Are You Paying Attention?
The Cognitive Costs of Speech Perception Adaptation

Rachel Sabatello*[1,2] & Dr. T. Florian Jaeger[1,3]

[1]Department of Brain & Cognitive Sciences, The University of Rochester
[2]Department of Psychology, The University of Rochester
[3]Department of Computer Science, The University of Rochester

Author's Note

This project was funded by the University of Rochester Wiesman Summer Fellowship grant from the in Brain and Cognitive Sciences department, and a Research and Innovation Grant (RIG) from the University of Rochester.

*Corresponding author email: rachel.sabatello@gmail.com

Abstract

Spoken language production is highly variable, despite being one of the most prevalent forms of human communication. However, listeners can often understand newly encountered talkers when hearing them speak for the first time. In this experiment, we will investigate the role of attention in speech perception and adaptation by limiting the participant's available attentional resources. Specifically, if there are limits to the automaticity of speech perception, then we can expect listeners to adapt their speech perception to a talker only when they are directing their attention towards that talker's verbal stream. To explore this question, we expose participants to two simulated talkers simultaneously with different pronunciation variants. Participants were instructed to attend to one of the two talkers throughout the exposure phase of the experiment, and then report if that talker is saying a word or a nonword in a lexical recognition task. We then compared how participants perform on a lexical discrimination task for both simulated talkers, as well as the talker the participants were instructed to attended to across pronunciation variants. Unexpectedly, we found a lack of significant perceptual adaptation to either talker, suggesting that implemented paradigm had inhibited perceptual recalibration towards either talker. Potential explanations for this result are discussed further.

Keywords: speech perception, perceptual adaptation, attentional resources, cognitive load

Are You Paying Attention?
The Cognitive Costs of Speech Perception Adaptation

Spoken language is highly variable despite being one of the most prevalent forms of human communication in daily life; Even talkers with similar language backgrounds tend to differ in how they produce speech sounds, for example how they may distinguish /s/ (the "S" sound) from /ʃ/ (the "Sh" sound). Still, listeners can often understand newly encountered talkers when hearing them speak for the very first time. How they are able to flexibly categorize speech sounds despite the presence of natural variation is known as the *lack of invariance problem* (Appelbaum, 1996). Variation in speech presents a unique challenge for cognitive processing that is solved seemingly automatically: Our brains learn how talkers speak, and then apply this information to construct expectations about speech they encounter in the future (Kleinschmidt & Jaeger, 2015). This cognitive process often occurs without the listener even noticing. However, the absence of awareness brings into question potential limitations of this ability: do listeners passively sponge information from speech in their environment, or must they direct their attention towards a talker to learn how they speak?

A large body of research suggests that perceptual learning does not require conscious effort, nor is it inhibited by environmental distractions (Zhang & Samuel, 2014) or exposure to multiple talkers (Cummings & Theodore, 2022). However, earlier research has also found that listeners may use the context of speech when adapting, for instance the effects of stereotypes on initial expectations and labeling in phoneme mapping (Zhang & Samuel, 2014). There is even evidence that listeners consider causality when learning how talkers speak, for example accommodating a talker chewing on a pen while talking (Kraljic & Samuel, 2011). These latter findings could suggest that listeners store information based on a perceived utility. Therefore, in our experiment when there is competition for auditory processing resources, we expect that processing of "useful" information— information framed as being more relevant to the participant's task—will take precedence.

This theory is echoed in the findings of Dr. Samuel's 2016 paper, which strongly supports that adaptation is involuntary and robust to distractions unless the competing task requires some form of categorization of the auditory information. In Experiment 1, participants were exposed to two simulated talkers. One talker produced speech with a phonetically shifted /s/-/ʃ/ and was always presented before the 2nd talker. The 2nd talker did not produce a phonetic shift. Dr. Samuel found that when the onset of the second talker interrupted the 1st talker and the participant performed a lexical recognition task for the second talker's speech, the participant did not exhibit adaptation. This was also true in Experiment 3, where the second talker was replaced with an environmental sound (e.g., a doorbell) and the participant was asked to identify this sound. However, when the lexical recognition task was performed for the first talker's speech in Experiment 1, the participant did exhibit adaptation despite still being interrupted by the second talker (Samuel, 2016). By engaging in a categorization task that requires information from one of multiple competing audial streams, the perceived utility of that stream is elevated above the others. This suggests that adaptation is only automatic given that cognitive resources are available (i.e., not occupied with processing another talker's speech).

In our study, we investigate the role of attention in speech perception and adaptation by limiting the participant's available attentional resources. To this end, we expose listeners to two simulated talkers speaking simultaneously and test the effects of directing the listener's attention to one talker on the listener's ability to adapt to both talkers. This study expands on the findings of Samuel 2016 in several ways. Our paradigm removes the stimulus onset asynchrony (SOA) previously employed in Samuel 2016, resulting in both talkers' speech overlapping perfectly. This would increase the difficulty of the task—when speech naturally coincides, it rarely occurs simultaneously and is not usually limited to one word—and theoretically allow us to draw

conservative conclusions about the role of attention in perceptual adaptation. By doing so, we can also investigate listeners' ability to separate the talkers and maintain attention, whereas Samuel 2016 focuses primarily on the latter.

Our paradigm also introduces an atypical talker in place of the second talker in Samuel 2016. Both talkers have been engineered to have distinct voices and to produce inversely atypical in their speech (i.e., one talker produces their /s/ sounds more like "Sh," and the other talker produces their /ʃ/ sounds more like "S"). Participants were instructed to attend to one of the two talkers throughout the exposure phase of the experiment, and then select if that talker is saying a word or a nonword in a lexical recognition task, like the task in Experiment 1a in Dr. Samuel's 2016 paper. We then adopted a lexical discrimination task similar to that used in the aforementioned paper to measure if listeners have learned to expect the atypical pronunciation from one or both talkers. Removing the SOA between the two talkers introduced in the original paper also negates the potential influence of order on a listeners' ability to hone in on a stimulus.

The goal of this experiment is to investigate the automaticity of speech perception adaptation; specifically, how does directing attention to one talker compete with adaption to a second talker when both are speaking at the same time. If we are able to replicate the results of Samuel 2016 with this paradigm, providing evidence that perceptual adaptation to speech is contingent on a listener's attention being directed towards a given talker, then this paradigm may also be implemented in future experiments to determine what factors divert listener attention to one talker over others or environmental noise. Therefore, the omission of a SOA in the stimuli we employ in this experiment is critical for optimizing the future utility of this paradigm as it removes the potential temporal confound when observing participants' decision-making process.

We hypothesize that perceptual adaptation to speech is contingent on a listener's attention being directed towards a given talker. In this experiment, we therefore aim to simulate two distinct talkers that a listener will hear speak simultaneously. The listener will perform a lexical recognition task for one of these talkers (referred to as the Attended Talker henceforth), and then we will compare participants' adaptation to both the Attended Talker and the Unattended Talker. If perceptual adaptation requires a listener to be attending to the talker, then we would expect participants to only adapt to the attended talker. If perceptual adaptation does not require attention directed towards a given talker, then we would expect participants to exhibit adaptation for both talkers. We would expect to see the former trend in the data based on the results in Samuel 2016. If our hypothesis is not proven false, this paradigm could be used in the future to investigate features that may cause a listener to prioritize one verbal stream over another. By extension, this may have further implications for language learning and the possible effects of social biases on speech processing. If our hypothesis is proven false—contrary to our expectations based on prior research—then this novel finding could suggest that speech perception adaptation occurs automatically when the brain encounters any form of human speech if adaptation is observed for both talkers. Contradictorily, if our hypothesis is proven false due to no perceptual adaptation being observed for either talker, then this finding might lend support for the types of cognitive resources employed for word recognition and perceptual adaptation. Regardless of the findings, our results could have informative implications for how we theorize the brain actively collects, stores, and uses information to formulate expectations.

Methods

In this study, our goal is to determine if perceptual adaptation to speech is contingent on a listener's attention being directed towards a given talker. We therefore aim to simulate two distinct talkers that participants hear speak simultaneously. The listener will perform a lexical recognition task for one of these talkers (referred to as

the Attended Talker henceforth), and then we will compare participants' adaptation to both the Attended Talker and the Unattended Talker. If perceptual adaptation requires a listener to be attending to the talker, then we would expect participants to only adapt to the attended talker. If perceptual adaptation does not require attention directed towards a given talker, then we would expect participants to exhibit adaptation for both talkers.

Specifically, we measure participants' perceptual adaptation /s/-/ʃ/ productions. /s/-/ʃ/ exist on a continuum, spanning from an "s" sound (e.g., "**S**ock") to an "sh" sound (e.g., "**S**hock"). This continuum is determined by spectral energy, where /ʃ/ tends to be produced with more spectral energy than /s/ in English (X). However, the perceptual boundary between /s/-/ʃ/—when a "s" sound begins to be perceived as a "sh" sound and vice versa—is variable across listeners (Norris, McQueen, & Cutler, 2003), and potentially even within listeners. Earlier research suggests that listeners' adaptation to /s/-/ʃ/ production is talker-specific, meaning that listeners adjust their perceived boundary between /s/-/ʃ/ for each talker (Kraljic & Samuel, 2005).

We chose to the /s/-/ʃ/ continuum as our measure for this experiment because of its talker-specific quality; this ideally allowed us to simulate two talkers with inversely atypical /s/-/ʃ/ productions—e.g., Talker A produces a typical "s" sound and an atypical "sh" sound ("?sh", as in Publi**ss**er), while Talker B produce a typical "sh" sound and an atypical "s" sound ("?s", as in Dino**sh**aur)—without cross-talker contamination of perceptual adaptation (Cummings & Theodore, 2022; Kraljic & Samuel, 2007). Many earlier studies (Cummings & Theodore, 2022; Kraljic & Samuel, 2007; Trude & Brown-Schmidt, 2012; Luthra et al., 2021) have simulated two unique talkers within the same experimental exposure by presenting one talker as female-sounding, and another talker as male-sounding. We adopted this same strategy, as well as attempting to simulate different spatial positioning assigning each verbal stream exclusively to one ear, to simulate two distinct talkers.

Materials

This experiment utilizes two types of stimuli: *Exposure Items* and *Test Items*. Items refer to the sound pairings within each stimulus (i.e., S Word + Sh Word, ?s Word + ?sh Word, Word + Nonword), independent of the Voice that produces them. Exposure Items present two talkers per stimulus in the form of 2-alternative forced-choice (2-AFC) lexical recognition tasks, where participant must report whether Attended Talker says a Word of a Nonword. Exposure Items can be categorized as either Critical Items (/s/-/ʃ/ present) or Filler Items *(*/s/-/ʃ/ lacking). Test Items present one talker per stimulus and are in the form of 2-AFC lexical discrimination task where the participant must report if they perceive Sound A (ASI) or Sound B (ASHI).

Exposure Items

The experimental stimuli were adapted from the female stimuli set developed by Dr. Tanya Kraljic and Dr. Arthur Samuel (Kraljic & Samuel, 2005). These stimuli included 20 critical S Words that included a /s/ sound (e.g., Para**s**ite, pronounced /pærə**s**aɪt/) and 20 critical Sh Words that included an /ʃ/ sound (e.g., Ambi**ti**on, pronounced /æmbɪ**ʃ**ən/). The /s/-/ʃ/ sounds were presented within the first syllable in each word to optimize lexical access (Samuel, 2016). These same 40 words were also produced with an ambiguous sound, referred to as ?s Words and ?sh Words respectively, that would likely be perceived as if 's' and "sh" are being interchanged to the typical U.S. Native English speaker (e.g., Para**sh**ite, pronounced /pærə**ʃ**aɪt/; e.g., Ambi**s**on, pronounce /æmbɪ**s**ən/). Additionally, we used 64 filler words from this same stimulus set that did not contain any /s/ or /ʃ/ sounds, and 98 nonwords that followed the typical structure of English words and did not contain any /s/ or /ʃ/ sounds. A full list of all the words/nonwords used in this experiment is available in *Appendix A*.

Each of these recordings were transformed in PRAAT (Boersman, 2002) using PRAAT Vocal Toolkit (Corretge, 2022) to a typical male-presenting voice (format shift ratio: 0.8; new pitch median: 100 Hz) and a typical female-presenting voice (format shift ratio: 1.0; new pitch median: 180 Hz), based on the parameters determined in Luthra et al. (2001). Both the audio for our female voice and our male voice were transformed to closely mirror the voices that had been recognized as two distinct talkers in the aforementioned paper

The stimuli used in this experiment each contain a female voice in one ear and a male voice recording playing in the opposite. The recordings were paired using the audio editing application Audacity (Audacity Team, 2021). To create the Unambiguous Critical Items, each S Word was paired with an Sh Word. These word pairings were the same for the ?s Word/?sh Word pairings to create the Ambiguous Critical Items. The ambiguous words (?s Words and ?sh Words) were paired together rather than an ambiguous word to unambiguous word (e.g., "Beneficial" and "Parashite" or "Ambision" and "Coliseum") to avoid the participant being distracted by hearing an unexpected pronunciation when they should be attending to the other talker. The recordings paired within each Exposure Item were decided by listing the stimuli alphabetically and pairing the S Words at the top of the list with the Sh Words at the bottom of the list. For Filler Items, a Word was paired with a Nonword in a similar fashion. Exposure Item pairings were then subjectively evaluated for similarity, and pairs that were judged as sounding too similar were arbitrarily rematched. Items are numerically noted in *Appendix A.* We created four versions of each item, for a total of 80 potential critical stimuli:

1. The **(?)S Word/Word** spoken in the **Female Voice** in the **Left Ear**
2. The **(?)Sh Word/Nonword** spoken in the **Female Voice** in the **Left Ear**
3. The **(?)S Word/Word** spoken in the **Male Voice** in the **Left Ear**
4. The **(?)Sh Word/Nonword** spoken in the **Male Voice** in the **Left Ear**

Additionally, we added 200ms of silence at the beginning of each Item using the Insert Silence PRAAT script (Hirst & Daidone, 2018) to account for potential issues with Bluetooth connection that could cause the initial onset of both talkers to be inaudible.

Test Items

The Test Items were used to measure the effects of the experimental stimuli and were adapted from six tokens selected from the full 31-token ASI-ASHI continuum. Only six tokens are required because we are interested in participants' perceptual boundary between /s/ and /ʃ/; the point in this continuum where "asi" begins to be perceived as "ashi." Therefore, we will be focusing on six tokens in the middle of this continuum (tokens 13, 17, 18, 19, 20, and 24). As the tokens increase, the amount of spectral energy in the sound production increases, as should the probability that a Native U.S. English speaker would categorize the token as ASHI.

We adopted the 6-token from Experiment 1 of Cummings et al. (2022) rather than the 6-token subset developed by Kraljic and Samuel, (2005) because the Kraljic and Samuel subset seemed to be somewhat ASHI-biased in previous studies—participants tended to perceive ASHI more often across tokens—while Cummings et al. (2022) appear to be better-balanced around participants' perceptual boundary between ASI and ASHI.

The same gender transformation processes used to simulate the male and female voice for recordings in the Exposure Phase were applied to each of the six tokens. to produce two versions of the six Test Items: one 6-token subset in the simulated female voice, and one 6-token subset in the simulated male voice. By applying the same transformation process to the Test Items, our goal was for listeners to identify the Test Items as being spoken by the same talkers as the Critical Items and the Filler Items so we could measure adaptation in the Test Phase to the talkers from the Exposure Phase.

Design

This experiment is split into two main parts: an Exposure Phase and a Test Phase. The Exposure Phase contains a total of 80 unique trials: 10 Ambiguous Critical Trials, 10 Unambiguous Critical Items, and 60 Filler Trials. These trials are divided into 10 blocks, and trial presentation order was randomized within each block. The Test Phase contains a total of 72 trials divided into 12 blocks, and trial presentation order was randomized within each block. All Critical Words, Filler Words, or Filler Nonwords appear once and only once within each experiment, regardless of ambiguity of the /s/-/ʃ/ production in the word or the voice it is produced in.

Nuisance factors: Critical Items

The Critical Items presented during a given version of the experiment were determined by the Voice (Male of Female) the participant is instructed to attend to, and whether that Voice produces ?s Words or ?sh Words. To avoid repetition within the created stimuli, Critical Items were divided into Materials A (10 Critical Items) and Materials B (10 Critical Items). Material assignment was determined by listing the Critical Items alphabetically by the (?)S Word and assigning every other Item to Materials B. In this experiment, the Critical Items in Materials A were always the Ambiguous Critical Items (see *Figure 1*, below).

As a result, the ?sh Words in Materials A are produced in the same Voice as the S Words in Materials B (Talker A), and the ?s Words in Materials A are produced in the same Voice as the Sh Words in Materials B (Talker B). Our intended effect was to shift the participant's perception of Talker A's ASI-ASHI continuum towards ASI, and Talker B's ASI-ASHI continuum toward ASHI (see *Figure 2).*

As shown in *Figure 3* (below), each talker presented in the experiment would produce a total of 20 Critical Words, where all the ?s Words for Talker B and all the ?sh Words for Talker A would be shifted. Which Talker (A or B) would be the Attended Talker varied by experiment, as would the gender presentation of the Attended Talker's voice (Male of Female). In this experiment, the Critical Words for the Attended Talker were always presented in the participants' Left Ear because we did not expect this factor to significantly impact our results.

Materials A

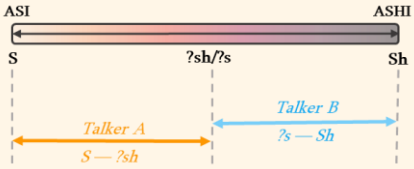
Sh	S
Talker A	*Talker B*
Ambition	Parasite
Machinery	Obscene
Brochure	Medicine
Official	Tennessee
Crucial	Peninsula
Pediatrician	Hallucinate
Flourishing	Arkansas
Reassure	Compensate
Graduation	Dinosaur
Vacation	Rehearsal

Materials B

Sh	S
Talker B	*Talker A*
Initial	Pregnancy
Beneficial	Democracy
Negotiate	Embassy
Commercial	Legacy
Parachute	Reconcile
Efficient	Personal
Publisher	Eraser
Glacier	Episode
Refreshing	Literacy
Impatient	Coliseum

Figure 1 (above): How Talker words are paired into Critical Items and grouped into Materials A (left) and Materials B (right) to assign the Ambiguous /s/-/ʃ/ to half the Critical Items.

Figure 2 (below): A visual illustrating how the /s/-/ʃ/ continuum is manipulated for Talker A and Talker B. More-ASI Items are represented in pink, and More-ASHI Items in purple. In this example, Talker A is assigned the ?sh sound and Talker B is assigned the ?s sound.



All Attended Talker Critical Items were presented in the same ear to avoid the possibility that perceptual adaptation is ear specific. Therefore, there were four potential Item configurations for the Critical Items.

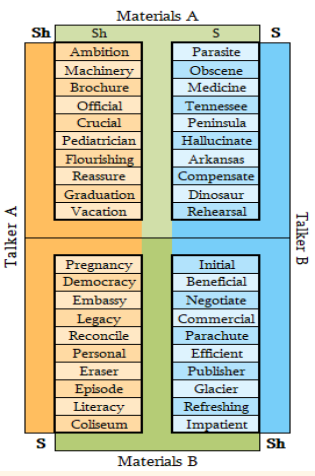


Figure 3 (left): A visual of how the Critical Words spoken by Talker A and Talker B are divided into Materials A and Materials B. In this experiment, the words in Materials A are produced with an ambiguous sound and the words in Materials B are produced with an unambiguous sound. During the Critical Trials, each Talker is always heard in the same ear (e.g., Talker A in the Left Ear and Talker B in the Right Ear, as shown to the left).

Nuisance factors: Filler Items

The Filler Items used in each version of the experiment were chosen so that, in total, half the items during the Exposure Phase presented the Attended Talker on the Left Ear, and half of the items presented the Attended Talker saying a Word.

This configuration served as a participation check: if participants were to remove the headset from one of their ears and only be exposed to one of the two talkers, they would select the correct response for the Attended Talker for 50% of the exposure trials. If the participant were to always respond Word (or conversely Nonword) for all exposure trials, they would select the correct response for the Attended Talker for 50% of the trials. If the participant reported their answers randomly without engaging in the lexical recognition task, they would be expected to report the correct response for the Attended Talker ~50% of the time.

To produce this configuration, the 60 Filler Items were divided into three sets of 20 Items: Set A, Set B, and Set C. To group the Filler Items into Sets, they were first listed alphabetically and then grouped by every third word. Each Set had four versions (see *Figure 5*, below).

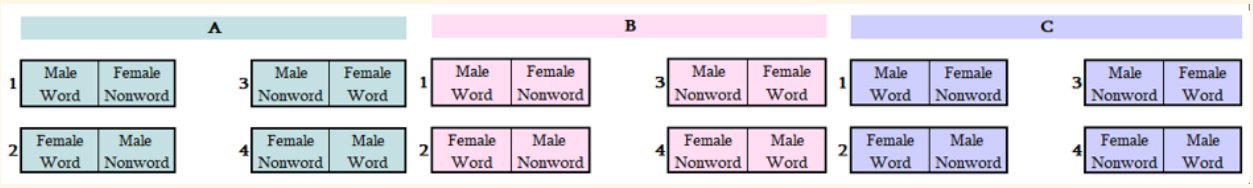


Figure 5: The 60 Filler Items (60 unique Word/Nonword pairs) used in this experiment were divided into 3 Sets (A, B, C) of 20 Filler Items. We created 4 Versions of each Set (1, 2, 3, 4) to produce every combination of factors—i.e., if the Male or Female voice speaking a Word or a Nonword, and if the voice is heard in the Left or Right Ear. The Version is noted by the number to the left of the set pair. The Ear presentation is represented by the horizontal position of the box within each pair (e.g., in A1 (Set A Version 1), the Female voice produces a Nonword in the Left Ear).

The Filler Items in Set A presented the Attended Talker producing a Word in the Right Ear. Set B presented the Attended Talker producing a Nonword in the Left Ear. Set C presented the Attended Talker saying a Nonword in the Right Ear. The result is eight potential Item configurations for the Exposure Phase (see *Figure 6*).

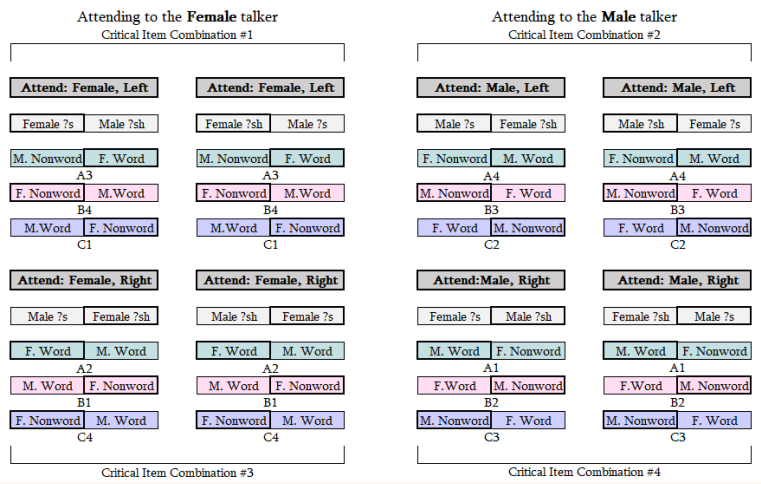


Figure 6: The Set Version assigned to each Critical Item combination. The 20 Critical Items in each experiment are shown in gray, Set A in teal, Set B in pink, and Set C in purple. The letter number combo below each pair (e.g., A3 below Set A in Combination #1 above) correlates to the Set and Version in Figure 4a. The ear the talker is presented in correlates to the position of the word in the pair (e.g., in Combination #1 above, the Female talker is presented in the Left Ear), and outline around the Attended Talker is bolded. Please See Appendix B for a full list of nuisance factor combinations.

The Exposure Items were then divided into ten blocks of eight items where each block included: 1 Ambiguous (?s/?sh) Critical Item, 1 Unambiguous (Sh/S) Critical Item, 2 Filler Items from Set A, 2 Filler Items from Set B, and 2 Filler Items from Set C. The order of presentation of Items is randomized within each block. In this experiment, we included two block different block orders our of the potential ten because we did not expect this factor to significantly impact our results.

Nuisance factors: Test Items

Each Test Block contains six trials: the 6-token Test Items in either the Male Voice or the Female Voice. The Test Items are presented in a random order within each block. The Voice that produces the Test Items changes every two blocks, to alleviate potential additional cognitive load from shifting between the expectations for each talker. Which Voice is presented first during the Test Phase varies across participant (see *Figure 7).*

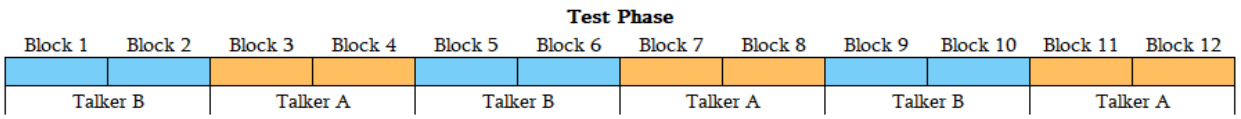


Figure 7: Each block has a total of 6 trials. Each item on the test continuum will play once in a randomized order before moving to the next block. The test continuum will be produced by both talkers 6 times each. Each talker will produce two blocks in a row before switching to the next to minimize additional strain on cognitive resources that may occur due to swapping between the two talkers. The talkers will alternate every two blocks. Which talker (male/female) speaks first will be counterbalanced within the experiment

Nuisance factors: response key mapping

Additionally, we counterbalanced the key mapping for reporting a response. Participants used the "X" key and the "M" key on their keyboard to report their responses. We alternated the key mapping for reporting a response as a Word or a Nonword during the Exposure phase, and for reporting a response as ASI or ASHI during the Test Phase. The result was 64 potential factor combinations:

Attended Talker Ambiguous Sound (?s/?sh) **x Attended Talker Gender** (Male/Female) **x Attended Talker Ear** (Left) **x Exposure Block Order** (1/5) **x Test Order** (A/B) **x Exposure Response Key Mapping** (Word/Nonword) **x Test Response Key Mapping** (ASI/ASHI)

Participants

A total of 64 participants (M = 32; age range 18-68, avg. = 34.483) were recruited from the online crowdsourcing program Prolific. This sample size is similar to that of the Experiment 1b in Samuel 2016; earlier similar studies typically recruited 30-50 participants (Samuel, 2016). Participants were self-reported English monolinguals, who were native to and spent most of their time in the United States before the age of 18. Additionally, participants had not participated in any other experiments launched by the Human Language Processing Lab on Prolific. The purpose of this criteria was to minimize variation in the perceptual boundary between S-Sh relative to the simulated talker's atypical pronunciation.

Before proceeding to the experiment, participants were asked to confirm that they spoke United States English most of the time before they were ten years of age in the experiment instructions. Participants were instructed to complete the study in one sitting, in a quiet room with minimal distractions. They were informed that they could not take the experiment multiple times or reload the experiment and were provided with reasons their work could be rejected. Each participant was provided with the participant consent form and were required to agree to proceed with the experiment. For all the provided instructions, please see *Appendix C.*

Additionally, participants were informed that over-ear or in-ear speakers (e.g., headphones) would be necessary for this experiment, oppose to external speakers (e.g., laptop speakers). These instructions were followed by a 6-trial 3-AFC tone comparison task (Woods et al., 2017) to confirm that participants were using proper, functional hardware. To continue to the experiment, participants were required to correctly identify the softest tone in five of the six trials.

Data Collection

Table 1: Participant recruitment

Recruitment (date)	*Start Time*	*Experiments Completed*	*Experiments Returned*	*Experiments Timed Out*	*Median Time*	*Notes*
02.19	15:45	10	15	0	16:18	Pilot 1: Test experiment Data collection error: response not recorded in CSV file
03.06	14:45	1	0	0	15:00	Pilot 2a: Correct recording error
03.06	16:48	1	2	0	15:00	Pilot 2b: Gender recruitment balance
03.06	20:23	53	48	1	16:11	Full Experiment: crashed on 1 participant before they were able to begin the experiment; list assignment error

Table 1: When participants were recruited, how many participants completed the experiment during a recruitment period, and how many participants did not complete the experiment. Also included is the medium time it took participants to complete the experiment during each recruitment and notes about the purpose and problems encountered during the recruitment period.

Exclusions prior to analyses

A total of 131 participants engaged with this study on Prolific. Of these participants, 48.85% (n = 64) completed the experiment. 67 participants requested to take the experiment but decided not to complete it. This low completion rate may be due participants regarding the task as too difficult for the reward, losing interest in the task, or failing the 3-AFC equipment check. 1 participant was excluded on 03.06 because they failed to complete the experiment in 56 minutes, the maximum time allotted by Prolific based on our estimated completion time of 15 minutes. Additionally, 1 participant on 03.06 contacted us and was excluded due to a technical malfunction that inhibited them from beginning the experiment.

Procedure

Participants were intended to be assigned to one of the 64 counterbalanced lists. However, some lists were assigned multiple participants and others were assigned none due to a technological error during the Full Experiment. A list of the number of participants assigned to each experiment version and were included in the analysis is available in *Appendix D*.

Practice Phase

After reading through the instructions, participants proceeded to the Practice Phase. This phase consisted of 4 filler trials. Participants were instructed to perform 2-AFC lexical recognition task for either the Male Talker or

the Female Talker by pressing either the "X" or "M" character on their keyboard. The Attended Talker and the response key mappings were kept consistent through both Practice and Exposure Phase. During the Practice Phase, participants were given feedback if they selected the correct response. In order to continue to the Exposure Phase, the participant had to correctly respond to all four filler trials consecutively.

Exposure Phase

Participants completed 80 2-AFC lexical recognition tasks for either the Male Talker or the Female Talker. We refer to this Talker as the Attended Talker. During each trial, participant heard either a Critical Item or a Filler Item and selected either the "X" or "M" key to report if the Attended Talker said a Word or a Nonword.

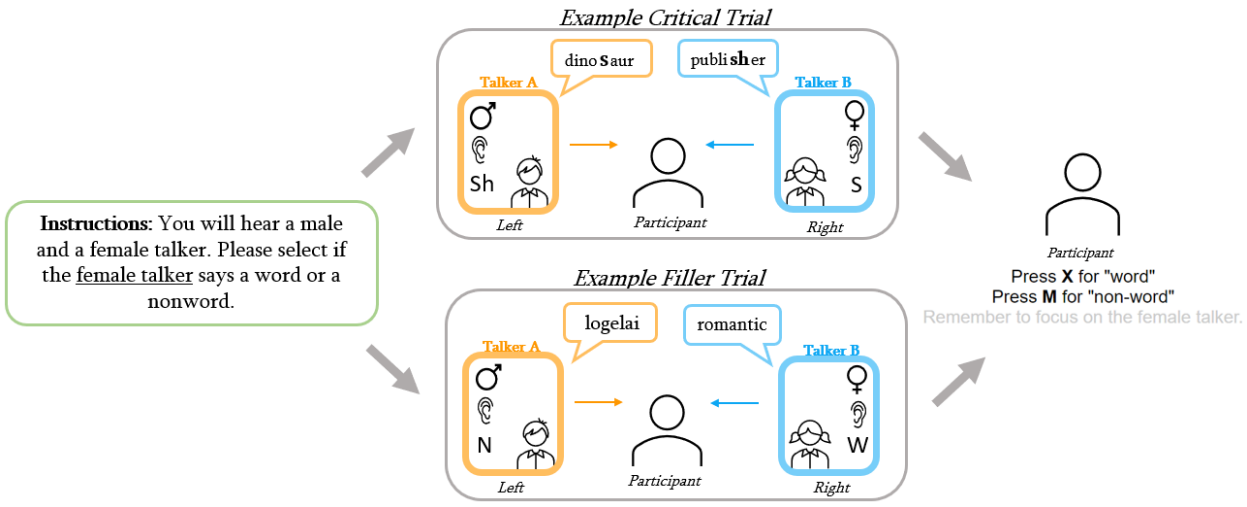


Figure 8: A visual diagram of a single trial. Participants will be instructed to attend to either the male (♂) or female (♀) talker at the beginning of the experiment (left, green box). One talker will be presented in the left ear, and the other in the right ear (pictured). Each subsequent exposure trial will feature two talkers, either in a critical trial (top) where both talkers produce a word that contains an S or Sh sound, or a filler trial (bottom) where one talker produces a word (W) and the other a nonword (N). The participant must then select if the Attended Talker produced a word or a nonword for each exposure trial (right, participant).

Test Phase

Participants then completed 72 2-AFC lexical discrimination task for both the Attended Talker and the Unattended Talker. Participants heard either the Attended Talker or the Unattended Talker produce a Test Item on the ASI-ASHI continuum. Participants were instructed to select the "X" key or the "M" key on their keyboard to report if they perceived the Talker as producing ASI or ASHI in the recording.

Exit Survey

Participants were prompted to complete a series of questions after finishing the Test Phase. Participants were instructed to respond truthfully, and that their responses would not affect their participation compensation. We first asked participants if they recalled a series of words from the experiment. Responses to this question are not addressed in this question, but the response data is available upon request. We then asked participants if they're audio stalled (yes or no), which gender talker they attended to in the Exposure Phase (male or female), and what type of equipment they used to listen to the stimuli (over-ear headphones, in-ear headphones, external speakers, or laptop speakers). Additionally, we asked participants to report their time zone and included the HLP lab demographic questionnaire.

We also posed a series of questions about their perception of the Attended Talker's and Unattended Talker's speech production, such as if the noticed anything odd about the Talker's pronunciation (free response), how they would describe the Talkers' pronunciation (faster than normal speaking rate, normal speaking rate, slower than normal speaking rate, higher than normal voice pitch, lower than normal voice pitch, relaxed, and/or serious), and if they noticed anything specifically about the Talkers pronunciation of /s/-/ʃ/ (both S and SH sounded normal, S sounded more like SH, Sh sounded more like S, and S sounded more like SH and Sh sounded more like S).

Results

Exclusion Criteria

Before collecting data, we outlined a set of exclusion criteria to eliminate participants who did not complete the Exposure Phase as instructed. After exclusions, 93.75% of participants (n = 60; Male = 29, Female = 28, NA = 3; Age μ = 34.036 years, Age σ = 10.966 years) remained for analysis. Participants averaged 4.842 incorrect responses to all Items (σ = 3.374). Participants averaged 3.667 incorrect responses for Filler and Unambiguous Critical Items (σ = 2.92), and 1.769 incorrect responses for Ambiguous Critical Items (σ = 1.012). A list of the number of participants analyzed per experiment version is available in *Appendix D*.

Self-Report Survey

Participants were excluded from analyses if they reported using audio equipment other than in-ear or over-ear speakers. In-ear or over-ear speakers are necessary to simulate different spatial positions of the two talkers, further distinguishing the voices. A total of 1 participant was excluded for this reason.

Participants were also excluded if they reported that they did not listen to the instructed Attended Talker during exposure. If perceptual adaptation is constrained by attention, then we would expect any participants who listened to the Unattended Talker to exhibit perceptual adaptation to the Unattended Talker. A total of 1 participant was excluded for this reason.

Lexical Discrimination Task Accuracy

Participants who performed poorly (< 80% correct) during the lexical discrimination task were excluded from analyses. Prior studies have excluded participants who have completed lexical discrimination tasks with an accuracy below 85-80%. We opted to exclude participants who were less than 80% accurate in their responses because of the anticipated difficulty of this task, and that some participants may have struggled to consistently recognize some of the ambiguous Critical Items as words. If a participant had randomly completed the lexical discrimination task, we would anticipate them to respond correctly approximately 50% of the time. One participant did exhibit this trend. A total of 2 participants were excluded for failing to accurately complete the lexical discrimination tasks at least 80% of the time.

Participants were also excluded if they failed to recognize more than 60% (6/10) of the ambiguous Critical Items as Words. Previous work has suggested that listeners adjust their perception of a talker's /s/-/ʃ/ production after as few as 6 observances of an atypical production. Therefore, this criterion was meant to confirmed that our labeling of the ambiguous sound category was recognized on at least 6 different instances by the participant. No participants met this exclusion criterion.

Response Time

While many previous studies have excluded participants or individual trials based on response time, we opted to not adopt this exclusion criteria. Typically, these criteria exclude participants who have completed either the entire experiment or a single trial in three standard deviations greater or less than the average participant experiment/trial completion time. As a result, data tends to be excluded from participants who responded very quickly to trials (i.e., may have randomly completed the task) or very slowly (i.e., may have gotten distracted during the experiment). We did not feel a need to implement either of these criterion because A) Participants would achieve ~50% accuracy for the lexical discrimination task if they responded at random or removed one ear from their headphone/earphones, and B) Prolific automatically excludes participants from providing data when they exceed a maximum allotted time based on our estimated time of completion and the value of our compensatory payment. The median completion time for each recruitment period is reported in *Table 1*.

Analysis

After exclusions, 60 participants remained. Of these 60 participants, 32 attended to a talker with an ambiguous S pronunciation (?s), and 28 attended to a talker with an ambiguous Sh pronunciation (?sh). We determined the average proportion of ASHI responses to each Test Item (13, 17, 18, 19, and 24) with a 95% confidence interval for when the Attended Talker produced ?s and when the Attended Talker produced ?sh in *Figure 8,* below.

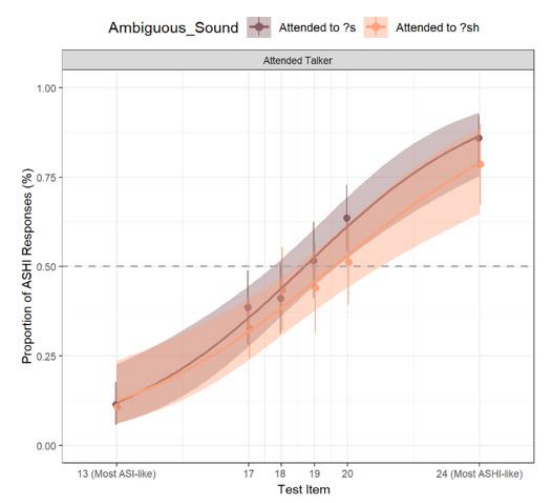
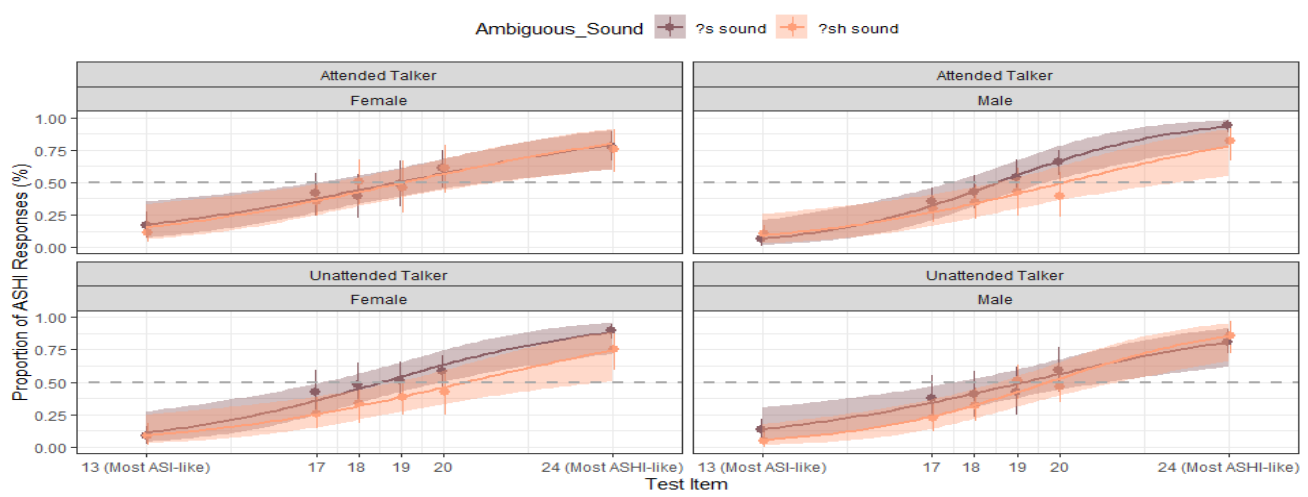


Figure 8a (left): Average proportion of ASHI responses by Test Item compared by the Attended Talker's ambiguous sound production (?s, purple or ?sh, pink). Each plotted point correlates to the calculated average number of ASHI responses for a given Test Item. The shading around the calculated psychometric curve represents a 95% confidence interval. The gray dashed line marks where the proportion of ASHI responses equals 50%; where the curves intercept this line represents the estimated perceptual boundary for the Talker, when participants are equally likely to respond either ASI or ASHI for that Test Item.

Figure 8b (below): Average proportion of ASHI responses by Test Item compared by the Attended Talker's ambiguous sound production, plotted by the talker gender and the attended/unattended status of the talker. There is no significant difference between the proportion of ASHI responses for the across the six test items for the Attended Talker or the Unattended Talker when compared within the Voice of the talker, suggesting that the Voice is not responsible for our null effect.



In *Figure 8a*, we expected the purple psychometric curve to be in the right-most position, and the pink psychometric curve to be positioned left-most: Participants who attended to the ?s talker were expected to perceive a typical /ʃ/ sound and an ?s sound between a typical /ʃ/ and /s/, eliciting more ASHI responses for the more ASI-like Items. Participants who attended to the ?sh talker were expected to perceive a typical /s/ sound and an ?sh sound between a typical /ʃ/ and /s/, eliciting less ASHI responses for the more ASHI-like Items.

Our pre-determined measure of significance before analysis is that the average proportion of ASHI responses for the Attended Talker producing ?s is not within the 95% confidence interval of the proportion of ASHI responses for the Attended Talker producing ?sh, or vice versa. Based on this criterion, we do not see a significantly different adaptation to the Attended Talker based on the talkers' pronunciation, suggesting perceptual adaptation did not occur to either ambiguous sound for the Attended Talker.

We then compared the average perceptual adaptation between the Attended Talker (represented in blue) and the Unattended Talker (represented in orange) within each experiment (*Figure 9,* below).

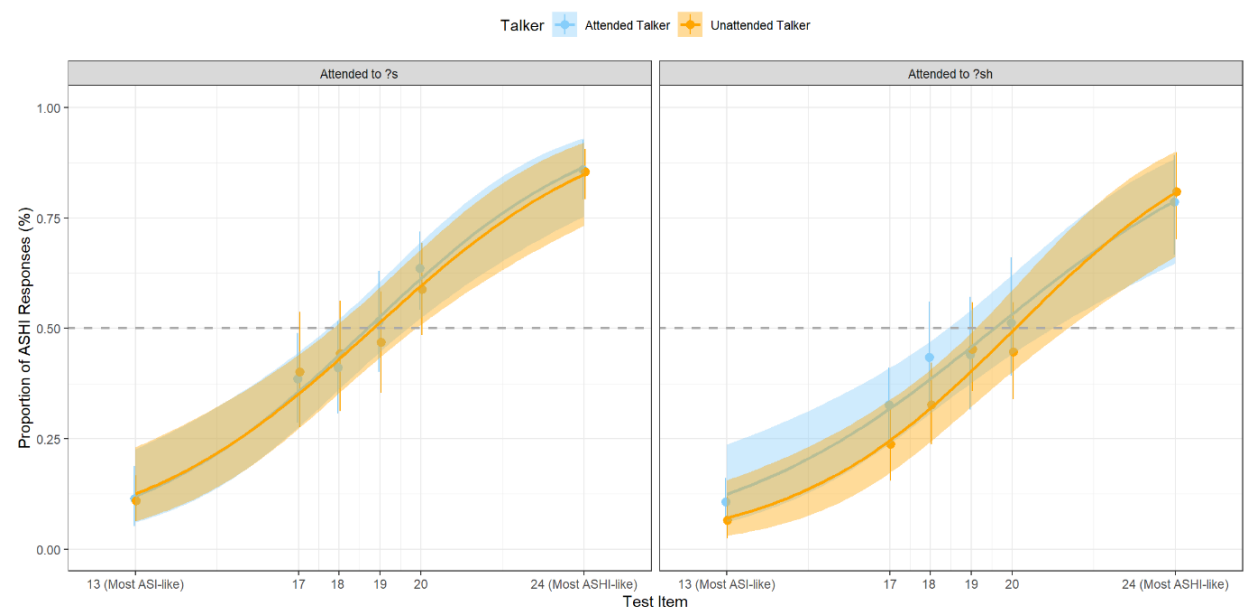


Figure 9: The average proportion of ASHI responses by Test Item for the Attended Talker (blue) and the Unattended Talker (orange). The responses are separated by the ambiguous sound production of the Attended Talker (?s, left; ?sh, right); In the left panel, the Attended talker produced an ?s, and the Unattended Talker produced an ?sh sound. The inverse is true in the right panel. Each plotted point correlates to the calculated average number of ASHI responses for a given Test Item. The shading around the calculated psychometric curve represents a 95% confidence interval. The gray dashed line marks where the proportion of ASHI responses equals 50%; where the curves intercept this line represents the estimated perceptual boundary for the Talker, when participants are equally likely to respond either ASI or ASHI for that Test Item.

Using the same pre-determined criterion for significance, we did not find a significant difference between how participants perceived the Attended Talker with an ambiguous pronunciation (either ?s or ?sh) or the Unattended Talker with the inverted ambiguous production (either ?sh or ?s). This result further substantiates that no perceptual adaptation occurred to either talker.

We also compared the average proportion of ASHI responses for each Test Item by test block, and compared the data separately based on the genders assigned to the Attended Talker and the Unattended Talker. This allowed us to determine if there was perceptual adaptation in some of the earlier test blocks that then assimilated over

time, or if there were differences in perceptual adaptation influenced by the voices assigned to the talkers. We do not see indications of either effect in *Figure 10,* on the following page.

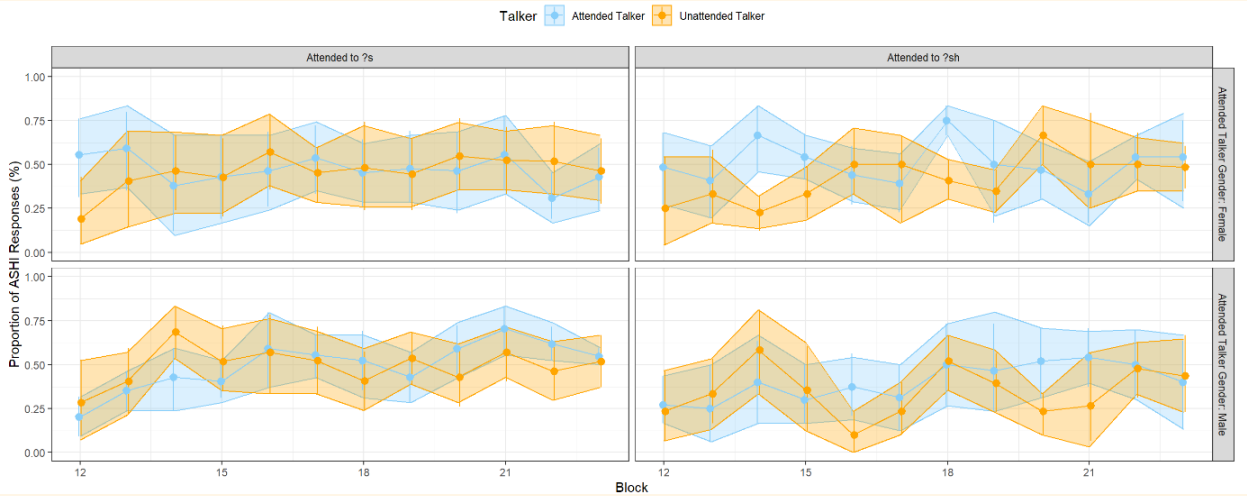


Figure 10: The proportion of ASHI responses across participants by block for both the Attended Talker (blue) and the Unattended Talker (orange). The points represent the proportion of ASHI responses per block, and the shaded regions represents a 95% confidence interval. Proportion of ASHI responses are separated by the whether the Attended Talker produced an ambiguous S sound (left) or an ambiguous Sh sound (right), and by the Attended Talker Gender (Male, top; Female, bottom). Generally, the proportion of ASHI responses did not vary significantly between blocks, nor were they impacted by the ambiguous sound produced by the Attended Talker or the Talker gender.

Discussion

Though our results suggest that participants failed to adapt their perception to either talker, they do potentially support that necessity of attentional resources for the speech perception adaptation. There are several possibilities as to why we did not observe adaptation to either talker, the most promising of which possibly support unintended effects of the paradigm and/or exposure tasks. If our results were caused by the presence of two simultaneous talkers degrading the perceived speech signal—possibly in conjunction with us employing a lexical recognition task during our Exposure Phase—then the additional strain on attention and/or other cognitive resources may be responsible for the lack of perceptual adaptation. Further research is necessary to determine the validity of this conclusion.

Improper task completion

The 2-AFC lexical recognition tasks

One possible reason that we do not observe perceptual adaptation to either talker is that participants did not complete the task as instructed. However, this explanation is highly unlikely. Our experiment design and pre-determined exclusion criteria should have omitted the data from any participant who selected their responses at random, attended to the incorrect talker, or only listened to the audio in one ear. Most of the participants reported their responses correctly at least 80% of the time—compared to the 50% threshold that would have suggested accuracy by chance—and all the participants included in this analysis correctly responded to at least 60% of the Critical Items, and at least 50% of the ambiguous Critical Items and 50% of the unambiguous Critical Items. This confirms that participants were accurately recognizing the Words presented in the experiment and should have recognized enough of the words in the Critical Items to have provided labels for the ambiguous sounds.

Furthermore, participant responses to the exit survey questions about the talkers' /s/-/ʃ/ production (*Tables 2a* and *2b,* below) suggest that some participants may have consciously noticed the atypical pronunciation in one or both talkers. Though accuracy in this form of reporting tend to be inaccurate—many participants may not be able to reliably recall the direction of the ambiguous sound (/s/ produced like "sh" or /ʃ/ produced like "s"), or if they had heard anything odd about the pronunciation in the Exposure Phase at all after having completed the Test Phase before the exit survey—the inaccuracy tends to be biased towards participants reporting not having noticed the (correct) ambiguous pronunciation.

Table 2a: Percent of participants who reported an atypical s-sh pronunciation for each talker

Talker	Atypical Pronunciation: Yes	Atypical Pronunciation: No
Attended	32.76%	67.24%
Unattened	25.86%	74.14%

Table 2b: Percent of the reported atypical pronunciations correct

	Correct Atypical Pronunciation: Correct	Correct Atypical Pronunciation: Incorrect
Attended	57.89%	42.10%
Unattended	40.00%	60.00%

In combination, these factors suggest that participants did complete the 2-AFC lexical recognition tasks in the Exposure Phase as intended.

The 2-AFC lexical discrimination tasks

It is also possible that participants did not complete the 2-AFC lexical discrimination tasks in the Test Phase as intended. Due to the nature of the discrimination task, participants' responses could not be assessed as correct or incorrect, like in the Exposure Phase. It is possible that participants lost interest in completing the task and

Figure 11: Color gradient panels illustrating the number of ASHI responses for each Test Item, separated by whether the Attended Talker produced an ambiguous S sound (left) or an ambiguous Sh sound (right), and whether the Test Item was produced in the voice of the Attended talker (top) or the Unattended Talker (bottom). Each column represents one participant; 32 participants attended to the talker who produced an ambiguous S sound, and 28 participants attended to the talker that produced an ambiguous Sh sound. Generally, the number of ASHI responses was high for the Test Item that were synthesized to sound most like ASHI (top row, lightest color), and low for the Test Item that were synthesized to sound most like ASI (bottom row; darkest color). However, the number of ASHI responses for the middle four Test Items (17, 18, 19, 20) appears more varied regardless of the ambiguous sound or the Talker.

began selecting answers randomly, but this possibility is not supported by the direction of the psychometric curves in *Figures 8* or *9*—as the Test Items became more ASHI-like, the proportion of ASHI responses increased—nor is it supported by the relatively consistent average proportion of ASHI responses for each Test Item across each block (*Figure 10).*We then compared the number of ASHI responses for each Test Item by participant. We separated this data by talker and by ambiguous pronunciation in *Figure 11,* below.

Based on this figure, the proportion of ASHI responses for each Test Item varied widely within participants who were exposed to the Attended Talker with the same ambiguous productions. Further analyses are necessary to determine if there is significantly more variation in the number of ASHI responses for each Test Item for the Unattended Talker than the Attended Talker.

A reason we may have encountered this level of variation across participants is that we did not screen our subjects for expected ASI-ASHI continuum categorization before they engaged in the experiment, which was done in Samuel 2016 and multiple other papers. The criteria used in Samuel 2016 was as follows:

> Following our usual procedures for recalibration experiments (e.g., Kraljic, Brennan, & Samuel, 2008; Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic, Samuel, & Brennan, 2008), in this and in all of the following experiments the labeling performance of each subject was initially screened to assure that the participant was able to systematically identify the members of the "asee" – "ashee" continuum. This screening eliminates any participants who made no effort to do the task, or who for some reason could not do so. Two exclusion criteria were used. The main criterion for exclusion was a failure to have at least a 35% difference in "sh" report between the most extreme "s" and the most extreme "sh" on the six-step continuum (as will be seen in the data below, the norm is a difference of 80–90%). Subjects who do not show at least this much ability to discriminate the endpoint members of the test series are effectively not doing the task that they have been asked to do. (p. 95)

Figure 11 suggests that many participants were able to discriminate the Test Items at the end of our ASI-ASHI continuum (13 and 24, respectively), but some were unable to do so. Further analyses would be necessary to determine if our collected data would follow the 35% difference criterion—though the only available data from this experiment is after exposure, which may have influenced the perception of participants who would have potentially met these criteria beforehand. We also used different test tokens on the ASI-ASHI continuum than those used by Samuel 2016, which may have also impacted our results.

Speech signal degradation

Given what we can gather from the participants' task performance, it appears unlikely that the lack of perceptual adaptation observed in this experiment was due to participants not completing the tasks as intended. Another alternative explanation for our observed results is that simulating both talkers simultaneously without a Stimulus Onset Asynchrony (SOA) like that implemented in Samuel 2016 may have caused the speech signal to become too degraded for participants to recognize fine-grained phonetic cues, a form of bottom-up processing. Instead, participants may have used top-down processing to accurately complete the lexical recognition task. This would mean that participants were able to process enough of the verbal stream to recognize the word being produced, therefore being able to correctly recognize the sound as a Word or a Nonword, without processing the ambiguous sound within the word. It is also possible that the ambiguous

sounds were attributed to environmental noise caused by the Unattended Talker, though we had tried to mitigate this possibility by pairing the ambiguous sounds together.[1]

Samuel 2016 Tasks

It may be important to note that in Samuel 2016 the task performed by participants varied across Experiments 1a and 1b: The participants completed a lexical recognition task in Experiment 1a, while the participants in Experiment 1b were tasked with counting the number of syllables the female talker produced. In Experiment 1a, the listener attended to the male talker who produced a typical pronunciation pattern and was the second talker to speak, interrupting the female talker. In contrast, the listener attended to the female talker—who always began speaking first and produced an atypical pronunciation—in Experiment 1b.

In subsequent experiments (2a, 2b, 2c, 2d, 2e), the participants were instructed to perform lexical recognition tasks for the male talker as the SOA increased, allowing the female talker to be heard more clearly. However, this was not the case in Experiment 3, where the female talker was interrupted by an environmental sound. In this task, the participant had to report whether the environmental sound was produced by a living or non-living thing. Though this task does not require lexical access to complete, perceptual adaptation to the female talker was observed to be inhibited at a SOA of 200ms. This is important because other earlier studies, for example Zhang & Samuel, (2018) have suggested that initial lexical access (identifying a word) is automatic, while maintaining competing lexical candidates requires is strained by additional cognitive load. If lexical recognition is a more automatic process—compared to counting the number of syllables in a word or evaluating the source of a noise, then we might fail to observe perceptual adaptation in this experiment—then participants may be automatically attempting to recognize both verbal streams as real English words when completing this task, therefore constraining the cognitive resources available for perceptual adaptation.

Implications, applications, and future directions

Though we have posed some possibilities, further research is necessary to determine the cause of our observed results. Potentially, this experiment may lend insight to theories about the relationship between cognitive load, lexical access, and phonetic processing. These results appear to support that speech perception adaptation is not an automatic process, suggesting a limit to the amount of noise a speech signal can be immersed in before fine-grained phonetic details are no longer processed despite the allocation of attentional resources. It is also possible that our results reflect the contexts in which perceptual adaptation would more-or-less naturally occurs in that had not been previously been investigated in a research setting to the best of our knowledge: If a listener is attempting to extract semantic information from a speech signal while also struggling to process the information being communicated in the interaction as a whole—e.g., A student trying to understand the material of a lecture that is being communicated in a thick, unfamiliar accent—then perhaps perceptual adaptation to the speech signal is constrained to some degree.

Future research may include replicating this experiment with a SOA <= 200ms, based on the findings of Samuel 2016, to continue probing the limits of speech perception adaptation in relation to attention. If there is perceptual adaptation to the Attended Talker but not the Unattended Talker, then this would provide additional evidence for speech perception adaptation being dependent on attentional resources. However,

---

[1] As a reminder, ?s and ?sh are the same ambiguous sound between a typical /s/ and typical /ʃ/. By presenting this sound in an identifiable English word, our goal was to label the ambiguous sound as either an /s/ or an /ʃ/.

implementing a SOA may impact participants' ability to adapt to the Unattended Talker, and may neglect potential differences between the cognitive process of honing in on a stimulus and having attentional resources be redirected. Future research may also seek to investigate how the complexity of a task may impact the processing and storage of speech information in perception adaptation, or the differences in the cognitive processes behind between initial lexical access and lexical competition. Both proposals have the potential application of bettering our understanding of how the brain processes speech and how we can possibly optimize our ability to do so.

Acknowledgements

A special thank you to Shawn Cummings, the Human Language Processing Lab, and the University of Rochester Brain & Cognitive Sciences department for all their help and support. Additionally, thank you to Dr. Tanya Kraljic and Dr. Arthur Samuel for their permission to use the stimuli they developed in their 2005 paper (Kraljic & Samuel, 2005).

References

Appelbaum, I. (1996). The lack of invariance problem and the goal of speech perception. In *Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP'96 (Vol. 3, pp. 1541-1544). IEEE.

Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. https://audacityteam.org

Boersman, P. (2002). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Corretge, R. (2022). PRAAT Vocal Toolkit. [Computer application]. https://www.PRAATvocaltoolkit.com

Cummings, Karboga, Yang, & Jaeger (2022). Causal Inference in Speech Perception. [Manuscript in preparation]

Cummings, S. N., & Theodore, R. M. (2022). Perceptual learning of multiple talkers: Determinants, characteristics, and limitations. *Attention, perception & psychophysics*, 84(7), 23352359.https://doi.org/10.3758/s13414-022-02556-6

Hirst, D., Qi, Z., & Daidone, D. (2018). PRAAT script Insert silence. https://www.ddaidone.com/uploads/1/0/5/2/105292729/insert_silence_at_start_of_all_files_in_folder.txt

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. https://doi.org/10.1037/a0038695

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, 121(3), 459-465.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1), 1-15.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. *Cognitive psychology*, *51*(2), 141-178.

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, & Psychophysics*, *83*, 2217–2228.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2), 204-238.

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, 88, 88–114. https://doi.org/10.1016/j.cogpsych.2016.06.007

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 27(7-8), 979-1001.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review, 28*, 1003–1014.

Woods, K.J.P., Siegel, M. H., Traer, J. & McDermott, J. H. (2017) Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*.

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1), 200–217. https://doi.org/10.1037/a0033182

Zhang, Xujin, and Arthur G. Samuel. "Is Speech Recognition Automatic? Lexical Competition, but Not Initial Lexical Access, Requires Cognitive Resources." *Journal of Memory and Language 100* (June 1, 2018): 32–50. https://doi.org/10.1016/j.jml.2018.01.002.

APPENDIX A

Words/Nonwords used in stimuli

Critical S Words

1. Ambition
2. Machinery
3. Brochure
4. Official
5. Crucial
6. Pediatrician
7. Flourishing
8. Reassure
9. Graduation
10. Vacation
11. Initial
12. Beneficial
13. Negotiate
14. Commercial
15. Parachute
16. Efficient
17. Publisher
18. Glacier
19. Refreshing
20. Impatient

Critical Sh Words

1. Parasite
2. Obscene
3. Medicine
4. Tennessee
5. Peninsula
6. Hallucinate
7. Arkansas
8. Compensate
9. Dinosaur
10. Rehearsal
11. Pregnancy
12. Democracy
13. Embassy
14. Legacy
15. Reconcile
16. Personal
17. Eraser
18. Episode
19. Literacy
20. Coliseum

Filler Words

1. America
2. Bakery
3. Ballerina
4. Blueberry
5. Bullying
6. Burglary
7. Camera
8. Continually
9. Directory
10. Document
11. Domineering
12. Dynamite
13. Eighty
14. Embody
15. Gardenia
16. Grammatical
17. Gullible
18. Hamburger
19. Honeymoon
20. Hurdle
21. Identical
22. Illuminate
23. Inhabit
24. Interior
25. Ironic
26. Keyboard
27. Knowingly
28. Laminate
29. Lengthen
30. Lethal
31. Liability
32. Lingering
33. Lobbying
34. Marina
35. Melancholy
36. Membrane
37. Metrical
38. Napkin
39. Negate
40. Nightmare
41. Ornament
42. Outnumber
43. Panicking
44. Parakeet
45. Pilgrim
46. Pineapple
47. Platonic
48. Predict
49. Purgatory
50. Tactical
51. Terminal
52. Therapeutic
53. Titanium
54. Turbulent
55. Tutorial
56. Umbrella
57. Undermine
58. Wealthy
59. Worldly
60. Wrinkle

Filler Nonwords

1. Youmgel
2. Wominid
3. Wojalto
4. Ungelnin
5. Tymolape
6. Tounamplem
7. Tilegkalo
8. Tamical
9. Ryligal
10. Rimkeluwar
11. Rengime
12. Rawamtee
13. Rakil
14. Pourilar
15. Ploupelai
16. Perkum
17. Omperoge
18. Nomerae
19. Niritaly
20. Nempring
21. Namuary
22. Mibgem
23. Meidnow
24. Makid
25. Logelai
26. Lilgrai
27. Lenediaw
28. Kloumidiger
29. Kermimer
30. Kelabidel
31. Kaldemia
32. Itempider
33. Inpaki
34. Imdalier
35. Ibirak
36. Hintarber
37. Halomimoc
38. Halken
39. Gerbualo
40. Gardimuallay
41. Ganla
42. Galliwinou
43. Gairelom
44. Emhoutic
45. Dilkuaund
46. Dadigal
47. Bowidai
48. Bikanian
49. Bamtell
50. Baliber
51. Anolipa
52. Anemer
53. Amalar
54. Alnadiro
55. Almikquary
56. Aknid
57. Ailounam
58. Aigi
59. Admunker
60. Acomining

Practice Words

1. Corridor
2. Heroine
3. Parable
4. Younger

Practice Nonwords

1. Nawinow
2. Neramgory
3. Kerkrum
4. Rikmaral

Appendix A: All audio recordings were produced by the same female talker for Kraljic & Samuel (2005). Recording numbers reflect the item pairs: Critical S Words 1-20 is paired with Critical Sh Words 1-20; Filler Words 1-60 are paired with Filler Nonwords 1-60; Practice Words 1-4 are paired with Practice Nonwords 1-4. To synthesize a typical male-presenting voice, we applied a format shift ratio of 0.8 to the recordings and new pitch median of 100 Hz to the recordings. To synthesize a typical female-presenting voice, we applied a format shift ratio of 1 and a new pitch median of 180 Hz to the recordings (Luthra et al., 2021).

APPENDIX B

Counterbalancing of non-nuisance factors

Critical Items						List	Attended Talker	Filler Items											
What qualities are assigned to the two talkers in the critical trials						*Versions of experiment*	*which Talker the participant performs the task for*	**Set Combo** *Set versions in 1*	**Set A** *Attended Talker Word: Opposite Ear*				**Set B** *Attended Talker Nonword: Same Ear*				**Set C** *Attended Talker Nonword: Opposite Ear*			
	Talker	**Shift**	**Materials**	**Gender**	**Ear**				**Version**	**Gender**	**Word**	**Ear**	**Version**	**Gender**	**Word**	**Ear**	**Version**	**Gender**	**Word**	**Ear**
1	**A**	?s	a	Female	Left	**1**	**A**	**1**	**3**	Female	Word	Right	**4**	Female	Nonword	Left	**1**	Female	Nonword	Right
	B	?sh	b	Male	Right					Male	Nonword	Left		Male	Word	Right		Male	Word	Left
2	**A**	?s	a	Female	Right	**2**	**A**	**3**	**2**	Female	Word	Left	**1**	Female	Nonword	Right	**4**	Female	Nonword	Left
	B	?sh	b	Male	Left					Male	Nonword	Right		Male	Word	Left		Male	Word	Right
3	**A**	?s	a	Male	Left	**3**	**A**	**4**	**4**	Male	Word	Right	**3**	Male	Nonword	Left	**2**	Male	Nonword	Right
	B	?sh	b	Female	Right					Female	Nonword	Left		Female	Word	Right		Female	Word	Left
4	**A**	?s	a	Male	Right	**4**	**A**	**2**	**1**	Male	Word	Left	**2**	Male	Nonword	Right	**3**	Male	Nonword	Left
	B	?sh	b	Female	Left					Female	Nonword	Right		Female	Word	Left		Female	Word	Right
5	**A**	?sh	a	Female	Left	**5**	**A**	**1**	**3**	Female	Word	Right	**4**	Female	Nonword	Left	**1**	Female	Nonword	Right
	B	?s	b	Male	Right					Male	Nonword	Left		Male	Word	Right		Male	Word	Left
6	**A**	?sh	a	Female	Right	**6**	**A**	**3**	**2**	Female	Word	Left	**1**	Female	Nonword	Right	**4**	Female	Nonword	Left
	B	?s	b	Male	Left					Male	Nonword	Right		Male	Word	Left		Male	Word	Right
7	**A**	?sh	a	Male	Left	**7**	**A**	**4**	**4**	Male	Word	Right	**3**	Male	Nonword	Left	**2**	Male	Nonword	Right
	B	?s	b	Female	Right					Female	Nonword	Left		Female	Word	Right		Female	Word	Left
8	**A**	?sh	a	Male	Right	**8**	**A**	**2**	**1**	Male	Word	Left	**2**	Male	Nonword	Right	**3**	Male	Nonword	Left
	B	?s	b	Female	Left					Female	Nonword	Right		Female	Word	Left		Female	Word	Right

Appendix B: A table depicting the possible Critical Item and Filler Item combinations. The factors we anticipate potentially confounding our results are the Attended Talker *Shift, Gender,* and *Ear.* For this experiment, the Attended Talker is always the talker assigned Materials A. The *Shift* column describes what ambiguous sound each talker produces (either ?s or ?sh), the *Gender* column describes the simulated voice assigned to the talker (male of female), and the *Ear* column describes which ear of the headset that the talker's voice is produced in. Every combination of *Shift* X *Gender* X *Ear* produces a total of 8 *List* combinations. Based on these factors, a *Set Combo* of Filler Items is selected. Each Set includes 20 Filler Items, with 20 Filler Words and 20 Filler Nonwords. The words within a given set share the same *Gender* X *Word X Ear* configuration. The *Gender* X *Word X Ear* factors applied to the Set is described by the *Version* column. The *Gender* column describes the simulated voice assigned to the talker (male or female), the *Word* column describes whether the talker produces Words or Nonwords within those 20 Filler Items, and the *Ear* column describes which ear the talker is heard in. The *Gender* of the talkers is consistent across all Filler Sets. The 20 Filler Items in *Set A* have the Attended Talker producing a Word in the opposite Ear assigned to the Critical Items. The Filler Items in *Set B* have the Attended Talker producing a Nonword in the same Ear as assigned to the Critical Items. The Filler Items in *Set C* have the Attended Talker producing a Nonword in the opposite Ear assigned to the Critical Items. As a result, each participant hears the Attended Talker produce a Word for 50% of the trials, hears the attended talker in the Right Ear for 50% of the trials, and hears the Attended Talker produce a Word in the Right Ear for 25% of the trials.

APPENDIX C

Participant Instructions

Attentional effects on listening

Thank you for your interest in our study! This is a psychology experiment about how people understand speech. You will listen to recorded speech and press a button on the keyboard to tell us what you heard.

Please read through each of the following requirements. If you do not meet all requirements, please do not take this experiment. You can click the names below to expand or close each section.

Experiment length

The experiment takes 15-20 minutes to complete, and you will be paid $3.20

Language requirements (grew up speaking American English)

You must be a native speaker of American English. **If you have not spent almost all of your time until the age of 10 speaking English and living in the United States, you are not eligible to participate.**

Environment requirements (quiet room)

Please complete this experiment in one sitting and in a quiet room, away from other noise. Please do NOT look at other web pages or other programs while completing this experiment. It is important that you give this experiment your full attention.

Headphone check

Please complete the following headphone test to make sure your audio setup is compatible with this experiment, and that your headphones are set to a comfortable volume.

First, set your computer volume to about 25% of maximum. Press the button, then **turn up the volume on your computer until the calibration noise is at a loud but comfortable level**. Play the calibration sound as many times as you like:

Trial 1/6

Which sound was the softest?

1st sound **2nd** sound **3rd** sound

Additional requirements

Please do NOT take this experiment multiple times, and do NOT reload this page. If you share an MTurk/Prolific account with others who have taken this experiment, please make sure that they have not yet taken this experiment. We cannot use data from reloaded or repeated experiments and won't be able to approve your work.

We use cookies and MTurk/Prolific qualifications to make it easy for you to recognize whether you have taken this experiment previously. If you accept our cookies and do not delete them, this should prevent you from accidentally taking the experiment more than once.

Reasons work can be rejected

If you pay attention to the instructions and **do not respond randomly** your work will be approved. **Please do NOT reload this page, even if you think you made a mistake.** We will not be able to use your data for scientific purposes, and you will not be able to finish the experiment. We anticipate some mistakes will be made, but those will NOT affect the approval of your work.

We will only reject work if you a) **clearly** do not pay attention to the instructions, b) reload the page, or c) repeat the experiment. We reject far less than 1% of all completed experiments.

Experiment instructions

The purpose of this experiment is to investigate listeners' ability to pay attention to a specific talker when there are multiple talkers speaking at once.

The experiment has two parts. In the first part, you will hear recordings of a female and a male talker speaking simultaneously. Your task is to **focus only on the female talker**. For each recording, you have to determine whether the female talker produced a word or a non-word.

In the second part, you will hear recordings from the same two talkers. This time, each recording will only contain speech from one talker at a time.

Informed consent

By accepting this experiment, you confirm that you have read and understand the consent form, that you are willing to participate in this experiment, and that you agree that the data you provide by participating can be used in scientific publications (no identifying information will be published). Sometimes we share non-identifying data elicited from you — including sound files — with other researchers for scientific purposes (your MTurk/Prolific ID will be replaced with an arbitrary alphanumeric code).

Begin the experiment

Once you press the green button, these instructions will disappear, so make sure you understand them fully before you click.

I confirm that I meet the requirements for this experiment, that I have read and understood the instructions and the consent form, and that I want to start the experiment.

Further (optional) information

Sometimes it can happen that technical difficulties cause experimental scripts to freeze so that you will not be able to submit a experiment. We are trying our best to avoid these problems. Should they nevertheless occur, we urge you to (1) take a screen shot of your browser window, (2) if you know how to also take a screen shot of your JavaScript console, and (3) email us this information along with the HIT/Experiment ID and your worker/Prolific ID.

If you are interested in hearing how the experiments you are participating in help us to understand the human brain, feel free to subscribe to our lab blog where we announce new findings. Note that typically about 1-2 years pass before an experiment is published.

APPENDIX D

	Attended Talker Attributes				Presentation Factors					
Version	Gender	Ear	Material	Label	Exposure (	Test Order	Exposure r	Test resp k	Participant	Excluded Participants
Condition.	F	L	A	?s	1	A	0	0	0	
Condition.	F	L	A	?s	1	A	0	1	1	
Condition.	F	L	A	?s	1	A	1	0	1	
Condition.	F	L	A	?s	1	A	1	1	0	
Condition.	F	L	A	?s	1	B	0	0	2	
Condition.	F	L	A	?s	1	B	0	1	1	
Condition.	F	L	A	?s	1	B	1	0	1	
Condition.	F	L	A	?s	1	B	1	1	1	
Condition.	F	L	A	?s	5	A	0	0	1	
Condition.	F	L	A	?s	5	A	0	1	1	
Condition.	F	L	A	?s	5	A	1	0	1	
Condition.	F	L	A	?s	5	A	1	1	2	P. 533
Condition.	F	L	A	?s	5	B	0	0	1	
Condition.	F	L	A	?s	5	B	0	1	1	
Condition.	F	L	A	?s	5	B	1	0	1	
Condition.	F	L	A	?s	5	B	1	1	1	
Condition.	F	L	A	?sh	1	A	0	0	0	
Condition.	F	L	A	?sh	1	A	0	1	1	
Condition.	F	L	A	?sh	1	A	1	0	1	
Condition.	F	L	A	?sh	1	A	1	1	1	
Condition.	F	L	A	?sh	1	B	0	0	1	
Condition.	F	L	A	?sh	1	B	0	1	0	P. 548
Condition.	F	L	A	?sh	1	B	1	0	3	
Condition.	F	L	A	?sh	1	B	1	1	3	
Condition.	F	L	A	?sh	5	A	0	0	0	
Condition.	F	L	A	?sh	5	A	0	1	1	
Condition.	F	L	A	?sh	5	A	1	0	0	
Condition.	F	L	A	?sh	5	A	1	1	0	
Condition.	F	L	A	?sh	5	B	0	0	1	
Condition.	F	L	A	?sh	5	B	0	1	2	
Condition.	F	L	A	?sh	5	B	1	0	0	P. 461
Condition.	F	L	A	?sh	5	B	1	1	1	
Condition.	M	L	A	?s	1	A	0	0	1	
Condition.	M	L	A	?s	1	A	0	1	1	
Condition.	M	L	A	?s	1	A	1	0	1	
Condition.	M	L	A	?s	1	A	1	1	1	
Condition.	M	L	A	?s	1	B	0	0	1	
Condition.	M	L	A	?s	1	B	0	1	0	
Condition.	M	L	A	?s	1	B	1	0	1	
Condition.	M	L	A	?s	1	B	1	1	1	
Condition.	M	L	A	?s	5	A	0	0	2	
Condition.	M	L	A	?s	5	A	0	1	1	
Condition.	M	L	A	?s	5	A	1	0	1	
Condition.	M	L	A	?s	5	A	1	1	1	
Condition.	M	L	A	?s	5	B	0	0	1	
Condition.	M	L	A	?s	5	B	0	1	1	
Condition.	M	L	A	?s	5	B	1	0	1	
Condition.	M	L	A	?s	5	B	1	1	1	
Condition.	M	L	A	?sh	1	A	0	0	1	
Condition.	M	L	A	?sh	1	A	0	1	1	
Condition.	M	L	A	?sh	1	A	1	0	1	
Condition.	M	L	A	?sh	1	A	1	1	1	
Condition.	M	L	A	?sh	1	B	0	0	1	
Condition.	M	L	A	?sh	1	B	0	1	1	
Condition.	M	L	A	?sh	1	B	1	0	1	
Condition.	M	L	A	?sh	1	B	1	1	0	P. 467
Condition.	M	L	A	?sh	5	A	0	0	1	
Condition.	M	L	A	?sh	5	A	0	1	1	
Condition.	M	L	A	?sh	5	A	1	0	1	
Condition.	M	L	A	?sh	5	A	1	1	1	
Condition.	M	L	A	?sh	5	B	0	0	1	
Condition.	M	L	A	?sh	5	B	0	1	1	
Condition.	M	L	A	?sh	5	B	1	0	0	
Condition.	M	L	A	?sh	5	B	1	1	0	