestyle (Geenberg, 2014). Despite clear social associations being given to structural mergers, (Eckert and Labov, 2017, pg. 482) go on to discuss the lack of structural *commentary* about structural changes: “the merger of /i/ and /e/ before nasals is more likely to be noted as ‘He says *pin* for *pen*’ than ‘He says *pin* and *pen*’ the same.” This focus on lexical items or specific pronunciation of lexical items is taken as evidence that the structure of the merger is invisible to speakers.

While there is clear evidence from nearly every speaker interviewed in the PNC who provides metalinguistic commentary about language that their evaluation is attached to the phonetic form rather than the phonological structure, it does not necessarily follow that the phonological structure does not attract implicit social evaluation. The evaluation given to the PIN-PEN merger in California is one example of a case where listeners do provide social evaluation of a structural feature, even if they are not themselves aware of the structural component to their evaluation.

The phonological restructuring of /æ/ in Philadelphia provides a useful case study for investigating the observability of structure. As we have seen in Chapters 2 and 4, the tense and lax phonetic targets of a PHL speaker and a NAS speaker are almost identical. If listeners evaluate only the phonetic form of an allophone rather than the abstract structure of it, this predicts that listeners will provide a similar evaluation to a PHL and NAS speaker whose phonetic targets are similar. In this chapter, I present two experiments designed to test different aspects of the social evaluation of PHL and NAS. In Experiment 1 (§5.2), I employ a Matched Guise technique to test for the overall social evaluation of PHL and NAS, finding that listeners do identify a PHL speaker as more *accented* than a NAS speaker. In Experiment 2 (§5.3), I take a closer look into how listeners evaluate the different conditioning factors that make up PHL and NAS, finding that listeners' explicit acceptability scores are best described along structural, rather than phonetic, lines.

5.2 Experiment 1: Matched Guise

Since its development by Lambert et al. (1960) (see also, Anisfeld et al., 1962; Lambert et al., 1965; Lambert, 1967), the Matched Guise technique has been a widely used tool for obtaining implicit attitudes towards language. The basic concept of a Matched Guise experiment is to provide participants with two (or more) recordings. The participants do not know that the two samples of speech are from the same person, and are asked to judge the speaker of each recording along a number of social dimensions. As outlined in Gaies and Beebe (1991), Matched Guise tasks have two main purposes:

1. to elicit reactions to particular features indirectly, rather than having participants express

opinions about the features themselves

2. to control all variables other than the features in question

The Matched Guise technique has been applied to a wide range of sociolinguistic features, including obtaining participant attitudes toward specific languages in multilingual settings (see, e.g. Edwards, 1983; Lambert et al., 1965; Wölck, 1973; Gibbons, 1983), dialectal differences (Strongman and Woosley, 1967; Giles et al., 1992a; Elwell et al., 1984; Ohama et al., 2000; Arthur et al., 1974; Cargile, 1997), and has been particularly useful in obtaining attitude reactions to raciolinguistic dialects (Purnell et al., 1999). In addition to linguistics, social scientists have used the Matched Guise approach to investigate participant evaluation of visual cues (Elwell et al., 1984), including race (Dixon et al., 2002; Rubin and Smith, 1990), and age (Giles et al., 1992a).

Sociolinguists have also used the Matched Guise technique to investigate the social evaluation of more fine-grained linguistic features, such as speech rate and pitch variation (Addington, 1968; Brown et al., 1985; Ray et al., 1991; Giles et al., 1992b; Apple et al., 1979; Ray and Zahn, 1999). The ability to synthetically manipulate a recording has also made it possible to investigate listener attitudes towards specific features: these features can be manipulated within a single recording, mitigating the potential effect of phonetic differences in instances recorded.

As a first step towards investigating whether listeners evaluate the abstract organization of PHL distinctly from the abstract organization of NAS, a Matched Guise task provides a controlled way to elicit listeners' implicit evaluations. It is particularly important to investigate implicit social evaluation, given that the evidence drawn on in Eckert and Labov (2017) is primarily explicit in nature. Here, instead of asking whether participants comment on abstract structure, we rely on differences in social evaluations of a matched guise experiment as evidence of listeners' ability to evaluate abstract structure.

5.2.1 Participants

Participants were recruited through social media. Demographic data, including age, gender, race, and childhood zip code was collected. Only participants who reported living in a Philadelphia-area zip code between the ages of 1-18 were considered, resulting in a total of 52 participants.

Because the change in /æ/ occurred around 1983 in the community, participants born before this year were considered “older” and participants born after this year were considered “younger”. The data consisted of responses from 17 older and 35 younger participants.

5.2.2 Methods

Stimuli

Previous treatments of the Matched Guise technique have highlighted that task effects may play an important role in participants’ responses. Specifically, read passages differ from spontaneous speech in their prosody (Fowler, 1988; Blaauw, 1994), speech rate (Kowal et al., 1975), pause quantity and quality (Kowal et al., 1975; Guaitella, 1999), and tone boundaries (Howell and Kadi-Hanifi, 1991). These linguistic differences translate into differences in participant behavior: Smith and Bailey (1980) demonstrate that the difference in speech activity (whether it was read or spoken spontaneously) influences speaker perceptions. Furthermore, recent research on the effects on non-standard speech in experimental settings (e.g., Perry et al., 2017) reveal that nonstandard speech may be processed differently based on participant expectations. Because reading is a task associated with education, providing participants with one supraregional standard guise (in the form of NAS) and one local nonstandard guise (in the form of PHL) in read form is likely to introduce a potential task mismatch effect. In other words, participants may rate the PHL guise more harshly because it is seen as an unacceptable way to *read* rather than an unacceptable way to *speak*. Furthermore, the primary interest at hand is whether PHL and NAS receive distinct social evaluations in everyday interactions (not in read speech).

However, as any researcher who has attempted to use natural sociolinguistic interview data in an experiment can attest, finding passages from naturalistic sociolinguistic interviews that can be used for experimental purposes is a difficult feat. Many interviews are conducted in noisy settings, making acoustic manipulation very difficult and unnatural sounding. In addition, the researcher must find a section of the recording that contains the appropriate number and phonological conditions of the variable under investigation. For very frequent features, such as *ing-in* variation or t/d deletion, this may be possible. As highlighted in Chapter 4, however, test tokens of /æ/ occur

relatively infrequently in natural speech.

With a goal of including natural-sounding oral narrative stimuli that can be easily acoustically manipulated and also includes the right proportions of test /æ/ tokens, I adapt an oral narrative found in a sociolinguistic interview from the IHELP corpus. Because its baseline was an oral narrative, the story maintains a cadence of spoken – not read – speech. The narrative was modified to include more test /æ/ tokens, with special care towards ensuring that the PHL guise and the NAS guise each contained 9 tense tokens and 15 lax tokens. A trained phonetician read the story twice: once with all tense /æ/ tokens and once with all lax /æ/ tokens. To ensure that listeners did not obtain external social cues independent of /æ/ realization, both the PHL guise and the NAS guise used the same baseline recording of the story. All /æ/ tokens were spliced into this baseline story, meaning that all test tokens for both guises were comprised of spliced /æ/. The text for both guises is provided below. Tokens that would be tense under PHL are in bold, tokens that would be lax under PHL are in italics, and tokens that would be tense under NAS are underlined.

I got in a lot of trouble that night. And I didn't do anything wrong! Okay.

There was a big blizzard, and we didn't have **class**, so we all went down to Jake's to hang out there and play in the snow.

My mom was like "Don't bring your phone out", because I had just gotten a brand new phone. So she was like "Don't bring it, because if you manage to ruin it, your dad's not gonna be happy."

So I left it at Jake's house because I didn't wanna damage it.

So we were hanging out in the snow all day. He has like a little canyon behind his house that we were sledding in and stuff. So this **lasted** for like hours.

We got back to Jake's house **after** that, changed because our pants were all snowy, and went out again.

I get home that night, and I find out that my parents had called my cell phone like a hundred times, and it was this whole big thing. So I called her back and she started

going *bananas* on me. I started *laughing*, like “You told me not to bring my phone out!”

And then she got really *angry* that she hadn’t heard from me all day. It was pretty *bad*.

And then supposedly I was grounded, but that *lasted* like a day because she doesn’t stay *mad* at me for very long.

Task


Participants each heard only one guise (either PHL or NAS), and were asked to rate the speaker on a number of social dimensions based on what they had heard using a Likert scale, as shown in Figure 5.1.

Participants were able to listen to the story as many times as they liked. The social attributes selected for the Likert scale were chosen to match the broad social characteristics reported Campbell-Kibler (2007). While Campbell-Kibler (2007) ran several pilot studies to determine the most relevant social characteristics for her subjects, here I adopt the reported list of social characteristics as a broad insight into the social evaluation of the phonological structure of PHL vs. NAS. Future work may investigate a more nuanced set of social characteristics, but this is beyond the scope of the current dissertation.

In addition to the Likert scale ratings for social characteristics, participants were also provided with a free-form response box asking “How old do you think Brittany is” and a second free-form response prompting participants for additional reactions (see Figure 5.1).

5.2.3 Analysis and Results

Participant ratings were analyzed using ANOVA, with story guise as the first independent variable. Because 21 attributes were tested for, resulting *p*-values were Bonferroni corrected. Because the changing /æ/ system in Philadelphia is a change in progress, and because NAS is most prevalent in elite circles for younger speakers, we anticipate that participant age will be an important factor in participant ratings. Specifically, a speaker growing up before the advent of NAS in elite schools will be expected to have a different overall rating of the NAS guise than a younger speaker, for whom



0:00 / 1:02

Brittany is telling a story to some of her friends. After listening to the story, please tell us what you think Brittany is like, according to each of the following characteristics. You may listen to the story as many times as you want.

	Not at all						Very
Friendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Intelligent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wealthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Arrogant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sincere	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Honest	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Judgmental	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Competent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Aggressive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nosey	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hard-working	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trustworthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trendy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Attractive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Annoying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Approachable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tough	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spoiled	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How old do you think Brittany is?

What else do you think Brittany is like?

Figure 5.1: Screen shot of Matched Guise Task.

NAS may serve as a strong indicator of social class or social mobility. In the figures presented here, responses are binned by age of participant, with a date of birth of 1983 as the break point. Because 1983 was the changepoint in the speech community (Chapter 2), where NAS began to emerge in the production of Philadelphian speakers, this presents a sociophonological argument for binning participant age by this date. It is expected that on average, speakers born before this date acquired language in a PHL-only environment while speakers born after this date acquired language in a radically different environment which included two allophonic /æ/ systems as the input.

5.2.4 Results

For the majority of attributes, /æ/ system did not have a significant effect. I include a brief plot of these non-significant attributes in Figure 5.2, which provides some insight into the overall social evaluation of the speaker (regardless of /æ/ guise). Immediately apparent is the effect of story context: this is a narrative about a speakers' parents not grounding her, and we see she is somewhat unsurprisingly rated high on *spoiled*. This young sounding female voice also is rated as *approachable*, *friendly*, *sincere*, *trendy*, and *wealthy*. She is not considered *hard working*, *aggressive* or *shy*.

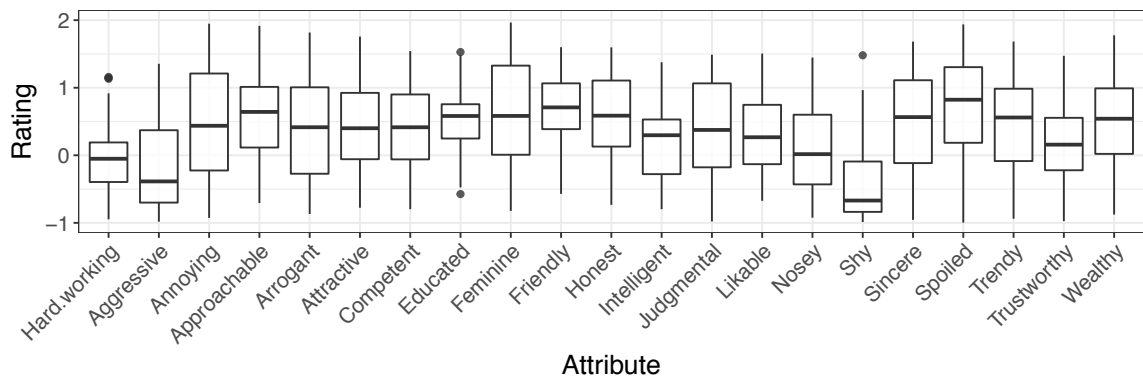


Figure 5.2: Non-significant attributes from the Matched Guise task.

Against the overall social characteristics attributed to the speaker, there is a single trait that is affected by story guise: *accented*. This result is shown in Table 5.1.

	Estimate	Std. Error	t value	Pr(> t)
Accented				
Story (PHL)	0.88	0.26	3.38	0.02*
Age (younger)	-0.52	0.39	-1.3	.99
Story:Age	-0.52	0.57	-0.89	.99

Table 5.1: ANOVA results for *Accented*; *p*-value presents Bonferroni correction.

5.2.5 PHL is rated as more Accented than NAS

As shown in Table 5.1, a PHL guise has a strongest effect on the standardized coefficient for *accented* ratings, with an estimate of 0.88. This serves as an important sanity check on the sociolinguistic awareness of the participants. Unlike a supraregional standard like the NAS system, Philadelphia English is a nonstandard regional dialect, and is interpreted and often maligned by the general public as an “accent”. Philadelphia English was included as a contestant in Gawker’s 2014 “America’s Ugliest Accent” competition (Evans, 2014), and dozens of sociolinguistic interviews in the Philadelphia Neighborhood Corpus contain metalinguistic commentary by Philadelphians about the Philadelphia accent. It is not surprising, therefore, that Philadelphian participants from both age groups rate the PHL guise as more *accented*.

That Philadelphians of both age groups rate the PHL guise as more *accented* speaks to their ability to detect linguistic variation. However, it does not necessarily follow that an identification of linguistic variation equates to social evaluation of that variation. We may expect, for instance, that a Philadelphian aware of the social patterning of PHL and NAS across school systems may rate a NAS guise as more *wealthy* or more *educated*, and a PHL guise as adjectives that align with social evaluation of working class speakers, such as *aggressive* or *hard working*. The lack of social adjectives assigned to the PHL or NAS guise suggests that this change has not attracted overall social meaning. However, as we have seen, listeners are still able to identify PHL as sounding more *accented*; we may then turn to the question of how listeners rate the six main conditioning factors governing the allophony of /æ/. For this, we turn to a Magnitude Estimation task.

5.3 Experiment 2: Magnitude Estimation

While Experiment 1 demonstrated that PHL and NAS are identifiably different, the next step is to ask exactly how the six main conditioning factors governing the allophony of /æ/ contribute to listener evaluation. We can conceive of several levels to the phonological architecture which may be the target for acceptability judgments. First, we've seen in Chapter 4 that PHL and NAS behave as two variants of an overall /æ/ parameter, with these two systems competing wholesale in production. The uniformity of these systems in production might lead a reader to expect similar uniformity in acceptability: in other words, we might expect to see all the PHL phonological contexts rated alike and all the NAS phonological contexts rated alike.

However, as we have seen in the results from previous Matched Guise experiments, there is additional evidence that phonetic variation such as speech rate (Brown et al., 1985) or F0 (Levon, 2014) may also be the target of evaluation. Adding to this, we have seen that allophones have also been found to be the target of evaluation: Labov (2001) found Philadelphia speakers negatively rating only the tense forms of /æ/, rather than the system as a whole. This has been taken (e.g., Eckert and Labov, 2017) as evidence that social evaluation targets a surface form (i.e., the phonetics of hyper-tense *bad* differently from the phonetics of phonetically mitigated lax *bad*) rather than the underlying grammar; the evidence provided in §5.2 suggests instead that social evaluation may target any number of levels of phonology: the abstract parameters governing an allophonic split, as we have seen in §5.2, an allophone (Labov, 2001), and the phonetics (Brown et al., 1985; Levon, 2014).

Here, I investigate how the phonological conditioning factors differentiating PHL and NAS are rated, using a modified version of a Magnitude Estimation task (Sprouse, 2007, 2011; Bard et al., 1996; Cowart, 1997; Featherston, 2005). Magnitude Estimation is a task quite widely used in experimental syntax (Sprouse, 2007), in which participants are encouraged to rate items in comparison to a reference item. For example, participants may be told a reference line is length 100, and asked to rate subsequent lines by comparing them to the reference, as shown in Figure 5.3.

The goal of a magnitude estimation task is to capture a perceptual scale, rather than a physical scale. For instance, while doubling the lumens of a light will double its physical brightness, par-

Reference:	_____
Length:	100
Item 1:	_____
Length:	200
Item 2:	_____
Length:	50
Item 3:	_____
Length:	300

Figure 5.3: Magnitude estimation of the length of a line. From Sprouse (2007)

ticipants do not react in a linear way to this increase; such a light is rated as brighter but not by double. Bard et al. (1996) adapted this task to acceptability judgment data, allowing participants to rate sentences with marginal acceptability along a gradient and non limited scale. Here, I adapt this method to acquire acceptability judgments of phonetic realizations. I present participants with auditory stimuli and ask them to rate each stimulus in comparison to a reference stimulus. The task and stimuli are reported in more detail below.

5.3.1 Participants

Participants were the same as in Experiment 1; participants completed the Matched Guise task first, then went on to complete the Magnitude Estimation task.

5.3.2 Methods

Stimuli

Stimuli consisted of 96 tokens total, comprised of 50% test tokens containing a target /æ/ word and 50% filler tokens that did not contain /æ/. Of the test words, each participant heard a tense and a lax form of each word. Lists were presented in four blocks, and were prerandomized so that a participant did not hear a tense and a lax token of the same token within a single block. Likewise, each list contained no more than three test tokens in a row. Stimuli were recorded in a

sound-attenuated sound booth. A tense and a lax form were recorded for each /æ/ word, meaning that no stimuli had to be acoustically manipulated.

Task


The experiment consisted of a training and a test phase. During the training phase, participants were introduced to the concept of magnitude estimation with the line task presented in Figure 5.3. After this training phase, participants entered the phonological ratings phase. They were presented with a reference stimulus (*chocolate*) and told that it received a rating of 100 for being “well pronounced.”

Participants were then asked to rate stimuli for how “well pronounced” they sounded, using the reference stimulus rated 100 as a reference. An example is provided in Figure 5.4. Each page of the experiment contained 24 tokens, and the reference stimulus was repeated at the beginning of each page. This task included one important modification from the classic Magnitude Estimation paradigm: rather than allowing participants to input any unbounded value, they were asked to slide a slider somewhere between 0 and 150 for pronunciation value⁹. The experiment was run through Qualtrix and results were analyzed using R.

5.3.3 Analysis and Results

The results of the Magnitude Estimation task suggest a somewhat complicated social evaluation of /æ/ conditioning factors, which differ between the older participants and the younger participants. Here, I split participants into age groups based on the community-wide sociolinguistic patterns found in Chapter 2. Older speakers are defined as any speaker born before 1983, which was selected as the best changepoint in the community-wide data from the PNC and IHELP corpora. Older speakers would have largely acquired their language in a PHL-only environment, while younger speakers would have acquired language in a mixed environment consisting of both PHL and NAS.

⁹A pilot study giving participants a blank line for response resulted in a majority of ‘99’ answers, presumably because participants wanted to finish the experiment as quickly as possible, and typing ‘99’ provides a quick response. Changing to a slider bar resulted in a much wider range of responses.



Part 2: Judging Words

In the last of the experiment you used numbers to estimate the lengths of lines. In this part you will use numbers in a similar way to judge how good some English words sound to you.

As with the lines in Part 1, you will first hear a reference word with a judgment of 100 like this:

chocolate

 100

You should listen to the word, and determine how well pronounced it sounds to you. For each word, you will assign a number to show how well pronounced you think it sounds.

Try to use a wide range of numbers and to distinguish as many degrees of pronunciation as possible.

Try not to dwell on any one word for very long; instead, try to go with your first reaction for each word.

There are 95 words for you to judge. The reference word will be repeated every 8 words in **bold** for your convenience.

chocolate

 100

bad good

0 10 20 30 40 50 60 70 80 90 100 101 201 301 401 501

flat

Figure 5.4: Modified Magnitude Estimation task rating the “well pronouncedness” of words against a reference word with score 100.

Older participants downgrade tense PHL

I begin by analyzing the results of the older speakers rating PHL-consistent tokens. We expect this data to align with the findings of (Labov, 2001), who found Philadelphian listeners negatively rating the tense allophone of /æ/ but not the lax allophone of /æ/. We see in Figure 5.5 a direct replication of these findings, with these older listeners downgrading tense /æ/ tokens and rating lax /æ/ tokens quite highly.

	Estimate	Std. Error	t value
(Intercept)	0.55	0.28	2.03*
Realization (tense)	-0.69	0.18	-3.85**
Gender (male)	0.41	0.54	-.76
Realization(tense):Gender(m)	-0.61	0.35	-1.75

Table 5.2: Tense PHL tokens downgraded by older speakers.

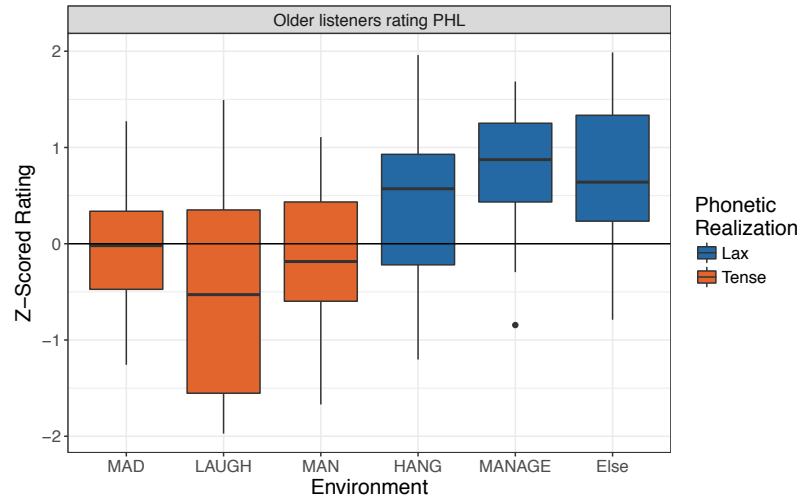


Figure 5.5: Older listeners downgrade tense PHL tokens.

A mixed effects model of this data with main effects of Realization (tense or lax) and Gender (male or female) and random intercept for participant is presented in Table 5.2, which finds a significant effect of tense realization on the evaluation of these tokens. This data serves primarily as a validation of the experiment: we find that the older participants rate PHL tokens consistently with the data reported in Labov (2001). In §5.3.3, I explore the systemic properties of this evaluation in more detail.

Younger participants learn two evaluation systems

I turn next next to the results from younger participants, meaning any participant born after 1983. While we do not have production data from participants, we can reasonably expect that these younger participants would have been exposed to both PHL and NAS in the community. The results from Chapter 2 demonstrating the social stratification of NAS in the elite non-Catholic schools in Philadelphia combined with the different social evaluations of PHL and NAS found in the Matched Guise experiment in §5.2 furthermore suggest that we might see a different pattern of overt ratings for PHL-consistent and NAS-consistent tokens from younger participants than from participants born before 1983. In other words, as the production of the community is in flux, younger

participants may in turn adjust their overt ratings of pronunciations in line with the changing community norms.

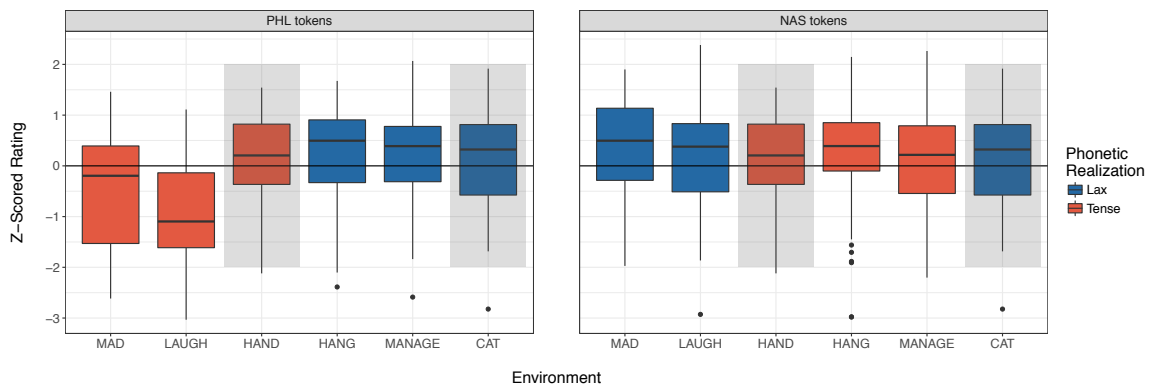


Figure 5.6: NAS rated highly by younger speakers (right); tense PHL downgraded (left).

Figure 5.6 shows the results from younger participants rating PHL-consistent tokens (left) and NAS-consistent tokens (right). Note that the HAND condition and the CAT conditions are the same in both facets, since HAND is produced as tense by both systems and CAT is produced lax by both systems; these boxplots have been grayed out as a visual aid to this fact. Let's first address the NAS-consistent tokens (right panel). Younger participants rate all NAS-consistent tokens highly, regardless of phonetic realization. This suggests that younger speakers have adopted a systemic evaluation of NAS: namely, that NAS-consistent tokens are all positively evaluated. Turning to the PHL-consistent tokens, we find that the younger participants have also learned the traditional community evaluations of PHL-consistent tokens, with tense realizations downgraded and lax realizations rated highly. Note that the only violation of this generalization is in the high ratings young speakers give to the MAN class, which I analyze as interference from participants' positive NAS evaluations.

These results suggest that younger participants are applying two evaluation systems. As an evaluation system, this means that younger speakers first apply a positive rating to any tokens that are NAS-consistent. This is relatively unsurprising, given the overt nature of this task: we have seen in Chapter 2 that the social patterning of NAS in Philadelphia resembles a change from above, in which the incoming NAS system is expected to be evaluated positively. That NAS is rated positively

may predict all PHL-consistent tokens to be downgraded. However, this is not what we see. After applying a NAS-positive evaluation, participants then also apply a PHL evaluation system to any remaining tokens. That is, any tense tokens of MAD or LAUGH are rated low, in accordance with the PHL evaluation system. Tense tokens of HAND remain high, as they have already been highly evaluated using the NAS evaluation. Finally, lax tokens of HANG and MANAGE get rated highly, also in accordance with the PHL evaluation system. In other words, participants have learned a NAS evaluation as well as the traditional community norms for evaluation of PHL-consistent tokens, which results in a high rating for lax PHL tokens and a low rating for tense PHL tokens. These results are confirmed by a parsimonious mixed-effects model (Bates et al., 2015), which I describe in detail below.

Older speakers evaluation of conditioning factors

I turn finally to the older speakers' ratings of NAS, comparing these ratings to their ratings of PHL. Again, the HAND and CAT class words are grayed out, as a visual reminder that these two classes share conditioning between PHL and NAS, and are therefore given the same ratings. Here, a somewhat surprising picture emerges (Figure 5.7). We see here that participants are rating tokens according to their conditioning factor, rather than according to their phonetic realization or the system they are consistent with. In other words, older speakers rate tokens MAD, LAUGH, and HAND negatively regardless of whether they were produced as tense or lax. Likewise, older participants rate tokens of HANG, MANAGE, and CAT positively regardless of phonetic realization.

Mixed effects modelling

Here, I present the results of a parsimonious mixed-effects model fit and optimized separately for the younger participants and the older participants. In both models, I begin with a maximal model with the following fixed effects.

Realization Realization was treatment coded as a binary factor (Tense or Lax), with Lax as the reference level. This was chosen as a reference level due to the evidence that PHL speakers treat lax realizations as a default and tense as a negative value (Labov, 2001).

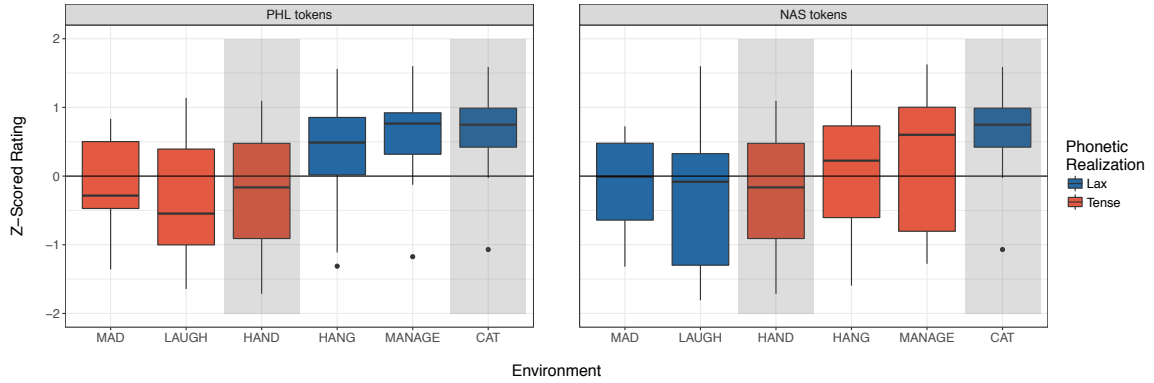


Figure 5.7: Older speakers rate MAD, LAUGH, MAN tokens low and HANG, MANAGE, CAT tokens high.

System Conformity The overlapping conditioning factors between PHL and NAS require some thought for the model, because they could potentially be analyzed as either PHL-consistent or NAS-consistent but not both. I resolve this by splitting the “system” parameter into two fixed effects: Conformity to PHL and Conformity to NAS. Conformity to PHL was coded as a (1) for tense tokens of HAND, LAUGH, MAD, and lax tokens of MANAGE, HANG, CAT, and a (0) elsewhere. Conformity to NAS was coded as (1) for tense tokens of HAND, MANAGE, HANG and lax tokens of LAUGH, MAD, CAT, and a (0) elsewhere. This enables us to test ratings of tense HAND and lax CAT as members of PHL as well as NAS.

Conditioning Environment Conditioning Environment was treatment coded, with six levels (HAND, LAUGH, MAD, MANAGE, HANG, CAT). Here, CAT was selected as the reference level because its lax production can be considered the default, unmarked variant.

PHL Conditioning This effect was included to test the effect suggested by the results in Figure 5.7 that older speakers downgrade the PHL tense-producing conditioning environments as a whole rather than the tense realization of those environments. PHL Conditioning, unlike Conformity to PHL, represents the conditioning environments only and not the realization of those environments. For PHL Conditioning, HAND, LAUGH, and MAD received a (1) and MANAGE, HANG, and CAT

received a (0).

Gender Models were tested with three different methods of coding participant self-reported gender¹⁰. The first method of coding gender was to sum code, given an assumption that males and females may produce different evaluations but that neither gender should be considered the reference level. However, in cases of language evaluation, it is not clear that sum coding gender is the best approach. There is an argument to be made that because the change from PHL to NAS has been described as a Change from Above, in which we expect women to lead in production, women may also lead in evaluation. For this reason, a second version of each model was also run with treatment coded gender with female as the reference level. Finally, because PHL is associated with an “accented” local dialect feature, there is the possibility that PHL-consistent tokens may be rated by participants as carrying covert prestige (Trudgill, 1974), which may predict that males positively evaluate PHL-consistent tokens. In all three versions of coding Gender, Gender did not improve model fit and was subsequently removed.

There is a large redundancy in including terms for Realization, Conformity to PHL, Conformity to NAS, Conditioning Environment, and PHL Conditioning in the same model. Conditioning Environment is colinear with PHL Conditioning, and the interaction of Conditioning Environment and Realization is colinear with Conformity to PHL and Conformity to NAS. For this reason, the maximal model and several of the near-maximal models were rank-deficient. All terms were tested for model fit using AIC and BIC comparison.

The results of the parsimonious mixed effects models for the younger listeners are consistent with the analysis provided above. Younger listeners have learned to downgrade tense tokens, but positively evaluate tense tokens that are consistent with NAS. In other words, younger listeners exhibit the operation of two evaluation systems: one in which NAS tokens are positively rated, and a second in which tense tokens that are inconsistent with NAS are negatively rated. It is worth pointing out that this is a slight break from the traditional rating pattern for PHL, since

¹⁰Participants were given a free-form response box for gender, to allow for queer and non-binary participants to self-identity. Participant responses fell categorically into a ‘male’ (‘m’, ‘M’, ‘man’, ‘male’) or ‘female’ (‘f’, ‘F’, ‘female’, ‘femail’, ‘woman’) response.

	Estimate	Std. Error	t value
(Intercept)	0.27	0.08	3.36**
Realization (Tense)	-1.00	0.12	-8.56***
PHL (true)	-0.06	0.07	-0.87
NAS (true)	-0.03	0.09	-0.34
Conditioning(PHL)	-0.02	0.07	-0.25
Realization (Tense) : NAS (true)	0.94	0.13	7.024***

Table 5.3: Younger speakers downgrade Tense but positively rate Tense NAS tokens.

	Estimate	Std. Error	t value
(Intercept)	0.20	0.25	0.81
Realization (Tense)	-0.17	0.32	-0.54
PHL (true)	0.33	0.21	1.58
NAS (true)	0.06	0.27	0.22
Conditioning(PHL)	-0.66	0.21	-3.13**
Realization (Tense) : NAS (true)	0.01	0.41	0.03

Table 5.4: Older speakers downgrade PHL conditioning factors, regardless of phonetic realization.

the traditional evaluation is to negatively evaluate all tense PHL tokens (including HAND tokens), while the results from the younger listeners demonstrate that tense HAND tokens are considered to be part of a NAS system and subsequently rated positively.

The results from the older listeners are somewhat more complicated. While we see the expected pattern of downgrading tense PHL tokens and upgrading lax PHL tokens, it is not clear how to interpret their evaluation of NAS-consistent tokens. Rather than rating all lax tokens of /æ/ positively and all tense tokens negatively, as would be expected if it is the phonetic production listeners evaluation rather than the phonological context, we in fact see older speakers not rating NAS tokens by their phonetic output. Instead, older speakers rate all conditioning factors that would be tense under PHL (MAD, LAUGH, HAND) as negative regardless of the phonetic production of the tokens, and all conditioning factors that would be lax under PHL (MANAGE, HANG, CAT) as positive regardless of the phonetic production of the tokens. There are two possible explanations for these results.

The first explanation is that older listeners' evaluation is tied to the phonological conditioning factors rather than to the phonetic production. In other words, listeners learn that the conditioning factors MAD, LAUGH, and HAND are negative while MANAGE, HANG and CAT are positive. Whether these tokens are produced as tense or lax does not matter all that much, as it is the underlying phonological environment that is evaluated rather than the phonetic production of that phonology. This would suggest that what seemed on the surface in Labov (2001) to be a straightforward case of participants negatively evaluating a tense production of an /æ/ allophone may instead have been participants evaluating the underlying conditioning of the allophone. This interpretation finds listeners evaluating the phonological system in a systematic way, contra the expectations in Eckert and Labov (2017).

A second explanation may be that older participants have several competing social evaluations available. First, any tense PHL token gets negatively evaluated while lax tokens are taken to be neutral or positive. Second, any tense token that conforms to NAS only may either be unnoticed or may be associated with a positive accent and so receives a high rating. This accounts for the positive ratings of HANG and MANAGE regardless of phonetic output. Finally, listeners would also need to apply an additional socially-motivated negative evaluation for lax productions of traditionally tense PHL tokens (MAD and LAUGH class), perhaps as a negative response to tokens that sound out-group. So any tense tokens of MAD and LAUGH are negatively evaluated because of the traditional evaluation, but lax tokens of these classes are also negatively evaluated because they don't sound Philadelphian enough.

5.4 Discussion

In this chapter, I have attempted to shed some light on the phonological target of social evaluation. In §5.2, Philadelphian participants were found to identify a PHL guise as distinctly more *accented* than NAS, using a Matched Guise paradigm. With the addition of the Magnitude Estimation results, I find that not only are Philadelphians at least implicitly aware of their sociolinguistic environment, but also that their explicit evaluations of "well pronouncedness" fall out from a structural rather than phonetic evaluation. Young Philadelphians exhibit the operation of two evaluation standards

in their responses: the new NAS system is rated positively overall while the older PHL system tokens receive the expected downgrading of the tense forms. The responses from older Philadelphians provide what is potentially the biggest surprise: here, we find that the target of listener evaluation may be the abstract conditioning factors, rather than the phonetic output of those conditioning factors. These findings reveal two important points: First, it suggests that abstract phonological structure may act as the target of social evaluation. Secondly, it reinforces the importance of diachronic work: what appeared synchronically to be participants rating the phonetic output of an allophone is revealed diachronically to be a potential case of participants rating the underlying phonological structure rather than the phonetic realization of that structure.