RESEARCH REPORT

Real-time speaker evaluation: How useful is it, and what does it measure?

Martha Austen                                            Kathryn Campbell-Kibler
The Ohio State University                                    The Ohio State University

Indexical associations are a crucial construct in third-wave variationist work, but little is understood about how perceivers incorporate indexical information over the course of sociolinguistic perception. In classic speaker evaluation, participants listen to a stimulus and report evaluations after listening, limiting our access to the moment-to-moment process of updating social percepts. Studies developing in-the-moment tools have combined methods development with substantive theoretical questions, hindering assessment. We test a continuous evaluation tool using a gestalt style shift and the English variables (ING) and like. The tool captures the expected reactions but has poor time granularity and very high variability. Divergence between slider responses and after-the-fact ratings suggests that the tasks may depend on a different mix of processes, underlining the multiplicity of sociolinguistic cognition processes.*

Keywords: sociolinguistic cognition, sociolinguistic perception, speaker evaluation, methods, sociolinguistic monitor, (ING), vernacular like

1. INTRODUCTION. Despite substantial evidence that speakers can invoke and understand complex indexical cues through sociolinguistic variation (e.g. Zhang 2005, Zimmerman 2013), little is understood about the cognitive processes through which individual sociolinguistic features are synthesized by perceivers into a percept. The primary tool theorized in variation has been the sociolinguistic monitor (Labov 1993, Labov et al. 2011), which manages self-regulation of one’s own speech as well as perceptions of others.

Campbell-Kibler (2016) argued that the sociolinguistic monitor is unrealistically broad in covering both self-regulation and the perception of others and further that variationists do not need specialized modules for sociolinguistic cognition in order to account for sociolinguistic behavior, beyond independently needed constructs already theorized for social and linguistic processing. For our current purposes, the most relevant such construct is the person-perception system. Social psychology research has suggested that humans maintain specialized structures for storing information about other people (Young & Bruce 2011) and that some processes create and/or update such impressions, while others draw on them to perform tasks like evaluation (Belmore 1987). While work in person perception has primarily focused on the perception of facial photographs and lists of traits or behaviors, sociolinguistic cues can easily be included in these models. Many open questions remain, however, about how a given token of a linguistic feature is incorporated into an impression.

These open questions have prompted recent work exploring in-the-moment rating tasks, in which listener responses are collected during—rather than after—the speech stimulus (Hesson & Shellgren 2015, Montgomery & Moore 2018). While this work has reported interesting and interpretable results, we cannot know how much to trust these tools without first establishing the accuracy and precision of in-the-moment results.

* We’d like to thank our research participants for their time and research assistants Nick Bednar, Shasteny Cabrera Roldan, Ben Elsbrock, Taylor Gifford, Yvon Gonzalez-Bergner, Jubileen Kombe, Hannah Mackenzie, and Navya Nalla for their help. Comments from Lynn Clark, Lauren Hall-Lew, and an anonymous referee substantially improved this work, and we are grateful to them as well. This research was supported by the National Science Foundation, grant 1655014. Supplementary materials for this research report can be accessed at http://muse.jhu.edu/resolve/150.

Printed with the permission of Martha Austen & Kathryn Campbell-Kibler. © 2022.
Across three experiments, we test a slider bar tool, with participants instructed to move the bar as they listen to reflect their ongoing impressions of a talker. Experiment 1 uses global stylistic shifts (disfluent → fluent → disfluent) to establish the task’s efficacy and to estimate response times. Experiments 2 and 3 test its responsiveness to well-studied linguistic features, namely the English variables (ING) and like. For each experiment, we examine patterns of listener response across the entire stimulus and directly following each shift or feature. We also analyze the relationship between in-the-moment responses and after-the-fact global ratings. To the degree that after-the-fact ratings match in-the-moment ratings from late in the stimulus, the accuracy of in-the-moment ratings as a ‘current state of the percept’ tool is supported. To the degree that they do not, the tools may draw on different processes or differently weighted mixes of the same processes.

Our results confirm that the slider task captures listener reactions to both stylistic shifts and individual features. However, data variability is very large, and response times are slow and variable (four to seven seconds post-cue). After-the-fact ratings show strong correlations with mean in-the-moment ratings, with smaller contributions from end-of-evaluation ratings in two of the three cases, suggesting that the in-the-moment ratings represent reactions to recent input more than an evolving global impression of a speaker.

2. Background. Labov (1993) proposed that a key cognitive tool for sociolinguistic variation is the sociolinguistic monitor, a module separate from the linguistic grammar, which socially evaluates the speech of others, in addition to evaluating and correcting the output of the speaker’s own grammar. Campbell-Kibler (2016) argued that the sociolinguistic monitor is unnecessary and that sociolinguistic behavior could be captured with already theorized cognitive structures, namely: a linguistic grammar with associative links to nonlinguistic structures, including individuals, groups, and situational contexts (Johnson 2006, Sumner et al. 2014); a person-perception system without direct access to the grammar but with a more limited ability to observe linguistic behavior in the self and others (Young & Bruce 2011); a self-regulation system (Wagner & Heatherton 2015) able to observe linguistic output but not to access the grammar directly; and the broad systems underlying functions like general memory and executive function (Diamond 2013), needed for, among other things, metalinguistic beliefs and verbal reflection on sociolinguistic meanings.

Sociolinguistic perception may be hypothesized to be the aspects of person perception that are informed by the observation of linguistic behaviors. The literature on person perception shows strong evidence that individuals have specialized systems for managing information about other people, which prioritize character traits (Newman & Uleman 1993) and allow existing perceptions to influence the interpretation of new information (Asch 1946).

We know very little about the time course of linguistic contributions to this system, including how quickly information from a new linguistic feature may be incorporated into a percept and whether potentially relevant associations may be briefly available before being eliminated by a social-reasoning process, analogous to associations found during the speech-perception process (e.g. Janse & Quené 2004). In the current study, we attempt to clarify this time course and develop tools that may offer future insight.

A handful of recent studies have explored the moment-to-moment development of sociolinguistic perception by manipulating token distributions of a single variable. Labov et al. (2011) exposed listeners to multiple recordings of the same performance,
differing in the proportion of -in and -ing tokens. They reported a logarithmic pattern of perceptions in which larger effects of (ING) were seen between guises with small numbers of -in tokens. Wagner and Hesson (2014) replicated this pattern to some extent, using similar methods and the same stimuli, but Levon and Fox (2014), using similar methods but new stimuli, found no effect for (ING) at all for UK listeners and a linear response pattern for TH-fronting. In all three cases, listeners heard all guises side by side, and thus were likely to have a high degree of explicit awareness of the linguistic feature under study. Examining aspiration of /s/ for Puerto Rican Spanish, García et al. (2016) likewise found no support for a logarithmic pattern.

Other recent studies have probed the time course of sociolinguistic evaluation by developing tools for capturing responses as the stimulus unfolds. Hesson and Shellgren (2015), using a two-dimensional schema on an electronic drawing pad, found lower friendly and intelligent ratings immediately after a token of discourse marker like. The friendliness effect dissipated, while the effect on intelligence persisted and even grew over the course of a stimulus, suggesting that after-the-fact ratings may fail to capture some sociolinguistic percepts.

Watson and Clark (2014), using a slider bar along a single dimension, found that social differences between five UK and Irish accents persisted throughout the stimuli. Watson and Clark also reported multiple sources of variability and noise in the task, including response latency; slider movements began roughly three seconds after stimulus onset for a Liverpool accent but six or seven seconds for a Cardiff accent.

Watson and Clark (2013) used the same task to examine reactions to the nurse-square merger in the northwest of England, comparing reactions to guises with the merger realized in either the Liverpool or St. Helens fashion. While they found no difference between guises in ratings at the endpoint of the stimulus, analysis of the slider movements suggested that listeners reacted in the moment to some, but not all, of the nurse-square tokens. Based on these results, they suggest that the extent to which a given token affects listeners’ evaluations depends on the ‘microlinguistic context’, namely, the other linguistic features that were heard immediately prior to that token.

Exploring a different approach, Montgomery and Moore (2018) asked participants to listen to an audio clip and click whenever they heard something that made them ‘wonder where he [the speaker] is from (or confirms where you already think he is from)’. Participants were then given the chance to revisit each mouse click and explain their selection. Responses differed depending on whether the talker was talking about locally or nonlocally marked content, suggesting that the indexical contribution of linguistic cues depends on other available information about the talker. Taken together, these studies showcase both the potential of real-time studies to answer new questions about sociolinguistic perception and the lack of methodological consensus on how best to conduct such studies.

Outside of linguistics, research on how people perceive experiences suggests a difference between in-the-moment and after-the-fact evaluations, although it is not clear how similar we might expect evaluations of people vs. experiences to be. Fredrickson and Kahneman (1993) showed that holistic post-hoc evaluations of positive or negative experiences (e.g. holding one’s hand in a bucket of ice water) are predicted by an average of evaluations at the most intense point of the experience and at the last point. Duration appears to play so little a role in the assessment that most participants prefer longer unpleasant experiences if they include even a slight decrease in discomfort at the end of the session (Kahneman 1999). This suggests that one of the first steps in pursu-
ing continuous speaker evaluations is to probe the relationship between in-the-moment and end-of-stimulus evaluations, testing the assumption found throughout sociolinguistic work that the final rating via the continuous metric is equivalent to a holistic post-stimulus evaluation (Hesson & Shellgren 2015, Jones 2016, Labov et al. 2011). By comparing in-the-moment and after-the-fact ratings, we might also be able to clarify whether in-the-moment and after-the-fact tasks both tap into similar person-perception processes. An alternative possibility is that the slider task provokes a more targeted behavior, similar to phoneme monitoring, in which participants monitor the linguistic material for cues they associate with the evaluative task. In that scenario, the task would reveal associations between the stimulus and the evaluative dimension, but not necessarily anything about the process of updating a person percept.

In this study, we test a continuous evaluation tool similar to that of Watson and Clark. In experiment 1, we examine listener reactions to sudden, large-scale changes in the talker’s level of fluency. In experiment 2, we examine a well-studied sociolinguistic variable, (ING), and in experiment 3, we follow Hesson and Shellgren (2015) in looking at reactions to discourse marker like. In each case we ask whether the slider tool reveals the expected responses, which quantitative analyses best capture such effects, and the time window in which such reactions occur. Additionally, we examine relationships between the real-time responses and after-the-fact ratings collected from the same participants.

3. Experiment 1: a gestalt feature. Experiment 1 tested how listeners used the continuous slider tool in reaction to a single stimulus that started out relatively disfluent, became more fluent, and then became disfluent again. This stimulus was constructed such that the changes in levels of fluency were sudden and obvious, so that listeners’ reactions would be large and easily interpretable. This was designed to develop a basic understanding of how listeners used the slider tool, with a focus on how quickly they reacted to changes in the stimulus.

3.1. Participants. Participants were visitors to the Center of Science and Industry (COSI), a science museum in Columbus, Ohio. A total of fifty participants were recruited. Data from participants who reported hearing problems (n = 2) were discarded. The remaining forty-eight participants ranged in age from eighteen to seventy (median age: 25.5). Nineteen participants identified as female and twenty-nine as male. Forty-one of the participants were white, three were Black, two Latinx, one Asian, and one mixed race; 65% reported living in an Ohio zip code.

3.2. Stimulus. A white, female native English speaker from the Midwest was recorded in a quiet room using a Roland N225 R-05 flash recorder. She read a passage about how to make a grilled cheese sandwich, framed as a podcast for learners of English as a second language. To create less-fluent-sounding speech, the talker was first presented with the passage printed in a hard-to-read font on a busy background, and was told that we were interested in recording a version of the passage containing many speech errors. After the initial reading on the hard-to-read font, the speaker was asked to read the passage from a normally presented text, first in her ‘normal’ voice, then in an enthusiastic voice, and finally in a bored voice.

The resulting four recordings (disfluent, normal, enthusiastic, and bored) were divided into three portions of roughly equal length and played to participants who were asked for three adjectives to describe the speaker. From these adjectives, we selected ‘articulate’ for the study described here. In the second pilot, participants were asked to
rate how articulate the speaker sounded in each stimulus piece. The disfluent excerpts were rated as least articulate (mean: 44 points out of 100), and the enthusiastic chunks as most articulate (mean: 83 points). Based on this, we created a spliced **disfluent → fluent → disfluent** stimulus, sixty-five seconds in length, for the main study.

This procedure was used to create several different stimuli with similar global stylistic shifts, across different topics, including a radio announcer discussing classical music and a reading of a children’s story. Versions of the stimuli were also created using different orderings of the fluent vs. disfluent portions. We present detailed results here of a single example for ease of explication, but the patterns found, including the time windows, were similar across multiple stimuli.

### 3.3. Procedure

We piloted a number of tasks, including different input devices (mouse, arrow keys, iPad slider), with and without a visual scale, and varying instructions (asking for ‘in the moment’ or ‘overall’ impressions). Each of these variations on the task was tested on five to twenty participants and produced the same basic patterns of response. We elected to use an iPad with a visible slider for the main experiment because it was easy to explain, portable, and rarely needed clarification or participant redirection.

After recruitment and consent procedures, main study participants were asked to listen to the stimulus and continuously indicate how articulate the speaker sounded using a slider with the endpoints labeled ‘not at all articulate’ and ‘very articulate’ (see Figure 1). The slider always started in the middle position.

![Figure 1. The slider used to collect participant reactions.](image)

After they had finished listening to the stimulus and had submitted their responses, participants were asked by the experimenter to give their overall impression of the speaker on a scale of 1 to 10. Finally, participants were asked for their age, gender, race/ethnicity, and zip code.

Multiple analytic approaches were tested, in all three experiments, in order to capture patterns of change while avoiding problems caused by autocorrelation. These approaches are discussed below.

### 3.4. Results

Figure 2 shows mean participant ratings across the time course of the stimulus, with timepoints rounded to the nearest 250 milliseconds. These ratings show the expected pattern: participants initially rate the talker as sounding less articulate during the initial ‘disfluent’ segment, then as more articulate once the middle ‘fluent’ segment of the audio begins, and finally as less articulate when the final ‘disfluent’ segment of the audio begins.

To analyze response lag, we began by qualitatively examining the patterns from individual participants, shown in Figure 3. There is substantial individual variability, but a strong majority showed significant periods of waiting between rapid movement (see also Watson & Clark 2014). Based on this pattern, we attempted to identify clusters where many participants started to move the slider in the same direction.
For each participant, we represented each movement by its starting point and total size of change, shown in Figure 4. The dashed vertical lines at eighteen and forty-three seconds show the guise shifts. If large changes in speech style prompt widely shared reactions, we would predict clusters of changes in response to these two points and at the beginning of the stimulus. In the first ten seconds, twenty-three participants (55% of participants with movement then) moved the slider in the expected direction, beginning at an average of 4.5 seconds (SD = 1.8) after onset. In the ten seconds after the first style shift, twenty-six participants moved the slider up (74% of those with movement), at an average of 3.36 seconds after the guise shift (SD = 2.0). In the ten seconds after the second shift, sixteen participants (64%) moved the slider down, at an average of 5.9 seconds after the splice (SD = 3.0). Since individual speech segments are typically no longer than 300 milliseconds, this wide range of mean response times (4.2–5.9 seconds) suggests that this slider method is ill-suited for determining which linguistic cues listeners are responding to post hoc. There was also no obvious way to divide participants into ‘slow’ vs. ‘fast’ reactors; only twelve participants (25% of total participants) showed a move in the expected direction within ten seconds of both the first and the second transitions.
Next, we tested two analytical approaches to formally test for clusters or change patterns. First, we compared rolling windows of responses through the course of the trial, testing windows of 1, 1.5, 2, 2.5, and 3 seconds. The stimulus duration was divided into a series of windows of the tested width, and the average rating was calculated within each window for each participant. A difference metric was created by subtracting the value of the preceding window for that subject. A Bonferroni-corrected one-sample t-test was applied to the difference metrics for each window, resulting in a graph, as shown in Figure 5. Using this approach, shifts in response to the stylistic shifts were visible, although the wide range in response times remained. The clusters of perceptual changes in the first ten seconds of each stimulus segment suggest that participants are attending to the stylistic shifts but may be prompted by different individual cues. Using such a response pattern to identify triggers post hoc would be challenging. Results were similar across the different window sizes we tested.

Second, we performed a changepoint analysis. Watson and Clark (2014) applied changepoint analysis to individual response tracks, a technique similar to our first de-
scriptive approach. We experimented with applying changepoint analysis to the mean response track across all participants, in hopes of identifying common shift points. Due to the curved nature of the mean track, this analysis produced a very large number of identified change points that did not obviously map onto any specific linguistic cues.

From these analyses, we conclude that it is not possible to extract a precise response window for this type of data. Despite agreeing on the relative articulateness of the three guise segments, participants exhibit a wide range of time lags between the guise shift and their corresponding slider movement. It is unclear how much of this individual time variability is due to participants reacting at different speeds to the same linguistic cue and how much to different participants reacting to different cues. In experiments 2 and 3, we address this issue by using a matched-guise paradigm in which only one linguistic variable is manipulated.

3.5. After-the-fact ratings. To explore the relationship between in-the-moment and after-the-fact evaluations, we built a model predicting participants’ after-the-fact ratings as a function of various characteristics of their in-the-moment ratings, specifically the maximum, minimum, mean, median, end-of-stimulus, and end-of-evaluation values. We distinguish the final two because the trial was self-paced and some participants continued moving the slider after the stimulus ended.

We calculated these in-the-moment ratings beginning five seconds after the participant first began moving the slider, to ensure that all of the input values reflected assessment made by the participants rather than the slider’s starting position at 50 points. This was particularly crucial in the case of maximum and minimum ratings, as some participants kept the slider either above or below 50 points throughout the task. The minimum and maximum ratings distributions were skewed (right and left, respectively), so for the purposes of model building these were log-transformed, with the maximum rating subtracted from 101 first and minimum rating added to 1, to avoid 0 values. Mean ratings during each of the three chunks were also tested as predictors.

A linear regression model of after-the-fact ratings was built using a step-up approach, using log-likelihood comparisons with an alpha of 0.05 and AIC as selection criteria. In addition to the in-the-moment statistics described above, we considered participants’ age, gender, and ethnicity as potential predictors. Interactions were considered to be predictors only if the relevant main effects had been included in the model. Because different in-the-moment ratings were correlated with one another, we checked the variance inflation factor (VIF) at each step of the model to ensure that the included measures were not too tightly correlated to yield a reasonable model, using a cutoff of 2.5 as unacceptable high and checking each effect with a VIF greater than 2 in a simpler model to verify its continued significance.

The resulting model contained only the mean rating as a significant predictor of the after-the-fact rating, with higher mean ratings yielding higher after-the-fact ratings ($\beta = 0.057, SE = 0.009, t = 6.074, p < 0.001$). End-of-stimulus and end-of-evaluation in-the-moment ratings were not significant predictors once mean ratings were controlled for, distinguishing sociolinguistic evaluation from positive/negative experiences as described in Kahneman et al. 1993. No demographic factors were significant predictors.

At the suggestion of an anonymous referee, we built a new model, testing the in-the-moment means for each of the three stimulus segments instead of a single global mean. This model yielded only the mean for the third and final segment as a predictor for the after-the-fact rating ($\beta = 0.061, SE = 0.010, t = 5.841, p < 0.001$) and had an improved model fit (adjusted $R^2 = 0.4468$ vs. 0.4238).
3.6. Summary. Experiment 1 showed that the slider was successfully able to capture participants’ in-the-moment reactions to changes in overall levels of fluency, although time granularity was poor. Additionally, we found that after-the-fact ratings were best predicted by mean in-the-moment ratings, particularly the mean of the final third of the stimulus, and that a participant’s final end-of-stimulus rating did not predict their after-the-fact rating.

4. Experiment 2: (ING). Given the poor time granularity seen in experiment 1, we do not recommend using a slider tool as a method of discovering post hoc which linguistic cues listeners are attending to. However, the slider may be useful within a matched-guise paradigm, where the researcher wishes to examine real-time reactions to a specific linguistic feature. In these cases, the variable itself provides a specific time point on which to base an analysis. In experiment 2, we test this approach with a well-studied variable, (ING).

4.1. Background. We selected (ING) as the variable for this experiment because it is well documented and shows robust effects in both production and perception across studies (for an overview, see Hazen 2008). (ING) is typically described as the alternation between [ɪŋ] (orthographically represented as <ing>) and [ɪn] (orthographically represented as <in’>) in fully unstressed syllables, for example, walking vs. walkin’. In this research report, we use (ING) to refer to the variable, ING to refer to the velar variant, and IN to the alveolar variant.

In speech production in the US, high use of IN correlates with more casual speech styles (Fischer 1958, Hazen 2008, Labov 1966), and IN is used more often by speakers with lower socioeconomic status (Hazen 2008, Labov 1966:270–75) and by Black (Labov 1966, but see Hazen 2008), male (Fischer 1958, Labov 1966), and Southern (Hazen 2008, Labov 2001:90, Wolfram & Christian 1972:62) speakers. Perceptual studies examining (ING) have documented associations between IN and lowered perceptions of intelligence and education (Campbell-Kibler 2009, Tamminga 2017) and professionalism of newscasters (Labov et al. 2011), and higher ratings of accentedness of Southern talkers (Campbell-Kibler 2007) and casualness (Tamminga 2017).

4.2. Participants. Participants were initially recruited from the COSI science museum, as in experiment 1. After approximately a hundred participants were run, it became clear that any effect of (ING) would be small enough to require more participants than obtainable at COSI. We therefore recruited additional participants from the crowdsourcing site Prolific, paying $1.20 each to achieve an estimated $15/hour. Participants were limited to those listing their nationalities as USA and native language as English; we further excluded data from those who reported hearing disorders (n = 38). In total, 885 participants with useable data were recruited from Prolific and 314 from COSI, yielding 1,199 participants total. COSI participants were ages eighteen to seventy-one (median: 25), 89% white, 61% female, and 78% living in an Ohio zip code. Prolific participants were ages eighteen to seventy-six (median: 29), 74% white, 49% female, and were scattered across the United States.

4.3. Stimuli. Stimuli were taken from Campbell-Kibler 2007, for which study they had been digitally manipulated from spontaneous interview speech to create matched IN and ING guises. We selected four clips of a white man from North Carolina in his early twenties with a perceptible Southern accent; in Campbell-Kibler 2007, listeners judged the ING guises as more educated than the IN guises.

Four matched pairs of ten-to-twenty-second clips, on different topics, were concatenated to create a seventy-one-second stimulus containing a total of eleven (ING) tokens,
with three beeps marking the boundaries between each clip. Six guises were prepared: all ING, all IN, and two pairs of alternating IN-ING-IN-ING and ING-IN-ING-IN, one pair with the topics in reverse order. Each guise was heard by 200 participants, except for the all-ING guise with 199 participants.

4.4. Procedure. The same real-time slider task as in experiment 1 was used with the question ‘How educated does this person sound?’ and scale ends labeled ‘not at all educated’ and ‘very educated’. After the stimulus ended, participants were asked to rate on a 1–10 scale how educated the speaker sounded, verbally for COSI participants and using a drop-down menu for Prolific participants. Finally, all participants were asked for their age, gender, race/ethnicity, and zip code.

4.5. Results. Figure 6 compares mean real-time ratings for the all-ING and all-IN stimuli. At the large scale, these profiles offer little support for an effect of (ING) on slider ratings. Although mean ratings for the all-ING stimuli are higher than for all-IN, as expected, this difference begins before the first (ING) token is heard and does not increase as the stimulus progresses. Furthermore, the ratings contours never differ significantly from one another. Their shared contour seems to primarily reflect an initial reaction to the talker’s Southern accent, then reactions to the content of each of the four clips in the stimuli. The first and fourth clips concern recreational activities, while the second and third take more professional stances. The alternating ING-IN-ING-IN and IN-ING-IN-ING stimuli with the same order of clips (not pictured) also show the same overall ratings contour, with no obvious effects of (ING).

Thus, these data do not support a large cumulative effect of (ING) on slider ratings, but they cannot tell us whether individual tokens had a small or short-term impact on responses directly following each token. Because the effects of overall message content are so large, it is necessary to ‘subtract out’ these content effects in order to test this more fine-grained question. In the next section, we present two related methods for doing so.

4.6. Model 1: token-to-token slopes. To analyze the effects of (ING) while controlling for message content, we compared the slopes of movement after each token between the (ING) guises. For example, the first token of (ING) occurred early in the

![Figure 6. Mean responses to all-ING versus all-IN stimuli. Dashed vertical lines indicate the start of each clip, and white dots indicate the timing of (ING) tokens.](image-url)
stimulus when many participants were already moving the slider downward, likely in response to the talker’s Southern accent. To find the effect of a given (ING) token independent of this shared trajectory, we compared the mean slope of the ING guise minus the mean slope of the IN guise—predicting a positive difference if ING had the expected effect.

For our first analysis, we took the slope for each participant for each token of (ING) from the offset of that token until the onset of the next token, or for the final token until the end of the stimulus. Because the (ING) tokens were not evenly spaced throughout the stimulus, this window is variable, ranging from 0.75 to 15.25 seconds (mean: 5.91). The shorter windows may be too short for reactions to appear, particularly given the lag time found in experiment 1. Below we discuss a different approach with a constant window size of five seconds, to address this issue.

We then built a linear model predicting these token-to-token slopes, testing as predictors token guise (ING, IN), preceding token guise, stimulus type (all ING/IN, alternating ING first, alternating IN first), token number (integer count), and length of the slope window, as well as the interactions of each of these factors with token guise. We also tested two possible experimental influences: subject pool (COSI, Prolific) and experimenter (experimenter 1, experimenter 2, Prolific), along with their interactions with token guise. Likewise, we tested listener demographics of age (integer), ethnicity (Asian, Black, Latinx, mixed/other/unspecified, white), and gender (male, female, nonbinary/unspecified) and their interactions with token guise. A random intercept for participant and a random slope for guise over participant were also included, as well as a random intercept for stimulus, which combined the particular subpart of the stimulus that the token appeared in with the larger overall stimulus.

The best-fit model, shown in Table 1, contains main effects for guise and stimulus type, plus an interaction between guise and stimulus type. The interaction suggests that (ING) has an effect only in the alternating IN-first stimuli, with no effect in the all-IN/all-ING guises or the alternating ING-first stimuli (see Figure 7).

4.7. Model 2: post-token slopes. As noted above, the token-to-token slope meant that the slopes were calculated over variable timespans, which has the potential to muddy the results by mixing windows of different sizes. We thus tested a second model examining the five seconds following each token, based on the findings of experiment 1, where mean response times ranged from 4.2 to 5.9 seconds. This traded the disadvantage of variable windows for the disadvantage of having some tokens’ windows include another token, leading to overlap. This possibility was included as an additional predictor in our models (no additional token, IN, or ING), replacing window length. All other aspects of the model-building process remained the same.

The resulting model contained effects of guise, subject pool, and listener ethnicity (Table 2). Listeners rated the talker as more educated following an ING token, as ex-
expected. Unlike in our token-to-token model, guise did not interact with stimulus type.¹ Prolific participants tended to move the slider upward more; this effect did not interact with guise. As compared to the average across other ethnicities, white participants had more negative slopes (harsher judgments); this effect also did not interact with guise.

4.8. Debriefing Comments. When participants were asked what they thought the purpose of the experiment was, no participants mentioned (ING). After debriefing, only three of the 145 participants who left post-debriefing comments mentioned that they had noticed (ING). More common were comments that they had focused on the content of what the speaker was saying, his Southern accent, or occasionally on his ‘stuttering’ or use of filler words.

4.9. After-the-Fact Ratings. Our final step, as in experiment 1, was to examine the after-the-fact ratings and their relationship to the slider data. When we tested only the effect of stimulus, without taking in-the-moment slider ratings into account, there were no significant differences between (ING) guises (all ING, all IN, or mixed). However, the order of excerpts of the stimulus did have a significant effect ($\beta = 0.123$, $p = 0.017$), shown in Figure 8: the lower-rated order (labeled ‘forward’ in our model) was

¹ A model containing an interaction of guise by stimulus type had a lower (= better) AIC than our final model and a significant coefficient for this interaction, but failed to significantly improve on a model without the interaction ($p = 0.062$).
that in which the final clip heard was the most casual and regionally marked clip, which prompted the lowest intelligent and educated ratings in Campbell-Kibler 2007. This finding suggests that later content has a larger effect on after-the-fact ratings.

We next built a model predicting after-the-fact ratings from in-the-moment ratings, following the same procedure as in experiment 1. As in experiment 1, in-the-moment ratings characteristics were calculated beginning five seconds after a participant first began moving the slider. This excluded twenty-eight participants who moved the slider only once. (When these participants were excluded, the effects reported in the previous paragraph were still significant.) In addition to the in-the-moment summary characteristics for each participant (mean, median, maximum, minimum, end-of-stimulus, and end-of-evaluation) and participant demographic characteristics (age, ethnicity, and gender), we also tested as predictors subject pool (Prolific or COSI), guise (ING, IN, mixed), and stimulus order (forward, reverse). Maximum and minimum slider ratings were log-transformed, with the maximum rating first subtracted from 101.

The resulting model indicated that both mean and end-of-evaluation ratings positively correlated with after-the-fact ratings (Table 3). In addition, a higher maximum rating predicted a higher after-the-fact rating for Prolific participants, but not for COSI participants.

| EFFECT                                                                 | EST  | SE   | t-VALUE | Pr(>|t|)   |
|-----------------------------------------------------------------------|------|------|---------|------------|
| (intercept)                                                           | 3.256| 0.328| 9.933   | < 0.001 ***|
| mean rating                                                           | 0.048| 0.003| 14.248  | < 0.001 ***|
| end-of-evaluation rating                                              | 0.018| 0.002| 8.508   | < 0.001 ***|
| inverse max. rating (log)                                             | −0.026| 0.075| −0.340  | 0.734      |
| subject pool=Prolific                                                | 0.781| 0.306| 2.550   | 0.011 *    |
| GUESE (BASELINE = MIXED)                                              |      |      |         |            |
| guise=ING                                                             | −0.104| 0.231| −0.448  | 0.654      |
| guise=IN                                                              | −0.936| 0.224| −4.185  | < 0.001 ***|
| order = reverse                                                       | 0.013| 0.038| 0.351   | 0.726      |
| inverse max. rating (log) × subject pool=Prolific                    | −0.310| 0.082| −3.758  | < 0.001 ***|
| SUBJECT POOL × GUESE                                                 |      |      |         |            |
| subject pool=Prolific × guise=ING                                     | 0.073| 0.245| 0.297   | 0.767      |
| subject pool=Prolific × guise=IN                                      | 0.820| 0.238| 3.445   | 0.001 **   |

Table 3. Best-fit model of after-the-fact ratings for (ING). Note: a higher estimate = more educated.
Once the characteristics of a participant’s in-the-moment ratings were included in the model, the order of stimulus a participant heard was not a significant predictor of their after-the-fact rating; however, we still included order in the model to more accurately assess the effects of ING guise, since guise was partially correlated with order. ING guise interacted with subject pool: for COSI participants, the all-IN guise predicted significantly lower ratings than the all-ING or mixed guises, but for Prolific participants, guise had no (or a very small) impact.

4.10. Summary and discussion. In experiment 2, (ING) most clearly showed the expected effect for alternating IN-first guises, in spite of the fact that the single-variant stimuli did show the expected effect in participants’ after-the-fact ratings, when in-the-moment ratings were also included in the model. It is unclear whether this is because (ING) had no effect on participants’ in-the-moments ratings for the single-variant stimuli, or because we lacked enough statistical power to see the effect. If it is the latter, the necessary sample sizes are likely to be prohibitively large.

In examining the relationship between in-the-moment and after-the-fact ratings, we found that mean in-the-moment ratings most strongly contributed to participants’ after-the-fact ratings, as in experiment 1. However, unlike experiment 1, end-of-evaluation slider ratings (where the slider was left when the participant stopped the task) were also a predictor, roughly half as strong as the mean slider ratings. The interaction between (ING) guise and subject pool is particularly interesting, since it suggests that for the COSI participants (whose after-the-fact ratings were given verbally), the contribution of (ING) to their after-the-fact ratings is not fully captured by their in-the-moment ratings. When controlling for mean and end-of-evaluation ratings, COSI participants show an overall effect of (ING). This suggests that a gestalt evaluation may capture responses not captured in the slider task, but only when the gestalt task is distinct enough from the in-the-moment task.

5. Experiment 3: (like). Experiment 2 offered limited evidence that the slider tool was able to capture the in-the-moment effects of (ING). (ING) showed the expected effect in alternating stimuli, but not in single-variant stimuli—despite having the expected effect in the after-the-fact ratings of these stimuli. One possible reason that (ING) failed to show the desired effect in single-variant stimuli was that the effect was simply too small to see, given the noisiness of timing data. Therefore, in experiment 3 we selected a feature—discourse marker like—that we predicted would have a larger effect. We selected like because it is subject to overt comment (as is (ING), but with perhaps stronger stigma) and as a lexical item, it may be easier for listeners to note in the moment. If like provides clearer or more reliable effects than (ING), this may indicate where slider-based techniques might be most useful.

5.1. Background. D’Arcy (2007) distinguishes between different kinds of ‘vernacular’ like in English: quotatives (she was like … ), approximate adverbs (it was like three dollars), discourse markers, which occur clause-initially, and discourse particles, which occur clause-internally. She notes that these different forms of like correlate differently with different social groups in production—for example, women are more likely than men to use quotative like, whereas men are more likely than women to use discourse particle like.

In popular language ideologies, like is associated with adolescents, with women, particularly the characterological figure of the Valley Girl, and with American rather than British English (D’Arcy 2007). These beliefs seem to hold equally no matter what kind of like people are talking about, although this has not been tested empirically. In speaker-evaluation studies, this has translated into ratings linking stigmatized like to-
kens to younger (Buchstaller 2006, Dailey O’Cain 2000) and less educated or intelligent ratings (Dailey O’Cain 2000), but sometimes also to higher friendliness or attractiveness perceptions (Dailey O’Cain 2000). Some data suggest that judgments of like may depend on the age, or perhaps cohort, of the perceiver (Dailey O’Cain 2000) and on the specific like function used (Maddeaux & Dinkin 2017). Drager (2006, 2009) further suggests that the phonetic realization of like may complicate its indexical meanings in some communities.

Discourse like is one of the few features to have been documented as having an effect on listener perceptions of intelligence in a real-time study. Hesson and Shellgren (2015) found that a guise containing like was rated as less intelligent via in-the-moment responses than the like-free guise; the like guise was also initially rated as less friendly immediately after like, but this difference between guises faded by the end of the stimulus.

5.2. Participants. As in experiment 2, participants were recruited both from the COSI science museum (n = 46) and online through Prolific (n = 353). Data from participants who reported hearing disorders (n = 5) or from whom an after-the-fact rating was not collected (n = 1) were discarded, leaving 393 participants. The remaining COSI participants (n = 44) were ages eighteen to sixty-eight (median = 33), 73% white, 11% Black, 2% Latinx, 5% multiracial, 2% other race, 73% female, 26% male, and 86% living in an Ohio zip code. The remaining Prolific participants (n = 349) were ages eighteen to seventy-six (median = 30), 77% white, 7% Black, 7% Asian, 3% Latinx, 1% Native American, 5% multiracial, 55% female, 44% male, 1% nonbinary, and lived in various places across the US.

5.3. Stimuli. For experiment 3, we extracted a fifty-four-second sample containing ten instances of like from speaker #10 of the Buckeye corpus, a corpus of sociolinguistic interviews with speakers in Columbus, Ohio (Pitt et al. 2005); this speaker was a white man forty or older at the time of recording. Based on D’Arcy’s (2007) classification of the types of like, this excerpt contains six instances of discourse particle like, three instances of approximative adverb like, and one instance of discourse marker like. To create a matched stimulus without like, we replaced each token with a silence of the same duration as the like token. While the indexical meanings of contrasting like against a pause vs. against a full deletion are complex, the practical considerations of keeping the two guises time-aligned were taken to outweigh those concerns for the current study.

5.4. Procedure. Using the same input slider as in experiments 1 and 2, participants were asked to rate how ‘intelligent’ the talker sounded, following Hesson & Shellgren 2015. As before, participants were then asked to rate how intelligent that speaker sounded on a 1–10 scale, using the same procedure as in experiment 2. Finally, participants were asked for their age, gender, race/ethnicity, and zip code.

5.5. Results. Figure 9 shows that, as expected, participants hearing the like guise rated the talker as less intelligent as the stimulus progressed, compared to participants in the no-like condition. In contrast to experiment 2—where almost no participants mentioned noticing (ING)—many participants explicitly mentioned noticing like or ‘filler words’ when they were asked what they thought the experiment was about. To test whether this explicit noticing affected participants’ reactions to the stimulus, we coded for whether Prolific participants in the with-like condition referred to like or filler words in their post-task comments. These included responses to the question ‘What do you think this experiment was about?’, a general request for additional comments, and a post-debriefing request for comments. COSI participants’ metalinguistic comments were not recorded in as much detail and were not included in this analysis. Figure 10
shows the results of this analysis, suggesting that much of the apparent difference between guises was driven by participants who consciously noticed like.

5.6. Model 1: Token-to-Token Slopes. To assess the effects of individual tokens of like, we fit two models using the same slope-modeling procedures as in experiment 2. First, we constructed a linear mixed-effects model predicting a participant’s change in ratings between two tokens of like, with the main predictors of interest being stimulus guise and comment, with the levels ‘+like, mentioned noticing’, ‘+like, no mention’, ‘+like, COSI participant’ (as these participants’ responses about whether they noticed like were not recorded), and ‘−like’.

The model was built using a step-up approach with a random intercept for token. A random intercept for participant was not included because it did not account for any variance in the data. Fixed effects tested were guise with metalinguistic commentary, participant age, race/ethnicity, gender, and subject pool (COSI or Prolific), as well as the interaction between each of the other factors and guise. We also tested interactions between guise and token number, between guise and length of time from current to the
The only predictor to emerge as significant was guise ($p = 0.031$), with *like* making the talker sound less intelligent, regardless of metalinguistic comment.

5.7. Model 2: Post-token slopes. The second model analyzed slopes from the offset of the token to five seconds later, using the same method and predictors as in the previous section, except slope window length. Two effects emerged as significant: participant age and an interaction between token number and guise with metalinguistic commentary (Table 5). For age, younger participants tended to rate the talker as less intelligent, regardless of guise. For the interaction, Prolific participants who heard the *like* guise but did not comment on it were similar to those hearing the no-*like* guise for tokens toward the beginning of the stimulus in showing little change immediately after, unlike those who did report noticing *like* and the less-debriefed COSI participants, who showed a negative slope (Figure 11). This pattern was tempered by a trial effect, with later tokens showing less of a negative slope and less of a difference across guises.

### Table 4. Best-fit model of *like* token-to-token slopes. Note: a higher estimate = more intelligent.

| EFFECT                              | EST  | SE   | df     | t-VALUE | Pr(>|t|) |
|-------------------------------------|------|------|--------|---------|---------|
| (intercept)                         | 0.206| 0.408| 11.291 | 0.505   | 0.624   |
| guise=[−like]                       | 0.578| 0.268| 3918.997| 2.160   | 0.031 * |

5.8. After-the-fact ratings. In after-the-fact ratings, participants showed no effect of guise ($p = 0.06$), though the descriptive pattern rated the no-*like* stimulus as more intelligent (means = 6.43 vs. 6.09). Prolific participants (Figure 12) who mentioned *like* gave the lowest ratings, differing significantly from no-*like* participants ($p = 0.004$ in model coefficients) and nonmentioners ($p = 0.044$), who did not differ from each other ($p = 0.103$). This finding suggests that *like* affects after-the-fact ratings only if a listener notices it in a way accessible to verbal introspection.

To assess the relationship between a participant’s in-the-moment ratings and their after-the-fact ratings, we built a model predicting after-the-fact ratings from characteristics of in-the-moment ratings (mean, median, minimum, maximum, end-of-stimulus, and end-of-evaluation). We used the same procedure as in experiment 2, except that because end-of-stimulus and end-of-evaluation ratings were more strongly correlated with mean slider ratings, we tested the residuals of models predicting each of these based on mean slider ratings as predictors. As in the previous experiments, only ratings starting next token, and between guise and *like* type (approximate adverb, discourse particle, or discourse marker).

### Table 5. Best-fit model of *like* responses, based on slope from token to five seconds after. Note: a higher estimate = more intelligent.

| EFFECT                              | EST  | SE   | df     | t-VALUE | Pr(>|t|) |
|-------------------------------------|------|------|--------|---------|---------|
| (intercept)                         | −0.916| 0.952| 15.446 | −0.962  | 0.351   |
| guise & mention LIKE (baseline = [−LIKE]) |      |      |        |         |         |
| [+like, ?mention (COSI)]           | −5.344| 1.369| 3913.002| −3.904  | < 0.001 ***|
| [+like, −mention]                  | −0.526| 0.647| 3913.002| −0.814  | 0.416   |
| [+like, +mention]                  | −3.314| 1.244| 3913.002| −2.664  | 0.008 **|
| token number                        | 0.227 | 0.139| 10.382  | 1.632   | 0.133   |
| participant age                     | 0.024 | 0.012| 3913.002| 2.023   | 0.043 * |
| guise & mention LIKE (baseline = [−LIKE]) × token number |      |      |        |         |         |
| [+like, ?mention (COSI)] × token number | 0.733 | 0.221| 3913.002| 3.324   | 0.001 **|
| [+like, −mention] × token number    | 0.009 | 0.104| 3913.002| 0.088   | 0.930   |
| [+like, +mention] × token number    | 0.263 | 0.200| 3913.002| 1.312   | 0.190   |
five seconds after a participant started moving the slider were included; thirty-seven participants were excluded because they moved the slider only once. (When these participants were excluded, the after-the-fact ratings difference reported above between participants who did vs. did not mention noticing like was no longer significant.) In addition, subject pool (COSI or Prolific), race/ethnicity, age, and gender were considered as predictors. Once best-fit models were built testing these predictors, guise was again tested as a potential additional factor, to see if any contributions from guise to after-the-fact ratings remained, once the in-the-moment correlations were accounted for. VIFs were checked throughout the process, with a cut-off threshold of 2.5.

The final model, shown in Table 6, shows interactions between subject pool with mean and residualized end-of-evaluation slider ratings. Residualized end-of-stimulus rating, added in the stepping-up process, was no longer significant in the final model. For both mean and end-of-evaluation ratings, Prolific participants had stronger effects than COSI participants. Indeed, COSI participants actually showed a slight negative re-
Relationship between their final slider position, relative to their mean slider responses, and their after-the-fact rating. Which guise a participant heard (and, if they heard like, whether they mentioned hearing it) was not a significant predictor once in-the-moment ratings were included in the model.

| EFFECT                                           | EST   | SE    | t-VALUE | Pr(>|t|) |
|-------------------------------------------------|-------|-------|---------|----------|
| (intercept)                                      | 4.040 | 0.420 | 9.629   | < 0.001 *** |
| mean rating                                      | 0.024 | 0.009 | 2.724   | 0.007 **  |
| subject pool=Prolific                           | 1.560 | 0.461 | 3.384   | 0.001 **  |
| end-of-evaluation rating (residualized)          | -0.054| 0.012 | -4.388  | < 0.001 *** |
| end-of-stimulus rating (residualized)            | 0.011 | 0.007 | 1.634   | 0.103     |
| mean rating × subject pool=Prolific             | 0.048 | 0.009 | 5.056   | < 0.001 *** |
| subject pool=Prolific × end-of-evaluation rating (residualized) | 0.095 | 0.013 | 7.089   | < 0.001 *** |

Table 6. Best-fit model of after-the-fact ratings for like.

5.9. Summary and discussion. In summary, we find that for both in-the-moment and after-the-fact ratings, like negatively affects perceptions of talker intelligence. This effect is biggest for listeners who explicitly comment on the use of like or filler words. We found little evidence for the trial effects observed in Labov et al. 2011. While the slopes of post-token response windows did trend upward, showing less negative responses over time, this was significant only for the COSI respondents hearing like. All other categories, across guises, showed a similar slight upward trend, suggesting that the effect is unrelated to like.

The after-the-fact results again show the strongest correlation with mean in-the-moment ratings, with weak support for a correlation with end-of-evaluation ratings. As in experiment 2, we see again a subject pool effect, different in specifics but again indicating a weaker relationship between the in-the-moment and after-the-fact ratings for COSI participants, who reported after-the-fact ratings verbally, than for Prolific respondents, who used a drop-down menu.

6. Discussion. The goal of these three studies was to examine the validity, precision, and ease of use of the in-the-moment slider and to better understand social responses to individual linguistic features during the listening process. Across all three studies, we find that slider reactions do capture changes in response to both style shifts and sociolinguistic variables, but with considerable variability in the data, necessitating large, sometimes extremely large, samples.

The results of experiment 3 suggest that explicit awareness of a feature may correlate with slider responses, possibly more than with after-the-fact evaluations. The results for experiment 2 are more ambiguous, but the token-to-token slopes depending on the order of (ING) alternation suggests that slider responses may perhaps be more fragile, or more dependent on speech context, than after-the-fact ratings, which did not show such an effect. Note, however, that the five-second slopes analysis likewise showed only a main effect of (ING), across alternation patterns. These differences provide weak evidence that the two tasks may rely on different processes or different representations. It is also possible, however, that the high variability in the slider-based data makes the results more susceptible to factors that influence both tasks.

All three studies show strong correlations between mean slider ratings and after-the-fact ratings. Weaker evidence is found for correlations with end-of-evaluation ratings and maximum ratings. This suggests that a slider rating at any given point in time
should not be assumed to be equivalent to a listener’s current overall impression of a speaker, as it develops in real time. Instead, recent input is more heavily weighted in the slider task, while the after-the-fact task tends toward a more global view. In experiment 1, the final disfluent section of the stimulus pulled the slider ratings down by more than half the possible territory, from 62 to 29 points. The after-the-fact ratings, best predicted by the mean responses to this final segment, nonetheless show a mean of 5.54 out of 10. Subject-pool effects suggest that the specific evaluation task may play a role, since participants using a drop-down menu for after-the-fact ratings showed multiple examples of a closer relationship to their in-the-moment slider-based responses than did the participants who verbally reported their rating to an experimenter.

The finding that after-the-fact ratings correlate more strongly with mean in-the-moment ratings than with final ratings raises the possibility that the two tasks draw on different processes. As discussed previously, person perception is believed to depend on specialized cognitive structures accessed during recall-based tasks like evaluation. It is likely that after-the-fact speaker-evaluation tasks, like other common person-perception tasks, prompt perceivers to form social perceptions during the stimulus and to access the stored percept during the evaluation phase. It is less clear how in-the-moment reactions relate to this system. While responses may be based on a percept as it is updated, it is also possible that they are influenced directly by features available in the stimulus, using a more monitoring-like process. If this is the case, in-the-moment ratings may be able to capture useful and interesting information about participants’ reactions to specific linguistic features, but be unable to shed light on the dynamics of how person perceptions are updated over the course of speech perception.

Independent of the relationship of in-the-moment ratings to percept development, our findings also suggest that in-the-moment ratings may depend more heavily on explicit (i.e. verbally reportable) reactions than do after-the-fact evaluations. This suggests from a practical perspective that in-the-moment tools may be better suited to probing features amenable to verbal introspection.

The differences between after-the-fact ratings across subject pools further suggest that similarity in task details may influence ratings. This may relate to the details themselves—for example, direct effects of verbal vs. online evaluation—or may be due to participants feeling a greater need for self-consistency in their responses across more similar-seeming tasks.

We offer a somewhat pessimistic view on the potential for explicit in-the-moment evaluation work for illuminating the process of developing social percepts based on linguistic information. That said, reflecting on previous work that has used such techniques, we still see useful insights. The two major constraints on the tool are the lack of time precision, disallowing post-hoc feature identification, and the theoretical question of what processes are tapped. The latter point is not solvable methodologically, but also not prohibitive as long as we are careful about what conclusions are drawn. The former may be addressed in a variety of ways. Montgomery and Moore (2018) offer an innovative approach by asking listeners to reflect on their own responses in a second pass through the stimulus, indicating explicitly what features they responded to and allowing them to retract previous answers. Hesson and Shellgren (2015) use a matched-guise approach, as we did in experiments 2 and 3, ensuring tighter control over the linguistic features prompting the responses. In-the-moment response tools may well offer useful insight, particularly for features accessible to explicit verbal commentary, as long as the responses themselves are not expected to pinpoint specific linguistic features.
7. Conclusion. In conclusion, we find the slider task a useful but limited tool for capturing moment-to-moment reactions to sociolinguistic stimuli. The tool did show effects in the predicted directions in all three experiments. However, the variability in the data necessitates extremely high sample sizes and provides poor time granularity.

The relationships between the in-the-moment and the after-the-fact ratings suggest that speaker evaluation differs from ratings of high-emotion experiences as examined by Fredrickson and Kahneman (1993). While final in-the-moment ratings do show some correlation with global after-the-fact ratings, the mean ratings show stronger and more consistent effects. This suggests that in-the-moment ratings do not present a global view of a listener’s current understanding of a talker, but rather indicate perceptions based on a smaller time window. It is worth considering whether in-the-moment ratings may be based on or informed by a monitoring process distinct from the processes that form and update social percepts of talkers.

The evidence suggesting that distinct processes may underlie these two tasks combines with other evidence suggesting that a singular sociolinguistic monitor is insufficient to account for the complex range of sociolinguistic behavior documented in the literature. Campbell-Kibler (2016) proposed that we need at least a grammar with the ability to store contextual information, person perception, and self-regulation systems, as well as more general reasoning and memory systems, while noting that all of these systems are already necessitated by evidence from outside sociolinguistics. Austen (2020) and Campbell-Kibler (2021) have also provided some evidence that the mental links between linguistic features and other social objects may be different for different sociolinguistic behaviors, namely socially influenced speech perception vs. socially influenced speech production vs. linguistically influenced social perception. By offering evidence that in-the-moment and after-the-fact evaluations show different profiles of response, the current study further supports the hypothesis that sociolinguistic processing takes place across a range of cognitive systems, rather than being concentrated in a single module.

REFERENCES


Campbell-Kibler, Kathryn. 2007. Accent, (ING), and the social logic of listener perceptions. American Speech 82.32–64. DOI: 10.1215/00031283-2007-002.


Campbell-Kibler, Kathryn. 2021. The cognitive structure behind indexicality: Correlations in tasks linking /s/ variation and masculinity. Social meaning and linguistic variation: Theorizing the third wave, ed. by Lauren Hall-Lew, Emma Moore, and Robert J.


García, Christy; Abby Walker; and Kathryn Campbell-Kibler. 2016. Effects of linguistic environment on the social meaning of /s/ aspiration. Poster presented at the annual meeting of the Linguistic Society of America, Washington, DC.


Jones, L. Keeta. 2016. Updating evaluations of sociolinguistic variations in real time. Columbus: The Ohio State University qualifying paper.


Kahneman, Daniel; Barbara L. Fredrickson; Charles A. Schreiber; and Donald A. Redelmeier. 1993. When more pain is preferred to less: Adding a better end. Psychological Science 4.401–5. DOI: 10.1111/j.1467-9280.1993.tb00589.x.


WATSON, Kevin, and Lynn Clark. 2013. How salient is the nurse-square merger? *English Language and Linguistics* 17(2).297–323. DOI: 10.1017/S136067431300004X.


[Received 11 August 2020; revision invited 29 November 2020; accepted received 12 April 2021; accepted pending revisions 26 May 2021; revision received 7 July 2021; accepted 9 July 2021]