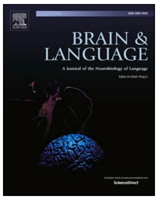
Brain & Language 222 (2021) 105022

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



When it’s not appropriate to adapt: Toddlers’ learning of novel speech patterns is affected by visual information

Katherine S. White [a,*], Olivia Daub [b]

[a] *The University of Waterloo, Canada*
[b] *The University of Western Ontario, Canada*

ARTICLE INFO

Keywords:
Toddler word recognition
Mouth obstruction
Perceptual learning
Accent adaptation
Alternative explanations

ABSTRACT

In adults, perceptual learning for speech is constrained, such that learning of novel pronunciations is less likely to occur if the (e.g., visual) context indicates that they are transient. However, adults have had a lifetime of experience with the types of cues that signal stable vs. transient speech variation. We ask whether visual context affects toddlers’ learning of a novel speech pattern. Across conditions, 19-month-olds (N = 117) were exposed to familiar words either pronounced typically or in a novel, consonant-shifting accent. During exposure, some toddlers heard the accented pronunciations without a face present; others saw a video of the speaker producing the words with a lollipop against her cheek or in her mouth. Toddlers showed the weakest learning of the accent when the speaker had the lollipop in her mouth, suggesting that they treated the lollipop as the cause of the atypical pronunciations. These results demonstrate that toddlers’ adaptation to a novel speech pattern is influenced by extra-linguistic context.

1. Introduction

Our environments are always changing. In some cases, it makes sense to learn the specific patterns of variation; in other cases, it might not. One domain in which this is particularly striking is in the processing of spoken language. Speech is highly variable. Across speakers, a single sound or word may be realized in different ways for many reasons, including the speakers’ ages and accents. Given this variation, accessing the correct word is not a straightforward task. The pronunciation of a word by one speaker might map onto a different meaning for another speaker. For example, one speaker might have a “pet” cat and another a “pit” cat. Or two speakers might use the same word form to express different meanings (e.g., “pit” for the animal vs. a deep hole). Although semantic and non-linguistic context can help resolve these ambiguities, a system that learns about such systematic variation across speakers can process it more efficiently in the future.

A growing body of work has shown that adult listeners cope with this variation in speech at least in part through perceptual learning, or adaptation (Samuel & Kraljic, 2009; Kleinschmidt & Jaeger, 2015). For example, if a speaker produces a *sh*-like sound ([ʃ]) in the word *compass*, instead of *s* ([s]), listeners will adjust their /s/ category representation for this speaker to encompass a broader range of sounds (those that typically lie near the boundary between /s/ and /ʃ/ (Norris, McQueen & Cutler, 2003). Adaptation to a speaker’s productions leads to more efficient future processing of speech from that individual talker or from people with similar ways of speaking. This type of process is thought to contribute to improvements in the recognition of unfamiliarly accented speech observed after exposure (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Maye, Aslin, & Tanenhaus, 2008).

However, adults do not show this kind of adaptation in all situations – and for good reason. Speakers often mispronounce words or transiently produce atypical pronunciations in certain situations (e.g., when they have a cold). Learning that these pronunciations are general characteristics of a speaker and changing category representations as a result would be maladaptive, as it would require subsequent unlearning during future interactions with the speaker. Fortunately, adults consider how strong the evidence is for a novel speech pattern before learning it (Kleinschmidt & Jaeger, 2015). For example, adult listeners use contextual information to judge whether an atypical pronunciation is characteristic of the speaker (i.e., reliable), or incidental (i.e., unreliable), and show more learning in the former case. In situations where there is contextual evidence consistent with an alternative explanation for an atypical pronunciation (such as when a speaker has a pen in her mouth; Kraljic, Samuel, & Brennan, 2008 or when the sound is

[*] Corresponding author at: Psychology Department, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada.
E-mail address: white@uwaterloo.ca (K.S. White).

https://doi.org/10.1016/j.bandl.2021.105022
Received 1 March 2021; Received in revised form 22 August 2021; Accepted 30 August 2021

phonetically motivated; Kraljic, Brennan, & Samuel, 2008), adult listeners do not update their category representations. Part of being an efficient listener, therefore, is using the evidence available to determine whether a novel pattern should be learned.

Like adults, toddlers can adapt to differences in pronunciation across accents, with this ability improving over development (Mulak, Best, Tyler, Kitamura, & Irwin, 2013; Potter & Saffran, 2017; Schmale, Cristia, & Seidl, 2012; Van Heugten, Krieger, & Johnson, 2015). For example, toddlers exposed to an accent in which the vowel /a/ is produced as [ae] (*block* → *black*) later recognize other words when the same speaker uses [ae] in place of what would be [a] in the toddlers' native dialect. In contrast, toddlers not exposed to the accent do not recognize these [ae] pronunciations later (White & Aslin, 2011). Toddlers also show improvements in recognizing words transformed by more complex accents following exposure (Potter & Saffran, 2017; Schmale, Cristia & Seidl, 2012; van Heugten & Johnson, 2014), though what specifically they are learning about the accents in these more complex cases is not known.

Although toddlers can learn a novel speech pattern (like a shift in the realization of /a/ from [a] → /[ae]), it is not clear whether their learning is affected by the strength and type of evidence they receive. Are toddlers, like adults, discerning learners, who take other aspects of the situation into account when evaluating how robust a novel speech pattern is likely to be? Although, to our knowledge, this has not been examined in the speech domain, children have been shown to learn differently depending on the nature of the evidence they are given in other domains. For example, during word learning, property extension, and causal reasoning tasks, children draw different conclusions depending on how much data is provided and how it was generated. In one such study, Xu & Tenenbaum (2007b) presented children with either one or three trials in which they saw referents of a novel word. They were then asked to select other referents of the word. Although children did not make strong assumptions about the extension of a word after a single trial, they did make strong assumptions after three trials. In particular, they made the most conservative hypothesis consistent with the data observed across the three trials (seeGweon, Tenenbaum, & Shulz, 2010 for similar behavior in a physical property extension task). For example, if the label was consistently applied to Dalmatians, children assumed it applied to Dalmatians and not to dogs more generally. These inferences are thought to arise because children assume (unless shown otherwise) that the data are not being sampled randomly and so the narrow range of exemplars is meaningful (Gweon, Tenenbaum, & Shulz, 2010; Xu & Tenenbaum, 2007a). Tasks of physical causal reasoning also reveal the sophisticated ways in which children evaluate different patterns of data. For example, if children are shown that two blocks activate a machine together and that one of those blocks does not activate the machine when it is presented alone, they infer that the other block causes the machine to activate. If they instead see two blocks activate the machine together and later see that one of the blocks does activate the machine alone, they infer that the other block does not (Sobel, Tenenbaum, & Gopnik, 2004). Together, these lines of work demonstrate that children reason about the most likely causes of the data they observe, taking into account how much data there are, how the data were generated, and whether some apparent causes 'explain away' others.

In the present work, we ask whether 19-month-old toddlers use extra-linguistic information to constrain speech learning. More specifically, we ask whether toddlers, like adults, are conservative learners, and will show less learning of a novel pronunciation when the visual context is consistent with an alternative explanation for it (as in Kraljic, Samuel, & Brennan, 2008). For toddlers, does the presence of one potential cause for the atypical pronunciations (a mouth obstruction) rule out another potential cause (a novel accent)?

There are at least two possible reasons why one might expect toddlers to be less conservative in their learning of speech variation than adults. The first is that young learners might not yet have realized that pronunciation changes occurring in certain contexts are unlikely to persist. For example, a child encountering a new individual for the first time may not realize that the peculiar way they are talking is a result of the fact that they have a cold, because they have not yet learned that colds alter a speaker's nasality. If the child learns (erroneously) that nasality is a general feature of that individual's speech, then they may have difficulty understanding that individual the next time they encounter them. Rather than learning that nasality is a general feature of the individual's speech, it would be better either to simply be more tolerant of the speaker's atypical pronunciations in the moment (listen "through" the pronunciations by relaxing the criteria for word recognition) or to learn the novel pronunciations, but link them to potential conditioning contexts (like the speaker's red nose), so that this knowledge can be applied in the future in the same contexts. Based on previous work, it is not clear which of the latter two approaches adult listeners take when they encounter a speaker talking with a pen in her mouth. However, they do not appear to learn that the novel pronunciations are generally characteristic of the speaker (Kraljic, Samuel & Brennan, 2008), likely because they are aware that mouth obstructions can alter a person's speech. If toddlers are unaware of the relationship between mouth obstructions and atypical productions, then they should perform differently than adults.

A second reason that we might expect toddlers to be less conservative learners of speech variation than adults is that their phonological systems are more flexible overall - their native language speech categories (and the link between those speech categories and the lexicon) are not as well established. Indeed, in both the lab and the real world, infants and young children appear to learn novel speech categories and patterns faster and more readily than adults, who may not be successful at all. For example, although infants show distributional learning of speech categories after only 2 minutes of exposure (Maye, Werker, Gerken, 2002), studies with adults use longer exposure periods (Chladkova & Simackova, 2021; Hayes-Harb, 2007; Maye & Gerken 2000) and sometimes fail to demonstrate learning even after these longer exposures (e.g., Wanrooij, Boersma, & van Zuijen, 2014). In the real world, children are more likely to acquire a new community accent than adults are. Therefore, it is entirely possible that young learners will learn novel speech patterns even in cases where adults do not. That said, a conservative learning strategy (in which the data and potential causes are evaluated) would seem to be particularly adaptive for toddlers, to prevent them from learning unreliable or transient patterns as they are building up knowledge of the native language. Adults, in contrast, could in principle afford to be less conservative, because small amounts of data should not cause significant changes in their representations.

To ask whether toddlers' learning of novel speech patterns is conservative, we tested toddlers in one of four (between-subject) conditions. In two of these conditions, the Characteristic and Incidental conditions, toddlers heard a consonant-shifting accent during exposure and saw a video of the speaker producing the accented words. In the Characteristic condition, the speaker held a lollipop to her cheek while she produced the words. In the Incidental condition, the lollipop was in her mouth during the pronunciation of the exposure words. Therefore, in the Incidental condition, toddlers had the type of contextual information (a lollipop in the mouth) that could indicate that the novel pronunciations were not necessarily characteristic of the speaker. In other words, the lollipop served as a potential alternative cause for the atypical pronunciations (rather than the speaker having a different phono-lexical system). Importantly, use of this contextual information to constrain learning in only the Incidental condition would require that toddlers understand that it is specifically a mouth obstruction (but not an object on the face) that can cause changes in speech productions.

If, like adult listeners, toddlers learn conservatively (and understand that mouth obstructions can cause pronunciation changes), then we predict that toddlers in the Characteristic condition will show learning of the accent, but that toddlers in the Incidental condition will not. If, on the other hand, learning is driven by the acoustic information alone, then we expect equivalent learning of the accent in these two conditions. We also included two additional conditions. The first was a No-face condition, to establish that toddlers could learn this type of consonant-shifting accent in the absence of visual information about the speaker. Previous work has demonstrated only that English-learning toddlers can learn the specifics of novel accents involving vowel shifts. Finally, a Control condition, in which toddlers heard only standard pronunciations during exposure, was included to examine toddlers' treatment of the accented test pronunciations in the absence of previous exposure to these pronunciations.

2. Methods

2.1. Participants

One hundred seventeen monolingual, English-learning toddlers between the ages of 18 – 20 months old ($M = 574$ days, $SD = 17$) were randomly assigned to one of four exposure conditions: accent in the absence of face information ("No-face"; $n = 25$), characteristic accent ("Characteristic"; $n = 30$), incidental accent ("Incidental"; $n = 31$), and no accent[1] control ("Control"; $n = 31$). Toddlers had a minimum of 90% exposure to English, as indicated by parental report. An additional 40 participants were tested, but their data were not included due to prematurity (1), vision issues (1), insufficient English exposure (2), parent interference (2), coding difficulty (5), equipment issues during recording or coding (8), and early termination due to fussiness/crying (21).

2.2. Design and stimuli

Toddlers were tested using the intermodal preferential looking paradigm, in which they were presented with visual stimuli and their looking behavior in response to a simultaneously presented audio stimulus was measured. All toddlers participated in both an exposure phase and a test phase. The nature of the exposure phase differed across conditions. The test phase was the same across conditions.

2.2.1. Exposure phase

During exposure, toddlers heard isolated words labelling three familiar objects. Depending on the counterbalancing condition, the labelled objects in exposure were either *brush*, *bear*, and *bottle* or *bunny*, *book*, and *balloon*. The remaining three objects were also presented during exposure, but were not labelled. The standard labels for all six objects begin with /b/ and were deemed to be highly familiar to our age group based on lexical norms (Dale & Fenson, 1996). A questionnaire given to parents after the study confirmed this. On a scale of 1 (not visually familiar, label unknown), 2 (visually familiar, label unknown), 3 (visually familiar, label familiar), and 4 (visually familiar, label highly familiar), the target words received a score of 3.54 (and this was similar across conditions)[2]. During exposure, toddlers in the No-face, Characteristic and Incidental conditions all heard the objects labelled with [d] onsets (i.e., accented), e.g., *dunny*; toddlers in the Control condition heard them labelled with typical [b] onsets (i.e., unaccented).

The auditory stimuli ([d]-initial pronunciations in the No-face, Characteristic, and Incidental conditions and [b]-initial pronunciations in the Control condition) were recorded by a female native speaker of English speaking in an infant-directed manner. These recordings were inserted into the exposure videos below (the speaker recorded the auditory stimuli while watching the video recordings, to ensure matching timing). The stimuli were adjusted in Praat (Boersma & Weenink, 2014) to be of equal perceived volume. The same auditory stimuli were used in the exposure phase for the No-face, Characteristic, and Incidental conditions.

In the No-face condition, toddlers were exposed to the words and their corresponding referents in the same way as in White & Aslin (2011), to determine whether we would find learning of a novel consonant accent under the same conditions previously used to test the learning of a vowel accent. This was done to ensure that, if toddlers failed to learn the accent in the Characteristic and Incidental conditions, this failure would not be attributable to our use of a consonant accent. In this condition, the six objects were arranged into three exposure displays (see Fig. 1 for a sample display). Each display contained four simultaneously presented objects (display 1: **brush**, **bear**, bunny, book; display 2: **brush**, **bottle**, book, balloon; display 3: **bottle**, balloon, **bear**, bunny). Every two seconds, one of the objects loomed (got larger and smaller) and was either labeled or was highlighted with "ooo" or "look". Half of the toddlers heard the bolded objects labelled (4 times each per display) and the other objects accompanied by "ooo/ look" (1 time each per display). The other half heard the reverse (bolded objects accompanied by "ooo/ look" and the other objects labelled). Therefore, across displays, participants heard 8 repetitions of the label for each labelled object. [3]The three displays were presented in random order across participants. See White & Aslin (2011) for more details about the exposure in this condition.

In the three visible speaker conditions (the Characteristic, Incidental, and Control conditions), toddlers were presented with a picture of a single object on each exposure trial and then a video of the speaker producing the associated label or "ooo/look". The object appeared alone at the bottom of the screen for 1500 ms prior to the appearance of the speaker and remained on the screen for 1500 ms after the disappearance of the speaker. The image of the speaker occupied visual angles of approximately 14 degrees vertically and 6 degrees horizontally. In the Characteristic condition, the speaker brought the lollipop to her face and rested it on her cheek prior to the naming of the object (see Fig. 1). In the Incidental and Control conditions, she instead brought the lollipop to her face and placed it in her mouth prior to and during the naming of the object. Therefore, in the Incidental condition, we provided toddlers with contextual information (a lollipop in the mouth) that could indicate that the novel pronunciations were not necessarily characteristic of the speaker. As in the No-face condition, three items were labeled during exposure (eight times total) and three were instead highlighted with "ooo" or "look" (twice total). The sets of items that were labeled/unlabeled were counterbalanced across participants as described for the No-face condition.

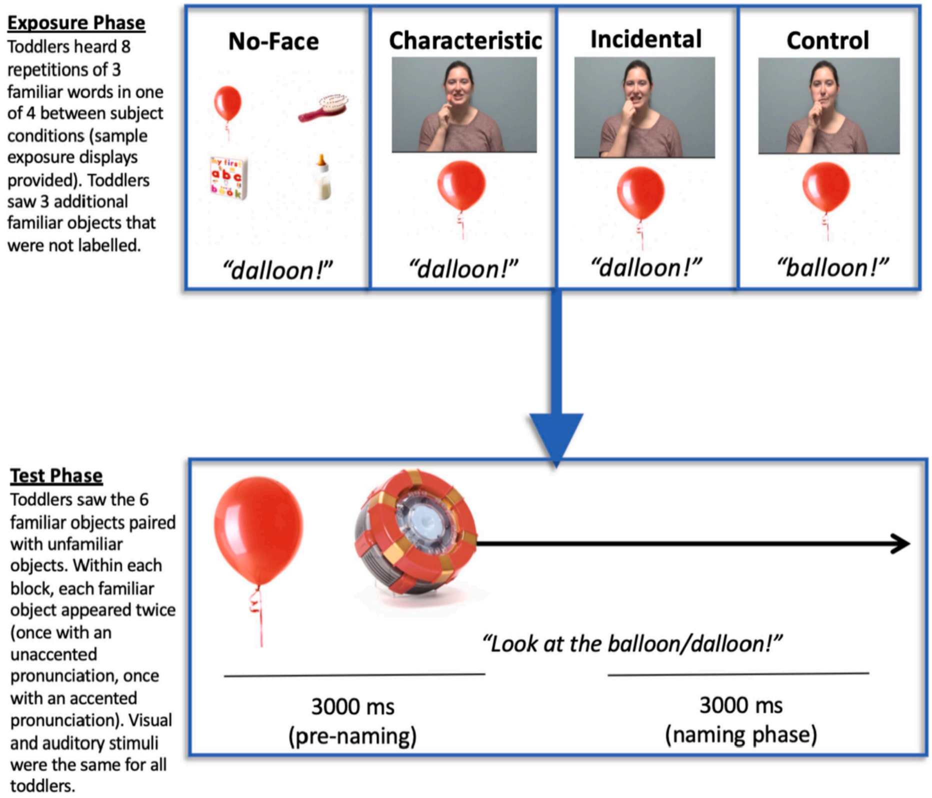
2.2.2. Test phase

All toddlers completed the same set of test trials. During each trial, two images appeared, each on one side of the screen. One of the images was an image from exposure (either previously labeled or unlabeled), the other an unfamiliar object (e.g., a vegetable spiralizer). Pairings of particular familiar and novel items remained fixed across trials. Unfamiliarity with novel objects was confirmed via our parent questionnaire. The first three seconds constituted a silent pre-naming phase. Three seconds after the images appeared, the same voice from the exposure trials labeled the familiar object in one of two sentence frames (the naming phase; "*Find the___*" and "*Look at the___*"). See Fig. 1 for a

[1] We recognize that there is no such thing as "unaccented" speech or speech with no accent. We use the terms unaccented and accented as shorthand to refer to familiarly accented and unfamiliarly accented speech, respectively.

[2] These means do not include the final 21 participants whose data were included, as these questionnaires are not currently accessible due to the COVID shutdown.

[3] In contrast to White & Aslin (2011), only a single token of each word type was used during exposure. Although this could have limited learning (more variability can lead to more robust representations and generalizations; Rost & McMurray, 2009), it did not. This suggests that it is the presence of multiple word types, not tokens, exhibiting the variation that is critical for accent learning.


Exposure Phase

Toddlers heard 8 repetitions of 3 familiar words in one of 4 between subject conditions (sample exposure displays provided). Toddlers saw 3 additional familiar objects that were not labelled.

Test Phase

Toddlers saw the 6 familiar objects paired with unfamiliar objects. Within each block, each familiar object appeared twice (once with an unaccented pronunciation, once with an accented pronunciation). Visual and auditory stimuli were the same for all toddlers.

Fig. 1. Schematic of exposure and test phase trials.

timeline of the test trials.

Toddlers in all conditions heard the labels for all six exposure objects, twice in unaccented form ([b]) and twice in accented form ([d]) (once in each sentence frame), for a total of 24 critical trials. The same auditory stimuli were used in all conditions[4]. Between trials, a wordless cartoon clip was played to ensure the child was looking at the screen prior to the onset of the next trial. The subsequent test trial was not initiated by the experimenter until the toddler's gaze was directed towards the screen. Across the four times each object pair appeared, the target object appeared half the time on each side of the screen (once per pronunciation type). There were no more than three targets on the same side of the screen in a row or more than three of a given pronunciation type or sentence frame in a row. Toddlers were randomly assigned to one of four randomization lists.

2.3. Procedure

Toddlers were tested in a sound attenuated room. The stimuli were presented on a 42-inch widescreen television and via two hidden speakers located at the base of the television; both were connected to a computer in an adjacent room. The child sat on the parent's lap approximately 80 cm from the television. During the session, the parent listened to music over headphones. The participants were monitored over a closed-circuit video feed that was recorded for later off-line coding; the camera was centrally located beneath the television and hidden behind a black curtain. The session lasted approximately 5 min in the No-face condition and 8 min in the other three conditions (because of the additional time added for the face presentations). Following the testing session, parents completed a questionnaire on their child's comprehension and production of the experimental items (both familiar and novel).

2.4. Analysis

Looking behavior was coded off-line by coders blind to the audio and condition using customized software at a rate of 30 frames per second (~33.33 ms/frame). For each trial, the proportion of time toddlers spent looking at the familiar object (out of the total time spent looking at the two objects) was calculated for both the 3-second pre-naming and 3-second naming phases. The naming phase was shifted to commence 300 ms after the onset of the target word, as per previous work with this paradigm and age group (White & Aslin, 2011). A difference score was computed (naming minus pre-naming), which indicates how much toddlers changed their looking to the labeled object from pre-naming to naming. A positive score indicates that toddlers increased their looking to the labeled object following naming, while a negative score indicates that they decreased their looking to the labeled object following naming. Trials in which toddlers did not look at the two objects for a combined 500 ms in each phase were eliminated from the analysis. In addition, because difference scores assessed a change from pre-naming to naming, trials in which both objects were not fixated during the pre-naming phase were also excluded. These criteria led to the exclusion of a total of 344 critical trials across all participants (12.3% of trials; similarly distributed across conditions).

3. Results

We first computed looking proportions for the pre-naming phase. The mean proportions of looking at the familiar target during pre-naming were 0.491 (*SD* = 0.051), 0.544 (*SD* = 0.057), 0.531 (*SD* = 0.055), and 0.545 (*SD* = 0.05) for the No-face, Characteristic, Incidental, and Control conditions, respectively. A one-way ANOVA found that there was a significant difference in pre-naming looking across

[4] Toddlers in the three speaker conditions also received an additional six filler trials, with labels that did not start with [b] or [d], to help maintain interest given the increased length of the exposure.

conditions, $F(3, 116) = 5.948$, $p = .001$, $\eta^2 = 0.136$. In particular, the average pre-naming proportion for the No-face condition was at chance ($t(24) = -0.889$, $p = .383$), whereas the pre-naming proportions were above chance for the other three conditions (Characteristic: $t(29) = 4.216$, $p < .001$, $d = 0.77$; Incidental: $t(30) = 3.132$, $p = .004$, $d = 0.563$; Control: $t(30) = 5.029$, $p < .001$, $d = 0.903$). A similar bias for name-known objects in this type of display has been reported previously (e.g., Schafer, Plunkett, & Harris, 1999; White & Morgan, 2008). Importantly, given the similarity across the three visual speaker conditions, any differences in the results cannot be attributed to differences during the pre-naming phase.

The analyses of interest were based on the difference in the proportion of looks to the familiar object between the pre-naming and naming phases. Inspection of the timecourse in the No-Face condition indicated that toddlers had returned to baseline levels of looking by 2500 ms post-target onset during the naming phase. Therefore, we used this naming period for our analyses (all conditions). See Table 1 and Fig. 2 for mean naming-pre-naming difference scores.

Although participants completed two blocks of 12 test trials, we report only the results of the first block of trials in the main text. When data collection began, we had one previous study demonstrating that condition differences (between toddlers who were and were not exposed to the accent) decreased in the second block of test trials (White & Aslin, 2011). This occurred because toddlers not exposed to the accent improved on their recognition of accented words during the test phase itself. The results of another study, conducted in parallel with the current study, then showed a similar pattern (Weatherhead & White, 2018). Therefore, we recommend going forward that studies utilizing this paradigm use only a single block of test trials. That said, the patterns remained similar when all trials were included (see Appendix).

3.1. *Between-condition comparisons*

We first conducted a repeated measures ANOVA on these difference scores, with Condition (No-Face, Characteristic, Incidental, and Control) as a between-subjects factor and Pronunciation Type (Unaccented, Accented) as a within-subjects factor. This ANOVA revealed a significant effect of Pronunciation Type, $F(1,113) = 11.893$, $p < .001$, $\eta^2 = 0.095$. There was no main effect of Condition, $F(3, 113) = 1.515$, $p = .215$ and no interaction of Condition × Pronunciation Type, $F(3, 113) = 1.998$, $p = .118$.

To more directly test our predictions, we then compared how toddlers treated the two types of pronunciations (unaccented and accented) across conditions. In this analysis, we subtracted the naming-prenaming difference scores for the accented pronunciations from the naming-prenaming difference scores for the unaccented pronunciations. In other words, we analyzed a difference of a difference (the change from baseline to test for unaccented vs. accented trials). These unaccented-accented scores are presented in Fig. 3. A smaller difference here indicates that accented pronunciations are being treated more similarly to unaccented pronunciations (conversely, a larger difference indicates a penalty for the accented pronunciations relative to unaccented pronunciations).

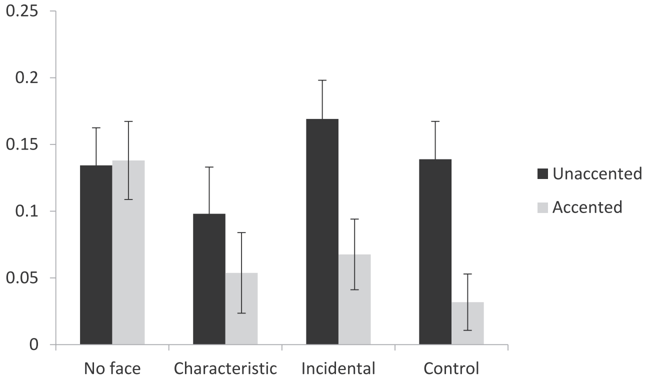


Fig. 2. Difference between looks to the labeled object in the naming phase vs. the pre-naming phase for unaccented and accented test pronunciations by condition (Block 1).

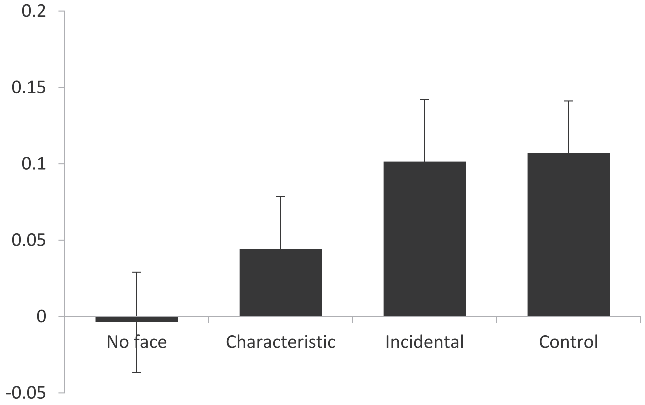


Fig. 3. Differences between unaccented and accented difference scores.

Table 1
Mean difference scores (standard deviations) by condition and word type for Block 1.

	Unaccented	Accented	Word type Difference
No-Face	0.134 (0.141)	0.138 (0.146)	-0.004 (0.164)
Characteristic	0.098 (0.192)	0.054 (0.165)	0.044 (0.187)
Incidental	0.169 (0.161)	0.068 (0.147)	0.101 (0.227)
Video control	0.139 (0.158)	0.032 (0.118)	0.107 (0.189)

Word type difference is the difference between scores for unaccented and accented pronunciations.

We predicted that toddlers in the No-face condition would show recognition of the accented words (and therefore, the smallest difference between unaccented and accented words). In contrast, we predicted that toddlers in the Control condition, who were not previously exposed to the accent, would show poor recognition of the accented words (and, therefore, the largest difference between unaccented and accented words). We were particularly interested in what would happen in the Characteristic and Incidental conditions relative to these two endpoints. If toddlers interpreted the mouth obstruction (but not the lollipop on the cheek) as an alternative explanation for the speech variation, then performance in the Incidental condition should be similar to the condition in which they were not exposed to the accent at all (the Control condition), whereas performance in the Characteristic condition should resemble that of the No-face condition. In our statistical analyses, rather than compare all of the conditions against one another (increasing the number of comparisons), we opted to compare all of our conditions to a single endpoint condition. We chose the No-face condition, because this condition is set up exactly the same as a previous experiment done with a vowel change (White & Aslin, 2011). Pairwise comparisons against the No-face condition revealed a significant difference between the No-face and Control conditions ($t(54) = 2.309$, $p = .025$, $d = 0.621$), a marginal difference between the No-face and Incidental conditions ($t(54) = 1.945$, $p = .057$, $d = 0.523$), and no difference between the No-face condition and the Characteristic condition ($t(53) = 1$, $p = .321$). The significant difference between the No-Face and Control conditions shows the effect of accent exposure: toddlers in the No-Face condition had heard the accented pronunciations prior to test, whereas toddlers in the Control condition had not. However, toddlers in both the Incidental and Characteristic conditions heard the same accented pronunciations in exposure as toddlers in the No-Face condition, yet they behaved differently. The pattern of results across the four conditions (Fig. 3) shows the smallest penalties for the accented pronunciations in the No-face and Characteristic conditions and larger penalties in the Incidental and Control conditions.

3.2. Within-condition analyses

We then compared the unaccented-accented difference scores for each condition to zero. These comparisons directly test whether toddlers mapped the accented pronunciations to the familiar objects as reliably as they did for the unaccented pronunciations. If a toddler treats unaccented and accented pronunciations the same (as might be expected if they have learned the accent), the difference between them should be small (a difference of 0 would indicate equivalent performance for the two pronunciation types). In contrast, if a toddler shows stronger recognition of the unaccented pronunciations, the difference will be larger (as might be the case if they did not learn the accent, or did not learn it as well). For toddlers in the No-face condition, there was no difference between the unaccented and accented pronunciations, $t(24) = -0.113$, $p = .911$. Therefore, toddlers in this condition (who had been exposed to [d] pronunciations prior to test) recognized the accented pronunciations as well as they recognized the unaccented pronunciations. The same was true of toddlers in the Characteristic condition: they did not demonstrate a difference in their recognition of unaccented and accented pronunciations, $t(29) = 1.296$, $p = .205$, suggesting that there was learning of the [d] pronunciations in this group as well.

In contrast, toddlers in the Incidental and Control conditions demonstrated significant penalties (less of an increase in looking to the target during the naming phase) for the accented pronunciations at test. The difference between unaccented and accented difference scores was significantly greater than 0, Incidental: $t(30) = 2.492$, $p = .018$, $d = 0.448$ and Control: $t(30) = 3.147$, $p = .004$, $d = 0.565$.

These results demonstrate that toddlers in the Incidental and Control conditions were disrupted by the accented pronunciations at test. For the Control condition, this is unsurprising, as the accented pronunciations were not heard during exposure. However, toddlers in the Incidental condition heard the same [d] pronunciations during exposure as the toddlers in the Characteristic and No-face conditions. Therefore, reduced recognition in the Incidental condition cannot be attributed to the acoustic information presented. This suggests that the presence of the lollipop in the mouth reduced their learning of these pronunciations.

4. Discussion

We learn novel speech patterns through exposure to auditory input. The present study explored whether toddlers' learning of such patterns is based entirely on acoustic information or whether non-linguistic contextual information can influence the learning process. Toddlers were exposed to a novel speech pattern (a consonant-shifting accent) either without or with a visible speaker. Toddlers learned the novel accent (later recognized the accented pronunciations) when there was no visible speaker, and to a lesser degree when the visible speaker had nothing in her mouth. However, when the speaker had a lollipop in her mouth, toddlers showed even less learning. In fact, toddlers' behavior in this condition was strikingly similar to their behavior in the Control condition, where toddlers had not heard the accented pronunciations before test. Overall, toddlers' learning (where learning is reflected by more similar behavior for unaccented and accented pronunciations) depended on whether the evidence suggested that the novel speech pattern was characteristic of the speaker: toddlers showed better recognition of accented words in the No-face condition and Characteristic conditions than in the Incidental and Control conditions (where there was an alternative explanation for the pronunciation shift, and no prior exposure to the shift at all, respectively). These results are consistent with approaches that characterize perceptual learning as an inference process that takes into account prior beliefs and the strength of the evidence in favor of updating (Kleinschmidt & Jaeger, 2015).

Previous work has demonstrated that infants' and children's learning of various types of information (words, category information, grammatical patterns) is affected by the nature of the evidence they observe. For example, Gerken (2006) found that infants will make the most conservative generalization about syllable sequences that is consistent with the data, and Xu & Tenenbaum (2007b) found that children were affected by the scope of the data during word learning – but only if it was provided by a knowledgeable teacher (see also Gweon et al., 2010). The current results show, in the domain of perceptual learning for speech, that toddlers may integrate information from various sources to determine the potential robustness of a pattern, and show more learning (updating) when the evidence is strong. In the present situation, increased support for one explanation of the data - a temporary disturbance due to the lollipop - weakened the evidence for another explanation - a characteristic feature of the speaker. This, in turn, led to less learning.

Of additional interest, the present results demonstrate that English-learning toddlers can learn a novel consonant shift as efficiently as a novel vowel shift. Previous work on toddlers' learning of specific accent shifts has focused on vowels (Newman, Morini, Kozlovsky, & Panza, 2018; Weatherhead & White, 2016; White & Aslin, 2011). A consonant place change was chosen here, because it is a plausible consequence of an obstruction in the mouth. Consonant shifts could be harder to learn for a variety of reasons. Vowels tend to vary considerably across dialects and accents (Labov, Ash, & Boberg, 2006), and in some learning situations, toddlers pay more attention to consonants than vowels (Nazzi, Poltrock, & Von Holzen, 2016). It is beyond the scope of the present work to evaluate whether English-learning toddlers are more (or less) sensitive to vowel vs. consonant variation. However, because toddlers in our No-Face condition learned an accent centered around a consonant shift, the weaker adaptation in the Incidental accent condition cannot be attributed to a general difficulty with consonant shifts.

Beyond the primary finding of interest (that toddlers attend to extra-linguistic information during speech learning), our results also suggest that toddlers may understand the link between mouth obstruction and potential changes in speech production. An alternative possibility is that toddlers simply refrain from learning about speech during marked, or atypical, situations (people do not typically speak with something in their mouths), and wait until the situation returns to normal. However, a lollipop placed on the cheek is also an atypical situation, yet toddlers appear to have learned the accent in this condition. That said, learning in the Characteristic (cheek) condition was numerically weaker than the learning in the No-Face condition. Whether this means that the simple presence of a lollipop led toddlers to wait for a return to normal, whether it introduced confusion, or whether the data are less reliable in this condition is unclear (although there was no significant difference in the number of trials per condition, more trials were lost on average per participant in the Characteristic condition – 3.6 vs. 2.7, 2.8, and 2.6 in the other conditions).

If toddlers do understand the relationship between mouth obstruction and speech quality, an open question is whether they have a more nuanced understanding of how particular contextual cues relate to particular types of acoustic variation. We presented participants with a salient visual cue to oral obstruction (a lollipop) and a consonant shift that was plausible given the nature of the obstruction (a change in place of articulation). What would have happened if a less plausible change to result from this kind of obstruction (e.g., in voicing) had co-occurred with the lollipop? Would toddlers refrain from learning a voicing change in this situation? What if the visual cue was a red nose, which might indicate a cold (and therefore, increased nasalization)? Would toddlers learn a place or voicing change in this situation, when it is less plausibly linked to the extra-linguistic cue? To our knowledge, these kinds of nuanced inferences have not been explored in adults, either. Such manipulations of plausibility will not only clarify how extra-linguistic information influences speech learning, but could also be a way of probing toddlers' knowledge of the specific links between acoustic variation and articulatory cues (Best, Goldstein, Nam & Tyler, 2016). The question will then be to determine when and how such knowledge is acquired – is it by accumulating data about articulatory causes and acoustic consequences through observation or through their own productions? Or is knowledge about these sorts of articulatory-acoustic relationships available early in development, even prior to experience? Even very young infants have access to basic cross-modal relationships between articulation and vowel quality (Kuhl & Meltzoff, 1984), and it is being increasingly recognized that speech perception is linked to both visual articulatory and sensorimotor information early in development (Kuhl, Ramirez, Bossler, Lin, & Imada, 2014; Yeung & Werker, 2013).

To summarize, in the present study, we demonstrate that toddlers' adaptation to a novel accent is not based on the acoustic signal alone. Instead, toddlers' learning is affected by the context in which the novel accent is presented. Our findings suggest that toddlers may be sensitive to contextual information that can cue stable vs. transient speech variation. Given the plasticity of the early speech system, such selective learning of stable patterns would be highly adaptive, preventing children from learning spurious patterns at an influential period of development.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Ashley Blayney-Hoffer, Joel LeForestier, Emily McIntosh, Shaquille Sealy, and Erin Kim for help with recruitment and testing, as well as Carragh Erhardt for appearing in the stimulus videos, and Ayah Taji, Chloe Whitham, Catherine Rocha, Katelyn Rowe, Laura Brooks, Justin Leung, and Kristen Thompson for assistance with data coding. We also thank the many families who volunteered their time to participate. This work was funded by an operating Grant from the Natural Sciences and Engineering Research Council of Canada to KSW.

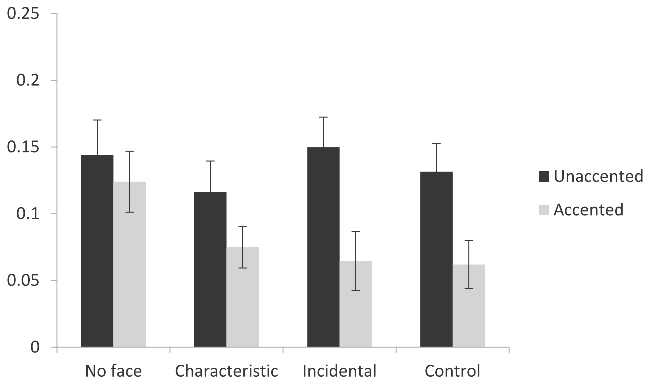
Appendix:. Overall results

Below we provide the results for both blocks of testing. The pattern of results is largely consistent with those provided in the text for Block 1 alone.

Mean difference scores (standard deviations) by condition and word type for both blocks

	Unaccented	Accented	Word type Difference
No Face	0.144 (0.131)	0.124 (0.114)	0.02 (0.137)
Characteristic	0.116 (0.127)	0.075 (0.086)	0.041 (0.122)
Incidental	0.15 (0.126)	0.065 (0.123)	0.085 (0.162)
Video control	0.131 (0.118))	0.062 (0.1)	0.07 (0.134)

Word type difference is the difference between scores for unaccented and accented pronunciations.

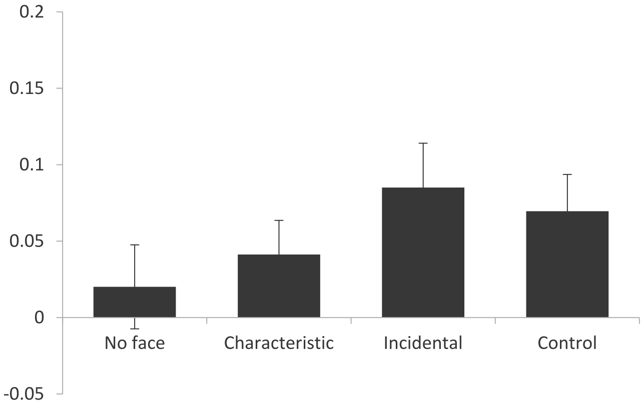


Difference scores for unaccented and accented pronunciation types by condition.

Between-condition comparisons

A repeated measures ANOVA on the difference scores with Condition (No-Face, Characteristic, Incidental, and Control) as a between-subjects factor and Pronunciation Type (Unaccented, Accented) as a within-subjects factor revealed a significant effect of Pronunciation Type, $F(1,113) = 17.330$, $p < .001$, $\eta^2 = 0.133$, but no effect of Condition, $F(3, 113) = 0.985$, $p = .403$ or interaction, $F(3, 113) = 1.204$, $p = .312$.

To test whether there was better performance on accented pronunciations in the No-face and Characteristic conditions than in the Incidental and Control conditions, we compared the conditions using the unaccented-accented difference scores. None of the individual pairwise comparisons were significant (including No-Face vs. Control: $t(54) = 1.358$, $p = .18$ and No-Face vs. Incidental: $\text{t}(54) = 1.596$, $p = .116$, comparisons that revealed a difference for the first block of trials).



Differences between unaccented and accented difference scores by condition.

Within-condition analyses

No-face condition There was no difference between the unaccented and accented pronunciations overall (t(24) = 0.732, p = .471), or in either block individually (b1: t(24) = -0.113, p = .911; b2: t(24) = 0.976, p = .339). In addition, both the unaccented and accented pronunciations were significantly above chance overall (t(24) = 5.518, p < .001, d = 1.104 and t(24) = 5.429, p < .001, d = 1.086) and for each block individually (b1: t(24) = 4.765, p < .001, d = 0.953 and t(24) = 4.716, p < .001, d = 0.943; b2: t(24) = 4.628, p < .001, d = 0.926, and t(24) = 4.133, p < .001, d = 0.827).

Characteristic condition There was no difference between unaccented and accented pronunciations in either block, although there was a marginal difference overall (t(29) = 1.853, p = .074, d = 0.338; b1: t(29) = 1.296, p = .205; b2: t(29) = 0.722, p = .476). In addition, both pronunciation types were recognized significantly above chance overall (t(29) = 5.007, p < .001, d = 0.914 and t(29) = 4.793, p < .001, d = 0.875) and in the individual blocks, with the exception of the accented pronunciations in block 1 (b1: t(29) = 2.801, p = .009, d = 0.511 and t(29) = 1.779, p = .086, d = 0.325; b2: t(29) = 3.077, p = .005, d = 0.562 and t(29) = 5.253, p < .001, d = 0.959).

Incidental condition The two pronunciation types were significantly different overall (t(30) = 2.929, p = .006, d = 0.526) and in both blocks individually (b1: t(30) = 2.492, p = .018, d = 0.448; b2: t(30) = 2.416, p = .022, d = 0.434). Both the unaccented and accented pronunciations were significantly above chance overall (t(30) = 6.611, p < .001, d = 1.187 and t(30) = 2.927, p = .006, d = 0.526) and for each block individually (b1: t(30) = 5.833, p < .001, d = 1.048 and t(30) = 2.554, p = .016, d = 0.459; b2: t(30) = 5.065, p < .001, d = 0.91, and t(30) = 2.090, p = .045, d = 0.375).

Control condition There was a significant difference between the unaccented and accented pronunciations overall (t(30) = 2.892, p = .007, d = 0.519) and in block 1, though it was no longer significant by block 2 (b1: t(30) = 3.147, p = .004, d = 0.565; b2: t(30) = 1.205, p = .238). Difference scores were significantly different from chance for both pronunciation types overall (t(30) = 6.219, p < .001, d = 1.117 and t(30) = 3.434, p = .002, d = 0.617), for the unaccented pronunciations in block 1 (t(30) = 4.9, p < .001, d = 0.88) and block 2 (t(30) = 4.809, p < .001, d = 0.864), and for the accented pronunciations in block 2 (t(30) = 2.483, p = .019, d = 0.446). However, this was not true of the accented pronunciations in block 1 (t(30) = 1.507, p=.142).

References

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology, 28*(4), 216–261.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [computer program]. *Retrieved from.* http://www.praat.org/.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729.

Chládková, K., & Šimáčková, Š. (2021). Distributional learning of speech sounds: An exploratory study into the effects of prior language experience. *Language Learning, 71*(1), 131–161.

Clarke, C. M., & Garrett, M. (2004). Rapid adaptation to foreign accented speech. *Journal of the Acoustical Society of America, 116*, 3647–3658.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*(1), 125–127.

Gerken, LouAnn (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*(3), B67–B74.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences, 107*(20), 9066–9071.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research, 23*(1), 65–94.

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.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*(1), 54–81.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19*(4), 332–338.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361–381.

Kuhl, P. K., Ramirez, R. R., Bosseler, A., Lin, J.-F.- L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences, 111*(31), 11238–11245.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English*. New York: Mouton de Gruyter.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckued wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*, 543–562.

Maye, J., & Gerken, L. (2000). In *Learning phonemes without minimal pairs* (pp. 522–533). Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. F., & Gerken, LouAnn (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.

Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. (2013). Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development, 84*(6), 2064–2078.

Nazzi, T., Poltrock, S., & Von Holzen, K. (2016). The developmental origins of the consonant bias in lexical processing. *Current Directions in Psychological Science, 25*(4), 291–296.

Newman, R. S., Morini, G., Kozlovsky, P., & Panza, S. (2018). Foreign accent and toddlers' word learning: The effect of phonological contrast. *Language Learning and Development, 14*(2), 97–112.

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

Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants' understanding of novel accents. *Cognition, 166*, 67–72.

Rost, G. C. & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. Developmental Science, 12, 339-349.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*(6), 1207–1218.

Schafer, G., Plunkett, K., & Harris, P. L. (1999). What's in a name? Lexical knowledge drives infants' visual preferences in the absence of referential input. *Developmental Science, 2*(2), 187–194.

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science, 15*, 732–738.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303–333.

van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General, 143*(1), 340–350.

van Heugten, M., Krieger, D. R., & Johnson, E. K. (2015). The Developmental trajectory of toddlers' comprehension of unfamiliar regional accents. *Language Learning and Development, 11*(1), 41–65.

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PLOSOne, 9*(10), Article e109806.

Weatherhead, D., & White, K. S. (2016). He says potato, she says potahto: Young infants track talker-specific accents. *Language Learning and Development, 12*(1), 92–103.

Weatherhead, D., & White, K. S. (2018). And then I saw her race: Race-based expectations affect infants' word processing. *Cognition, 177*, 87–97.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. Developmental Science, 14, 372–384.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language, 59*(1), 114–132.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10*(3), 288–297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5), 603–612.