**Title**
Preschoolers Have Difficulty Discriminating Novel Minimal-Pair Words.

**Journal**
Journal of Speech Language and Hearing Research, 65(7)

**ISSN**
1092-4388

**Author**
Creel, Sarah C

**Publication Date**
2022-07-18

**DOI**
10.1044/2022_jslhr-22-00029

# Preschoolers have difficulty discriminating novel minimal-pair words

Sarah C. Creel (screel@ucsd.edu)
UC San Diego
Department of Cognitive Science
UCSD/SDSU Joint Doctoral Program in Language and Communicative Disorders

## Abstract

**Purpose**: The primary aim was to assess whether children have difficulty distinguishing similar-sounding novel words. A secondary aim was to assess what task characteristics might hinder or facilitate perceptual discrimination.

**Method**: Three within-subjects experiments tested 99 3- to 5-year-old children total. Experiment 1 presented two cartoon characters each saying a novel word. Children were asked to report whether they said the same word or different words. Words were identical (e.g., deev/deev), dissimilar (deev/vush), differed in onset consonant voicing (deev/teev), or differed in vowel tenseness (deev/div). Experiment 2 added accuracy feedback after each trial to remind children of task instructions. Experiment 3 interspersed many "same" trials containing a repeating standard word to assess the role of bottom-up stimulus support on difference detection.

**Results**: The *d'* scores were highest for dissimilar words, next highest on different-vowel pairs, and lowest on different-consonant pairs. Performance was better with repeated standard stimuli (Experiment 3) than without (Experiment 1). Benefits for repeated task instructions (Experiment 2) were marginal. Exploratory analyses comparing current results to findings in a word-learning study using the same stimuli suggest an imperfect match to how easily children can learn similar-sounding words.

**Conclusions**: Overall, similar-sounding novel words are challenging for children to discriminate perceptually, though discrimination scores exceeded chance for all levels of similarity. Clinically speaking, same/different tests may be less sensitive to sound discrimination than change/no-change tests.

A critical achievement in learning a language is being able to tell apart fine gradations of sound that indicate differences in meaning in a language, that is, *minimal pairs*, such as the difference between /b/ and /p/ sounds in English. But how does the ability to distinguish sounds from each other develop?

A large literature on infant speech perception suggests the capacity to detect minute speech sound differences in very early childhood (e.g., Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola, & Nelson, 2008; Polka & Werker, 1994; Werker & Tees, 1984; though see Narayan, Werker, & Beddor, 2010; Polka, Colantonio, & Sundara, 2001, for exceptions). By about 17 months of age, infants show evidence of learning similar-sounding words in lab settings (e.g., Stager & Werker, 1997; Werker, Cohen, Lloyd, Casasola, & Stager, 1998). However, 3- to 5-year-olds, who are well past infancy, are slower to learn words than older children and adults (Snow & Hoefnagel-Höhle, 1978). This is especially true when the words to be learned sound very similar (e.g., Creel & Frye, under review).

Recent work in our lab (Creel & Frye, under review) suggests that children ages 3 to 5 years have much greater difficulty than adults do at learning novel words with subtle but linguistically relevant phonetic distinctions like *deev* and *teev*. In that study, children were asked to learn two labels for two novel cartoon characters. In Experiment 2 of that study, each child completed three rounds of learning and testing. In two of the rounds, they learned a pair of similar words such as *deev* and *teev* or *vayfe* and *veff*, and in the third (order counterbalanced) they learned one set of dissimilar words such as *boove* and *sudge*. Each name was heard 16 times, twice on each of 8 learning trials per word. After learning trials ended, they saw both characters at once and heard the name of one, and were asked to point to the named one while their eye movements were tracked. Pointing accuracy by children was about .60 for similar-sounding words, compared to about .80 for dissimilar-sounding words, and adult accuracy above .90, where .50 represents chance performance (see Table 2). These findings are consistent with earlier research (Barton, 1976; Garnica, 1971) suggesting that young children have difficulty telling apart newly learned minimal pair words.

One question that naturally arises from findings of minimal-pair word learning difficulty in young children is the extent to which such findings reflect actual perceptual difficulty. In contrast to findings of infant success with distinguishing minimal pair sounds and learning minimal pair words, a literature on children past infancy implies that perceptual development itself is a protracted process, with perception of speech (Hazan & Barrett, 2000, McMurray, Danelz, Rigler, & Seedorff, 2018; Nittrouer, 2001, 2002; Ohde & Haley, 1997) and nonspeech sounds (Buss, Taylor, & Leibold, 2014; Creel, 2016; Keller & Cowan, 1994) improving across early childhood (Creel, 2016; Idemaru & Holt, 2013) and, in some cases, into the teen years (Hazan & Barrett, 2000; McMurray et al., 2018).

This slow developmental course seems somewhat at odds with findings that infants excel at perceptual discrimination, a discrepancy that has been noted by numerous researchers (Buss et al., 2014; Creel & Quam, 2015; Holt & Lalonde, 2012; Lalonde & Holt, 2014; see also Keen, 2003). However, this superficial difference may be driven by substantial differences in the tasks that are used in infancy vs. childhood. First, infant tests typically involve an *implicit or involuntary response* (dishabituation, electrophysiological potentials). Testing of older children tends to have stronger cognitive demands in that they are tested explicitly—asked to point to pictures, reach for objects, or make overt judgments. This requires holding task set in mind and maintaining attention, which is more difficult for young children than adults. Second, infant tests generally contain *repeated presentations* of a perceptual standard stimulus, such as a repeating "ba" occasionally punctuated by a change stimulus ("pa"; dishabituation, conditioned head-turn, electrophysiological potentials). Repetition may strengthen the standard representation in working memory, allowing more sensitive detection of a change (Banai & Yifat, 2011; Holt & Carney, 2005, 2007; Keller & Cowan, 1994; Viemeister & Wakefield, 1991). Tasks used with older children make stronger demands on perceptual memory: standard stimuli are not repeated, and if the task involves word learning children may need to retain form-referent mappings for a longer time scale (minutes to days) rather than a few seconds. Thus, one open question is whether children past infancy might look more advanced at perception of speech sounds if they were tested in a manner that is more similar to infant tests.

## Previous research on young children's sensitivity to minimal pairs

Several earlier researchers have tested young children's sensitivity to minimal speech sound differences. Two major variations are whether familiar vs. novel words are used, and whether children are asked to recognize named entities vs. making judgments about sound patterns. Evidence on word recognition seems to depend on the relative familiarity of both minimal pair words. Newton, Chiat, and Hald (2008) showed children ages 2-7 years pictures of familiar objects with similar names (like *pea* and *key*) and measured both pointing accuracy and looking behavior. Pointing accuracy was high even in 2-year-olds and increased with age, and voicing differences showed the highest rate of confusion. Barton (1976) found that 2-year-olds in a picture-pointing task distinguished familiar minimal pair words well, but unfamiliar ones (still real words) less so. If only one side of the minimal pair is familiar, the outcome is quite different. Swingley (2016), testing 2-year-olds, and Creel (2012), testing 3-6-year-olds, found that children treated novel minimal pairs of familiar words as those words themselves: for example, children mainly pointed to a picture of a fish when hearing “fesh” even in the presence of a novel object (Creel, 2012). Krueger, Storkel, and Minai (2018) found a similar pattern for misarticulated words (e.g., “weaf” for *leaf*), which are similar to single-feature changes. Interestingly, in follow-up work, Krueger and Storkel (2020) found that when the task emphasizes difference detection by including dissimilar-nonword training, false-positive rates (selecting a leaf picture for “weaf”) drop somewhat, suggesting that detection of minimal pair differences in word recognition may be malleable in older children (ages 3-6 years in their study; see Swingley, 2016, for younger children).

Other researchers have used similarity or identity judgments to examine children’s sensitivity to speech sound differences, again using a variety of familiar and unfamiliar words. Gerken, Murphy, and Aslin (1995) asked 4-year-old children to respond when they heard a target word (such as “little”), measuring similarity of novel words in terms of children’s false alarms (responding erroneously that, say, “nittle” was the target). Novel words with more features overlapping “little” received more target-word responses. This suggests that children may sometimes fail to notice a minor phonological difference, as appears to happen in word recognition tasks when hearing a novel word similar to a familiar one (“fesh” for *fish*). Storkel (2002) equated for lexicality by using all familiar words and asked children to respond when a test word (e.g., “tough”) was similar to the target word (*tug*). Like Gerken et al. (1995), 3-5-year-olds were more likely to give “similar” responses when real words overlapped in more phonological features vs. fewer. Still, the use of familiar words introduces features of semantic similarity, which may interact with children’s similarity decisions and complicate interpretation of results. Garnica (1971) equated for lexicality in a different way, by using novel word pairs that differed in onset consonants: eight 2-year-olds had to identify which of two just-named objects was being referred to (for example, “This is Mr. Gak. This is Mr. Dak. Put Mr. Dak in the basket”). While this might be construed as a word-recognition task, it presumably relies on working memory and thus it is grouped here with other discrimination-judgment studies. Garnica (1971) found weak minimal-pair differentiation. These studies taken together suggest that children can probabilistically distinguish similar-sounding word forms from each other.

Holt and Lalonde (2012; Lalonde & Holt, 2014) report the developmentally earliest overt change detection responses, using pairs of novel words. They trained 2- and 3-year-olds on a change detection test over multiple (2-3) one-hour sessions, using an active response (such jumping to a “change” or “no-change” response area) and multiple repetitions of stimuli (change trials: *sa sa sha sha*; no-change trials: *sa sa sa sa*). Children detected changes even in the most subtle contrast, from *sa* to *sha* or vice versa, though performance was better on the easier vowel distinction (*bu* vs. *ba*). What is striking about these studies is twofold: first, that children as young as 2 years largely complied with the task; and second, that this early ability to respond may have been facilitated by repeated stimulus presentation. This is consistent with a literature on “multiple looks,” temporal summation, or anchoring (Banai & Yifat, 2011; Holt & Carney, 2005, 2007; Viemeister & Wakefield, 1991) that suggests that repeated presentations of a standard stimulus boost its strength in working memory and thus facilitate detection of a change. Note that Gerken et al. (1995) and Storkel (2002) also used repeated standard stimuli, perhaps aiding older children’s performance.

To summarize a varied pattern of results, it seems as though young children can distinguish minimally different familiar words very well and can differentiate minimally different novel word forms modestly well, performing better when perceptual support is present and memory demands are minimal. Cases where one item is familiar and the other is novel may be especially challenging. Task differences, use of familiar vs. novel words, and presence or absence of bottom-up perceptual support (stimulus repetition) may contribute to discrepant results.

## The current study

The current work asks whether young children have difficulty perceptually differentiating similar sounding (minimal pair) words, controlling for semantic (dis)similarity by using novel words and testing a larger word set than any previous developmental study I know of besides Garnica (1971), who tested only 8 children and presented only consonants, with no vowels tested. If children have perceptual difficulty, why—that is, what cognitive and perceptual factors contribute to perceptual discrimination itself? Does easing cognitive challenges improve perceptual discrimination, or does increased bottom-up perceptual representation strength improve perceptual discrimination? Experiment 1 presented children with a same-different task including dissimilar words and minimal pair words. Experiment 2 replicated Experiment 1, but with trial-by-trial feedback in the form of repeated reminders of the task to aid children in maintaining task set (see Creel, 2019, and Droit-Volet & Izaute, 2009, for helpful effects of feedback on tasks in children in the 3-5-year age range). Experiment 3 replicated Experiment 1, but with additional standard stimulus presentations to boost bottom-up perceptual support.

If children have perceptual difficulty distinguishing similar-sounding words, then different-consonant and different-vowel pairs in all studies should be detected less well (that is, lower *d'* scores) than dissimilar-word pairs. If previous findings of a "consonant bias" are replicated, then consonant-differing pairs in all experiments should be better detected than vowel-differing pairs. If difficulty in detecting minimal changes is lessened by feedback, then

Experiment 2 change detection should exceed Experiment 1. Finally, if difficulty is lessened by the presence of a repeating standard, then Experiment 3 change detection should exceed Experiment 1.

## Experiment 1

### Method

**Participants**. Children were recruited from local preschools and day cares, with consent from parents/caregivers and verbal agreement from children. This and following experiments were approved by the UC San Diego Human Research Protections Program. To obtain a target sample size of 36, we tested 40 children, and had to exclude two more after the fact for not meeting inclusion criteria, yielding a sample of 34. Six were excluded due to: child ended session early (2); not passing the training trials (2);[1] experimenters could not understand child (1); exposure to language besides English (1). Throughout, no speech development measures were administered, but if teachers indicated a child had a language delay their data were excluded. The final sample had an average age of 4.25 years (SD = 0.76, 23 female). None of the participants in this experiment had taken part in Creel and Frye (under review).

**Stimuli**. Stimuli consisted of 32 novel consonant-vowel-consonant nonsense words drawn from English phonology (Table 1) that were also used in Creel and Frye (under review). Production age of acquisition and phonotactic probability of stimuli are available in Supplementary Tables 1-3. Each word in the set (e.g., *deev*) was related to two other words by a

---

[1] Participants who did not meet training criterion (Exp. 1: 1; Exp. 2: 8; Exp. 3: 4) were all under 4 years of age, with two exceptions: a 4.5-year-old in this experiment, and a very young 4-year-old in Experiment 2. It is not clear why most children under 4 in Experiment 1 met criterion but 40% of those in Experiment 2 under 4 did not, as training in Experiment 1 and 2 were identical. Previous studies in my lab suggest that children under age 4 have difficulty grasping auditory same-different tasks with nonspeech auditory stimuli (only 44% of 3-year-olds met criterion in Creel, under review), so the current Experiment 1 may simply be an exception to this general pattern.

change in a single phonological feature: either consonant voicing (*teev*) or vowel tenseness (*div*). Words were originally recorded both in sentences and in isolated citation form (four repetitions each) by a female native speaker of American English from the region where children were tested. Recordings took place in an Industrial Acoustics sound-isolation chamber using a Beyerdynamic SoundStar MK II microphone. Word learning studies presented words in sentences, but error-free recordings from the second and third pass through the isolated-word forms were used here. Words were edited to remove leading and trailing silences. Extracted words were normalized to 70 dB SPL in Praat (Boersma & Weenink, 2014) and resaved as .wav files.

**Table 1. Novel word stimuli used in all experiments. Consonant voicing minimal pairs are to left/right, vowel minimal pairs are above/below. International Phonetic Alphabet appears in / / for clarity.**

| | | | |
|---|---|---|---|
| vosh /vaʃ/ | fosh /faʃ/ | beesh /biʃ/ | peesh /piʃ/ |
| vush /vʌʃ/ | fush /fʌʃ/ | bish /bɪʃ/ | pish /pɪʃ/ |
| vayfe /veɪf/ | fayfe /feɪf/ | boove /buv/ | poove /puv/ |
| vehf /vɛf/ | fehf /fɛf/ | buhv /bʊv/ | puhv /pʊv/ |
| zodge /zadʒ/ | sodge /sadʒ/ | dayge /deɪdʒ/ | tayge /teɪdʒ/ |
| zudge /zʌdʒ/ | sudge /sʌdʒ/ | dedge /dɛdʒ/ | tedge /tɛdʒ/ |
| zoof /zuf/ | soof /suf/ | deev /div/ | teev /tiv/ |
| zuhf /zʊf/ | suhf /sʊf/ | dihv /dɪv/ | tihv /tɪv/ |

**Procedure**. Studies were run in Matlab Psychtoolbox3 using custom scripts written by the author, with data output to text file after each response. Children were tested in a quiet area in their school or day care, providing some limits on distraction but allowing the possibility of both visual and auditory distraction. Auditory distraction in particular might obscure distinctions between the quietest speech sounds, such as /f/ and /v/ (although these were less confusable overall than /s/ and /z/; see Exploratory Analyses Spanning Experiments section below). Still, this represents a more realistic listening context than carefully-controlled lab conditions. Children wore child-sized wired KidzGear fold-flat travel headphones. As described by the manufacturer, these lightweight headphones have a frequency response of 20-20,000 Hz and a sensitivity of 80-90 dB. The provided volume limiter was not used. Loudness level was checked by researchers at the beginning of each testing session at a given preschool or day care. If the child requested, the volume was adjusted.

The study used animated cartoon creatures to make the task more engaging. There were two sets of trials: training trials (8), and test trials (56). Of the training trials, presented in random order, half contained two presentations of the same word (e.g., *bish-bish*) and half contained dissimilar word pairs (*bish-soof*). The purpose of training was to make certain that children understood the task. On each training trial, the experimenter pressed a key, at which point the left animated creature enlarged, "spoke" a word (its mouth was depicted as closed, then as open during the word, and closed after, timed to the exact start and end of the audio file), and then reduced to its original size.[2] Next, the experimenter pressed a key again, and the right creature

---

[2] While this provided a visual duration cue in addition to the auditory duration cue that might distinguish the words, there is some evidence that children in this age range are more sensitive to auditory duration than to visual duration (Zélanti & Droit-Volet, 2012).

did the same. The child was then asked if they said “the same word, or different words” and provided a verbal response that the experimenter entered via keypress. Accuracy feedback (“yes/no, those were different/the same”) was printed on screen after each training trial and the experimenter read it to the child. If accuracy in a block was fewer than 7/8 trials, the block was repeated, up to 5 times. Children who did not meet training criterion were excluded from analyses (see Participants section).

Test trials were identical to training trials except that children did not receive feedback. There were four types of test trials: “same” trials (16); different-consonant (e.g., *deev-teev*; 16); different-vowel (*deev-div*; 16), dissimilar-word (*deev-vush*; 8). On “same” trials, children heard two different recordings of the same word. Use of two recordings yields a baseline for false alarm “different” responses based on minor (non-phonological) acoustic differences. Each child heard all possible similar pairs (in one of two possible orders). Dissimilar pairs (16 total, 8 per child) were chosen to minimize phoneme overlap. Within each dissimilar pair (e.g. *deev-vush*), onset mismatched in place, manner, and voicing; vowels mismatched in place and tenseness; and codas mismatched in one or more features. Four lists were constructed (see Supplementary Table 4), each in random order with the constraints that: no more than three trials with a “same” response or with a “different” response could occur in a row; no more than three trials of the same type (such as consonant-differing) could occur in a row; no two consecutive trials could contain the same recording (for example, *deev-teev* could not be followed by *div-deev*).

**Analysis**. Each participant’s data were converted to d-prime (*d’*) scores using proportion hits for each type of different trial and proportion false alarms on same trials. In conversion, hit or false-alarm scores of 0 were converted to .01 and scores of 1 were converted to .99 to avoid values of positive or negative infinity when computing the *z* transform. Thus the maximum *d’* score was *z*(.99)-*z*(.01) ≈ 4.65, and chance performance represents 0 or equivalent rates of hits and false alarms, that is *z*(p)-*z*(p) = 0. Thus there were three *d'* scores for each participant: different-consonant, different-vowel, and dissimilar-word. The *d'* scores for the three levels of Similarity Type (different-consonant, different-vowel, dissimilar-word) were then subjected to a repeated-measures analysis of variance (ANOVA). Planned t-tests compared *d'* at the three levels of similarity. Follow-up tests compared each to chance, with Bonferroni correction.

**Predictions**. If children have difficulty perceptually differentiating minimal pair words, then there should be an omnibus effect of Similarity Type and the different-vowel and different-consonant conditions should show lower *d'* than the dissimilar-word condition. If children have particular difficulty differentiating vowels, then different-vowel *d'* should be lower than different-consonant *d'*.

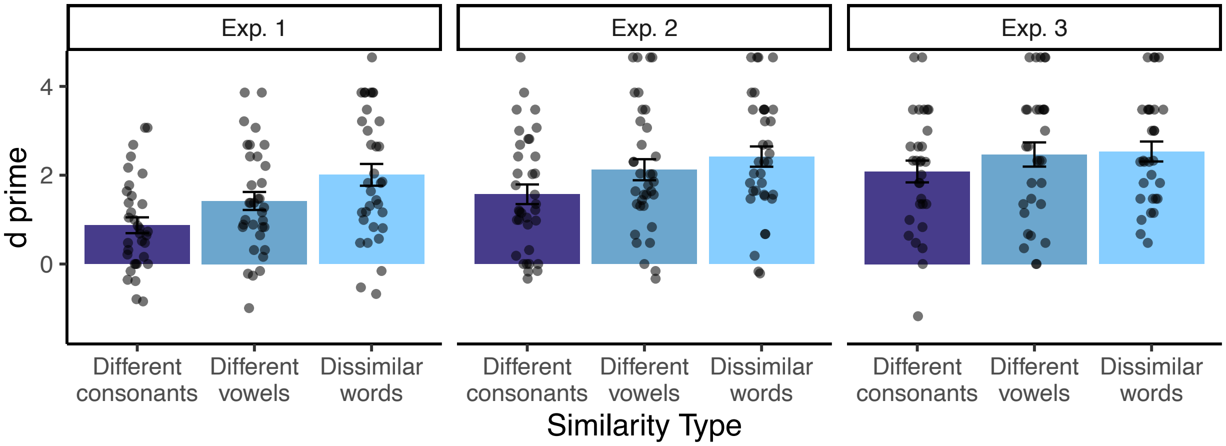


**Table 2. D-prime scores (SDs) from present same/different experiments on 3- to 5-year-olds, along with accuracy data from word learning studies using the same novel word set and comparison data from 2-year-olds in two change/no-change studies.**

| Study | Different consonants | Different vowels | Dissimilar words |
|---|---|---|---|
| **Word learning (proportion correct pointing responses)** | | | |
| Creel & Frye 1 (child) | 0.56 (0.23) | 0.63 (0.20) | . |
| Creel & Frye 1 (adult) | 0.96 (0.18) | 0.93 (0.22) | . |
| Creel & Frye 2 | 0.60 (0.27) | 0.63 (0.25) | 0.81 (0.24) |
| **Sound discrimination (d')** | | | |
| Exp. 1 | 0.87 (1.04) | 1.42 (1.18) | 2.01 (1.43) |
| Exp. 2 | 1.57 (1.31) | 2.12 (1.40) | 2.42 (1.35) |
| Exp. 3 | 2.08 (1.35) | 2.47 (1.49) | 2.53 (1.23) |
| Holt & Lalonde 2012[a,b] | 1.68 | 2.53 | 2.27 |
| Lalonde & Holt 2014[b,c] | 1.74 (1.73) | 2.27 (1.80) | 1.96 (1.21) |

[a] 2- and 3-year-olds. No SDs provided; average of scores across three age groups.

[b] Dissimilar words (training stimuli) always occurred first, perhaps contributing to heightened performance on the other contrasts due to practice effects.

[c] 2-year-olds.

## Results

The d' scores (Figure 1, Table 2) were subjected to a repeated-measures ANOVA, with Similarity Type as the independent variable. There was an effect of Similarity Type ($F(2,66) = 32.94$, $p < .0001$). Paired t-tests revealed a three-way distinction: dissimilar-word *d'* exceeded different-vowel ($t(33) = 4.83$, $p < .0001$) and different-consonant ($t(33) = 6.54$, $p < .0001$) *d'*, and different-vowel exceeded different-consonant *d'* ($t(33) = 4.67$, $p < .0001$). Nonetheless, all three conditions exceeded chance (zero) performance (all $t$>4.90, $p$<.0001 after Bonferroni correction).

## Discussion

Results suggest that children do have difficulty discriminating similar-sounding words, both consonant-differing and vowel-differing words, compared to a baseline of dissimilar-sounding words. Counter to predictions based on consonant biases in word learning, difficulty appeared greater for consonant voicing differences than vowel tenseness differences.

On the one hand, this fits the data pattern from word-learning studies (Creel & Frye, under review) in which different-consonant and different-vowel words were more difficult to learn to distinguish than were dissimilar words. However, it diverges slightly from that pattern in that vowels here were more *discriminable* than consonants overall, but in the word-learning studies vowel-differing words were not *learned* more accurately than consonant-differing words.

These data provide some initial support for the idea that word learning difficulty is predicted by differences in discriminability. They also suggest that children well past infancy are not perfectly tuned to subtle but contrastive speech sound properties. However, several alternative explanations exist. One such explanation is that children are quite good at distinguishing speech sound properties, but due to their cognitive immaturity they lose focus during the task and miss some of the subtler differences. Accordingly, the next experiment replicated the current one, but provided trial-by-trial task reminders during the test phase to maintain children's focus.

## Experiment 2

## Method

**Participants**. A new sample of children were recruited from the same pool as before. To obtain a target sample size of 36, we tested 53 children, and had to exclude an additional one after the fact after previous participation was discovered. Seventeen were excluded due to: not passing the training in 5 blocks (9); child ended experiment early (4); computer error (1); did not vary responses for 3+ training blocks (2); speech delay noted by teacher (1); had mistakenly completed the same study 10 months before (1). For the final sample (*N* = 35), age was 4.45 years (*SD* = 0.82; 15 female). One participant had completed a word-learning study using the same words one year before, but was included nonetheless as their results were similar to other children of the same age.

**Stimuli**. These were the same as in Experiment 1.

**Procedure**. The procedure matched Experiment 1 except that it was modified so that on all test trials, children were provided accuracy feedback on each trial. Test trial accuracy feedback was identical to that presented on training trials.

**Analysis and Predictions**. For the study itself, analysis and predictions were the same as in Experiment 1. An additional planned ANCOVA compared this experiment with Experiment 1 to assess whether performance improved in the current experiment due to the addition of task reminders. Due to age differences between studies (younger in Experiment 1), Age and its interactions were included as covariates in this analysis. If task reminders boost performance, then there should be an effect of Experiment, with higher *d'* values in Experiment 2.

## Results

An ANOVA revealed that Similarity Type was significant ($F(2,68) = 22.77$, $p < .0001$). Paired t-tests showed that dissimilar-word *d'* exceeded different-vowel ($t(34) = 2.70$, $p = .01$) and different-consonant ($t(34) = 6.32$, $p < .0001$) *d'*, and different-vowel *d'* exceeded different-consonant *d'* ($t(34) = 4.02$, $p = .0003$). All three conditions exceeded chance (zero) performance (all $t$>4.90, $p$<.0001 after Bonferroni correction).

An ANCOVA regressed out Age in years and its interactions with Experiment (1, 2) as a between-groups predictor and Similarity Type as a repeated measure. Age was significant ($F(1,65) = 35.00$, $p < .0001$), such that *d'* increased with age. However, it did not interact with any other factors. Similarity Type was significant ($F(2,130) = 53.54$, $p < .0001$). Finally, the effect of Experiment was marginally significant ($F(1,65) = 3.59$, $p = .06$), with higher overall *d'* in Experiment 2 than in Experiment 1. The interaction did not reach significance (F(2,130) = 1.59, $p = .21$).

**Discussion**

The current experiment replicated two findings from Experiment 1. First, children have heightened difficulty telling apart minimal-pair words, with lower *d'* for different-consonant and different-vowel trials compared to dissimilar-word trials. Second, children do *not* show increased difficulty differentiating vowel-differing words compared to different consonant words, with the opposite being true (different-consonant trials showed lower *d'* than different-vowel trials).

Analyses also asked whether children would show decreased difficulty in discriminating novel words when frequent reminders served to keep them on-task. The answer is a weak yes: a cross-experiment comparison indicated that children here showed marginally stronger discrimination overall than in the first experiment, which had no reminders. Still, the Experiment factor did not interact with Similarity Type, meaning that gains were not appreciably greater for similar-sounding trials than dissimilar-sounding trials. Thus, any improvement with reminders may represent across-the-board improvement on the task rather than specific improvement in similar-sounding word discrimination. The reader should bear in mind that these two studies were run at different time ranges and with some variation in personnel (though personnel were highly trained and each sample included children from multiple preschools or day cares for variety), and thus the comparison should be interpreted with caution.

Another point to note is that the improved performance in this experiment did not eliminate similarity-based discrimination difficulty. That is, children still had more difficulty telling apart the single-feature-differing words than the dissimilar words. This suggests that perceptual difficulty is not simply an artifact of lack of task focus and that children well past infancy experience challenges in differentiating similar-sounding words. Still, other explanations remain. A particularly salient one is that infant discrimination testing tends to take the form of many stimulus repetitions of a standard, followed by a change in stimulus, and the infant's change response is measured. In the current study, each child heard only one standard (the first word in a trial) before hearing the potentially changed stimulus (the second word in the trial). In the infant case, repeated presentations may strengthen the representation of a standard, making the change stimulus easier to detect. This raises the question of whether increasing the number of standard repetitions would improve change detection performance in the current task, and some evidence exists to support this possibility (Banai & Yifat, 2011; Holt & Carney, 2005, 2007; see Viemeister & Wakefield, 1991). Accordingly, the final experiment incorporated more standard stimulus repetitions into the design.

## Experiment 3

### Method

**Participants**. A new sample of children were recruited from the same pool as before. Our target sample size was 36, as before, but due to the onset of the Covid pandemic, data collection was cut short at 31. To maintain reasonable speed in scientific progress, the decision was to go forward with a nearly complete sample. To obtain 31, 46 children were tested. One additional participant was excluded after discovery of previous participation in Experiment 2. Fifteen were excluded due to: not passing the training trials within 5 repetitions (4); ended experiment early (4); computer error (4); experimenter error (2); exposure to a language besides English (1). The remaining participants (*N* = 30) averaged 4.49 years of age (*SD* = 0.74, 14 female). Three had previously taken part in a word-learning study using this word set 7-18 months prior, but were included as their results were similar to other participants.

**Stimuli**. These were the same as in Experiments 1 and 2, except that "same" trials simply repeated a single recording of a word rather than presenting two different versions of the same word. This was done to simplify design: if multiple standard recordings had been used, counterbalancing which ones occurred in what order across trials without creating confounds would be complex in a situation where several other factors required counterbalancing. False alarm rates were not markedly different than the other experiments, especially Experiment 2, suggesting this made little difference (.27, .21, .21 in Experiments 1, 2, and 3, respectively).

**Procedure**. Eight different lists were constructed. Each list had its own training trials and testing trials. As before there were 8 training trials, but in this case, the "same" trials all contained a single standard stimulus (for example, for list 1, this was always the same recording of *beesh*). The four "different" trials contained the standard as the first word (e.g., *beesh*) and a fixed distantly related word (e.g., *teeve*, which differs from *beesh* in place and voicing at onset and coda) as the second word. Design constraints limited the distance of the dissimilar words but all pairs differed by at least two segments (see Supplementary Table 5).

There were 56 test trials, presented without feedback. They were assorted differently than the previous experiments. There were 24 "filler" same trials, which contained a standard that repeated throughout the experiment. The number of filler same trials was chosen to allow frequent standard repetition without extending the length of the experiment past children's attentional limits. The remaining 32 trials were experimental trials. There were 8 experimental same trials, and 8 each different-consonant, different-vowel, and dissimilar-word. Experimental trials were divided into standard and nonstandard trials. For example, for list 1, the standard word in same trials was always *beesh*. There were two subsets of different pairs: ones that contained the repeating standard as the first word, such as *beesh/bish; beesh/peesh; beesh/teeve*; and ones that did not contain the repeating standard, such as *zoof/zoohf; zoof/soof; zoof/fayfe*. (In a different list, list 5, *zoof* was the standard and *beesh* was the nonstandard item.) From 2-5 participants were tested with each of the eight standards. The repeating standard, compared to the non-repeating standard, was used to assess whether greater bottom-up support for the standard increased difference detection. Trials were pseudorandomly ordered so that at least one filler same trial occurred every three trials to maintain representation of the standard. However, there were no more than three trials in a row that had identical responses (all same or all different). Note that the trial composition meant that only 43% of trials, not 71% as in previous studies, were "different" trials. According to classical signal detection theory (see Macmillan & Creelman, 1995), this is assumed to change a listener's bias (overall tendency to respond "different") but not detection sensitivity, that is, *d'*.

**Analysis and Predictions**. An ANOVA included Similarity Type and Standardness (standard [stimuli contained repeated standard], or nonstandard [did not contain repeated standard]) as repeated measures. In addition to predictions in the first two experiments, if participants benefit in distinguishing sounds when the sounds are compared to a frequently-repeated standard stimulus, there should be a significant effect of Standardness, with higher *d'* for standard vs. non-standard trials. In this analysis, proportions of hits and false alarms were converted to *d'* values, separately for standard and nonstandard conditions. Thus each *d'* value was based on 4 hits and 4 false alarms. To test the main effect of Similarity Type, data were collapsed over Standardness and *d'* was recomputed, using 8 hits and 8 false alarms.[3]

An additional planned ANCOVA compared this experiment with Experiment 1 to assess whether performance improved in the current experiment due to repetition of a standard stimulus. Due to age differences between studies (younger in Experiment 1), Age and its interactions were included as covariates in this analysis.

## Results

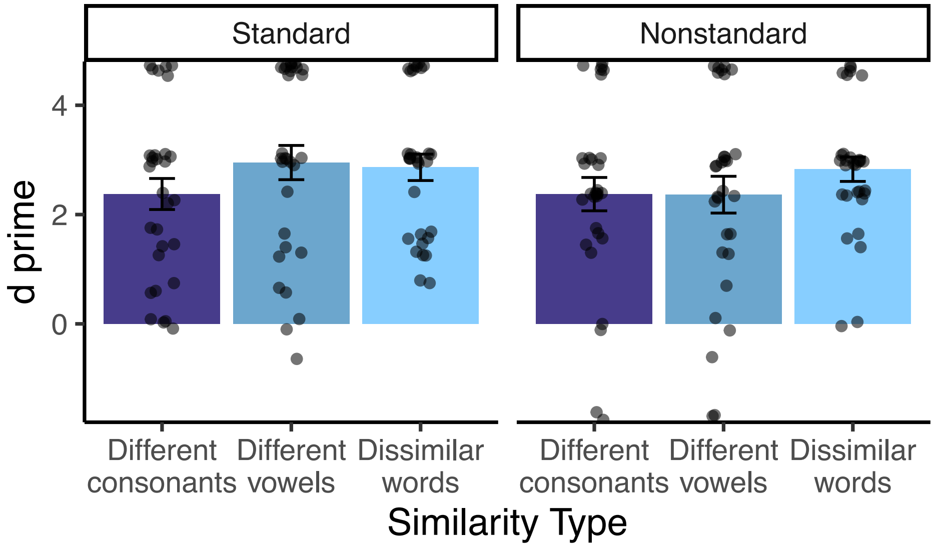
As a check, false alarm rates on the filler-same trials were calculated. False alarms were low for filler trials (.20, *SD* = .18) and similar to rates for standard (.20, SD = .24) and nonstandard (.23, SD = .25) experimental same trials. Filler trials were not analyzed further.

In an ANOVA with Similarity Type and Standardness as independent variables, Standardness did not approach significance (*F*(1,30) = 1.19, *p* = .28), but Similarity Type approached significance (*F*(2,58) = 2.84, *p* = .07). The interaction missed significance (*F*(2,58) =

---

[3] These are smaller numbers of raw hits and false alarms than in Experiments 1-2, which used either 8 different and 16 same trials (dissimilar) or 16 different and 16 same trials (different-consonant and different-vowel). To assess whether lower raw numbers of trials affected sensitivity to detect effects of Similarity Type, analyses were rerun on the first two experiments' data using only the first 4 or first 8 of each each trial type. Similarity Type remained strongly significant in all cases, and results of t-tests were largely unchanged (see Supplementary Table 6).

1.76, *p* = .18). To assess effects of Similarity Type as in previous experiments, data were collapsed over Standardness and paired *t*-tests were computed. These tests revealed that dissimilar-word *d'* exceeded different-consonant (*t*(29) = 2.32, *p* = .03) but not different-vowel *d'* (*t*(29) = 0.36, *p* = .72). Different-vowel and different-consonant discrimination did not differ (*t*(29) = 1.64, *p* = .11). All three conditions exceeded chance (zero) performance (all *t*>7.10, *p*<.0001 after Bonferroni correction).



An ANCOVA compared current results to Experiment 1. For this analysis, Experiment 3 data were collapsed across Standardness and *d'* values were recalculated. Age was significant (*F*(1,60) = 13.84, *p* = .0004). Similarity Type (*F*(2,120) = 22.05, *p* < .0001) was also significant. Experiment was also significant (*F*(1,61) = 8.78, *p* = .004), with higher *d'* in Experiment 3 than in Experiment 1. However, the latter two effects were qualified by a Similarity Type x Experiment interaction (*F*(2, 120) = 4.46, *p* = .01). Follow-up ANCOVAs on each Similarity Type suggested that the reason for the interaction was that the increase in *d'* in the current experiment was more robust for different-consonant (*F*(1,60) = 15.05, $p_{Bonf}$ = .0008) and different-vowel ($F(1,60) = 8.30$, $p_{Bonf} = .02$) conditions than for the dissimilar-word condition ($F(1,60) = 1.43$, $p_{Bonf} = .71$). In the model comparing the current experiment to Experiment 2, the effect of Experiment was not significant ($F(1,61) = 1.12$, $p = .30$).

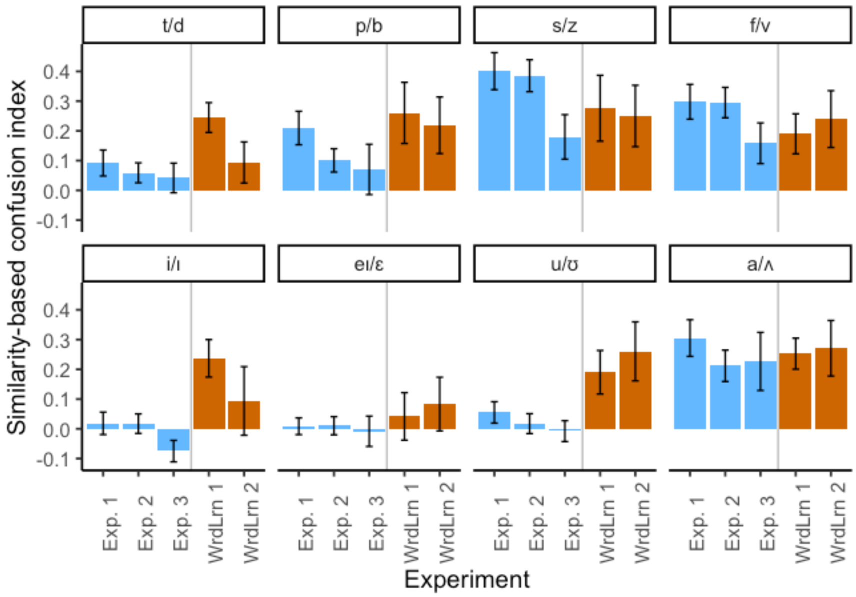
## Discussion

This experiment replicated the general pattern of similar-sounding pairs being harder to discriminate than dissimilar pairs, although this only held for different-consonant pairs, not different-vowel pairs. We also asked whether children continue to have difficulty discriminating similar-sounding words even in the presence of a repeating standard stimulus, which aimed to boost the perceptual representation of the standard. Within the experiment, whether a change trial included the repeating standard appeared to have no effect on discrimination, yet compared to Experiment 1, performance was improved, and more so for the minimal-pair trials than dissimilar-word trials. There was no significant difference between Experiment 3 and Experiment 2, suggesting that repeated task reminders (Experiment 2) and repeated standards (Experiment 3) are not different in their level of helpfulness.

It is slightly puzzling that performance would be better in the current study than in Experiment 1 given that the standard stimulus set—the one that matched the repeating standard—did not show better discrimination the non-standard set. Aside from a false positive, one possibility is that the repeated standards (or a reduction in total number of word tokens heard, 8 here vs. 32 in the other two experiments) lessened children's memory load and thus benefited all trials generally, not just those containing the repeating standard. A second possibility is that even the "nonstandard" standard occurred on a large number of trials (16), which may have conferred some repetition benefit. Future work should explore effects of various frequencies of stimulus repetition on children's discrimination performance.

## Exploratory Analyses Spanning Experiments

Several questions can be addressed more fully by combining data across studies. One question is whether children *can* readily distinguish similar sounds but only if they are extremely attentive. If so, then children with better attention should show minimal errors on similar-word trials, with similarity-based errors in the overall data set being generated by the less attentive children. While the current study does not have a direct assessment of attention, one can ask whether children who were the most "on-task" in terms of $d'$ on the easy trials (dissimilar word trials) escaped from phonological similarity difficulty. That is, do these on-task children show $d'$ values for minimal pairs that are equivalently high to $d'$ for dissimilar words? In short, the answer is no. I looked at the children across studies whose dissimilar-word $d'$ exceeded 3.5 (17 out of 99 children total). For different-consonant, different-vowel, and dissimilar-word conditions respectively, their mean $d'$ values were $2.76 \pm 1.04$, $3.62 \pm 0.97$, $4.28 \pm 0.41$. All of these differed from each other ($t(16) \geq 3.48$, $p_{Bonf} \leq .009$). This suggests that even the most on-task children experienced increased difficulty when hearing similar-sounding words.



These data also allow an exploration of the relationship between discriminability and learning of similar-sounding words, by comparing discrimination here to learning in Creel & Frye (under review), which used the same word set. Is ease of learning in Creel & Frye (under review) straighforwardly predicted by perceptual confusions here? At a high level, it is clear that words which are harder to tell apart perceptually (minimal pairs) are harder to learn than those that are easier to tell apart (dissimilar words). It does not seem to be the case that children have excellent perception but weak word learning. However, one difference from the word learning studies using these same words (Creel & Frye, under review) is obvious: in the current studies, consonant-differing words are harder to tell apart than vowel-differing words are. This differs from the word learning studies, where consonant-differing words were not significantly more difficult to *learn* than vowel-differing words. Further, in Experiment 3, vowel difference *d'* no longer fell below *d'* for dissimilar words. By contrast, in the word learning study that included a dissimilar-word baseline, vowel-differing pairs were significantly more difficult to learn than dissimilar-word pairs.

To get at the reasons for this discrepancy, I explored difficulty for the particular sound pairs used in the study. I calculated a similarity difficulty score for the current experiments as well as Creel and Frye (under review)'s word-learning experiments using the same novel words. For discrimination, this score was calculated for each participant by subtracting hits on similar trials for a particular sound (such as responding "different" when hearing "teev … deev") from hits on dissimilar trials. For word learning, this score was calculated by subtracting accuracy when learning similar-sounding labels for pictures (such as learning the labels "teev" and "deev") from that participant's accuracy in learning dissimilar labels for pictures.[4]

The resulting scores are depicted in Figure 3 (discrimination: left three bars in each subpanel; learning: right two bars). Briefly, some sounds that appear easy to discriminate nonetheless generate difficulty in word learning, particularly three of four vowel-differing word pairs (i/ɪ, u/ʊ, eɪ/ɛ) which are as easy to discriminate as dissimilar words (score near zero). Consonant-differing word pairs appear more difficult to discriminate than dissimilar words, particularly fricative voicing differences (s/z, f/v). Yet learning fricative-differing words is not markedly more difficult than learning stop-differing words. Thus, this exploration suggests that two words or sounds can be highly discriminable yet still present difficulty in word learning (see Creel, 2016, and Creel & Dahan, 2010, for related findings). Phonotactic probability was also calculated (see Supplementary Table 4) but it did not straightforwardly predict the pattern of discriminability.

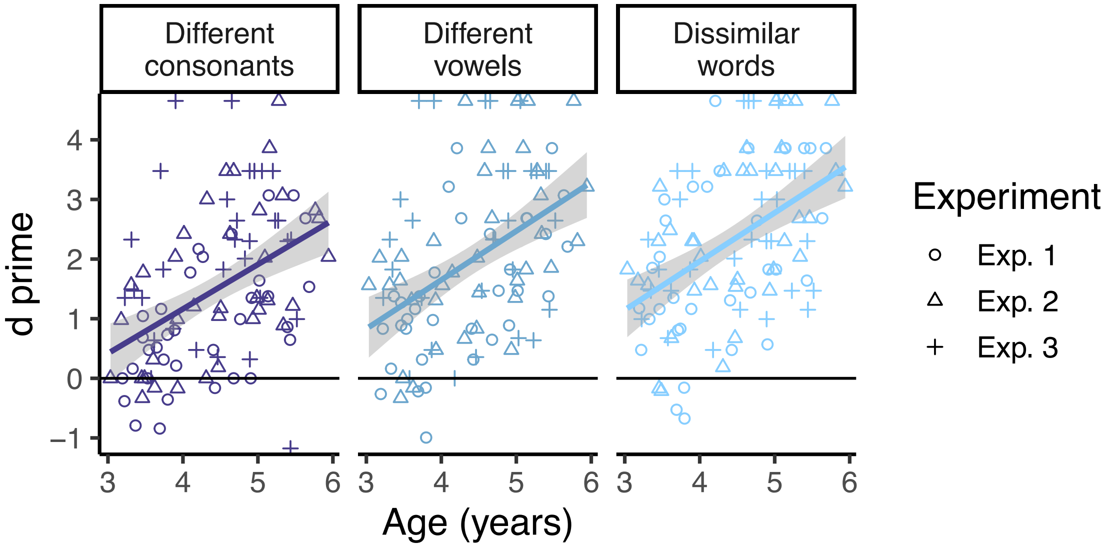
---

[4] Since Creel and Frye's (under review) Experiment 1 did not have a baseline dissimilar word learning condition, the mean of dissimilar word learning from Creel and Frye's Experiment 2 was used. The reader should keep in mind that there are fewer participants per cell in Experiment 3 (14-16 per sound contrast) and in the Creel and Frye (under review) word learning studies (8 per contrast in their Exp. 1, 12 per contrast in their Exp. 2), making estimates more volatile. Further, Creel and Frye's Experiment 1 used a different voice than used in the current studies, while their Experiment 2 used the same voice as the current studies.

## General Discussion

The primary question was whether young children find it hard to tell apart similar-sounding words. The answer is yes: children find it more difficult to report that two words are different when those two words differ in a minimal speech sound contrast than when they differ in multiple features. Further, this effect holds across a three-year age range: as evident in Figure 4, children are not approaching ceiling performance on similar-sounding words even near age 6 years. This is true both in terms of high absolute performance (maximum d prime of 4.65 in the studies here) and high relative performance (performance on similar-sounding words vs. in the baseline distinct-word condition).

Additionally, the study asked whether task reminders or bottom-up stimulus support improve performance. Performance differences between experiments suggested marginal improvement due to task reminders (Experiment 2) and significant improvement due to stimulus repetition (Experiment 3). Neither of these manipulations fully eliminated the increased difficulty of distinguishing between similar-sounding words. Task reminders did not eliminate similar-sound confusions for vowels or consonants (Experiment 2), nor did highly on-task children escape from sound confusion effects (Exploratory Analyses Spanning Experiments). Bottom-up stimulus support did not eliminate similar-sound confusion for consonants (Experiment 3).



## Theoretical implications

One interpretation of these findings is that perceptual memory is still developing. This is consistent with the improvements with age seen in the current study. It is also consistent with the apparent improvement in Experiment 3, which contained repeated stimuli, vs. Experiment 1, which did not repeat stimuli, in that stimulus repetition can bolster weaker perceptual memory. Evidence of ongoing perceptual development is consistent with a larger literature suggesting improvements in perception and production for multiple years during development (Gomes, Sussman, Ritter, Kurtzman, Cowan, & Vaughan, 1999; Hazan & Barrett, 2000; McMurray et al., 2018; Nittrouer, 2001, 2002; Ohde & Haley, 1997; for a protracted perceptual learning account see Creel, 2018; Creel & Quam, 2015).

A different interpretation of the findings here is that they reflect still-developing attention or general cognition. This is consistent with mild improvement in Experiment 2, which contained repeated instructions designed to focus attention on task, vs. Experiment 1, which did not repeat task instructions. However, those improvements do not eradicate the difficulty imposed by phonologically similar words. Further, one might reasonably argue that the “dissimilar” words here are not as dissimilar as they could be, given that all words used in this study are single closed syllables with singleton obstruent onsets and codas. Words can certainly be more dissimilar than this (for example, *manamana* vs. *doot*). Thus, even the improvements in distinguishing dissimilar words might partly indicate increasing perceptual acuity or perceptual memory rather than cognitive effects.

Findings have some implications for phonological development as well. It is particularly interesting that consonant voicing is difficult for children to differentiate, given that there are demonstrations of developmentally early perceptual sensitivity to voicing (Eimas et al., 1971; see also Rost & McMurray, 2009, 2010). Still, it fits with findings of Treiman, Broderick, Tincoff, and Rodriguez (1998), who found that consonant voicing distinctions were harder to detect than place distinctions, and Newton et al. (2008), who reported that young children have more difficulty distinguishing consonant voicing than consonant place. Close examination of data in Garnica (1971) also reveal particular difficulty with consonant pairs differing in voicing.

It is also interesting that, on average, vowels are easier to distinguish than consonants. The interest is that this finding seems discordant with a body of research on multiple languages including English which suggests that vowels are harder to tell apart or less “lexical” than consonants, that is, less indicative of meaning differences (Bonatti, Peña, Nespor, & Mehler, 2005; Nazzi, 2005; Nespor, Peña, & Mehler, 2003; see Creel, Aslin, & Tanenhaus, 2006, and Van Ooijen, 1996, for evidence in English-speaking adults; for review, see Nazzi, Poltrock, & Van Holzen, 2016). Developmental evidence in English is less consistent than that seen in other languages, with some researchers reporting better differentiation of consonant-differing words

(Nazzi, Floccia, Moquet & Butler, 2009) but others reporting no differentiation (Floccia, Nazzi, Delle Luche, Poltrock, & Goslin, 2014; Mani & Plunkett, 2007; Swingley, 2016).

One possible explanation for English adults but not children showing a consonant bias is that development of cues to each sound type vary in their rate of acquisition. Tense-lax vowel pairs are distinguished by first and second formants, and they also differ in duration, which is longer for tense vowels (Hillenbrand, Getty, Clark, & Wheeler, 1995). While young English-learning children (21 months) appear insensitive to changes in vowel duration (Swingley & van der Feest, 2019), English-learning infants have been shown to discriminate numerous vowel contrasts (see, e.g., Werker & Tees, 1999). Voiced and voiceless alternants of both stops (e.g. Lisker & Abramson, 1964; Raphael, 2005) and fricatives (Massaro & Cohen, 1977) are distinguished by voice onset time (VOT), the time between the mouth opening and phonation beginning, which is longer for voiceless variants; and to a degree by fundamental frequency (f0), which is higher at onset for voiceless variants. Sensitivity to VOT is evident in infancy (Eimas et al., 1971; see Galle & McMurray, 2014, for a review), though preschool-aged children appear insensitive to f0 cues to stop identity (Bernstein, 1983). One might think of frequency/spectral cues (formants) as dominant for vowels but duration cues (VOT) as dominant for consonant voicing, and infer that frequency discrimination may mature earlier than duration discrimination, leading to greater vowel weighting earlier on. Yet both duration and frequency discrimination appear to improve over the course of development (e.g., Jensen & Neff, 1993), making it difficult to ascertain how acoustic sensitivities alone might yield a shift toward consonant bias in adulthood in English speakers. A different possibility is that consonant informativeness, learned over experience into adulthood, leads to a stronger weighting of consonantal information (see Nazzi et al., 2016, for a review).

## Limitations

Some unresolved questions remain. One is why children in Experiment 3, the study which included a repeated standard stimulus, would perform better overall given that the repeated-standard trials did not show an advantage. It may be that the repeating standard relieved working memory resources and thus facilitated performance on all trials. Another possibility is that even the “non-standard” stimuli did in fact repeat, albeit less frequently, but perhaps enough to boost performance. This should be addressed in future work as it has implications for clinical implementations of same-different tests of speech processing.

Another open question with clinical import is how the current findings relate to Holt and Lalonde’s (2012; Lalonde & Holt, 2014) work with even-younger children, ages 2 and 3 years. They used a related paradigm, change/no-change, instead of same/different. They too tested children on novel words that were dissimilar (*u* vs. *ga*), vowel-differing (*ba* vs. *bu*), or consonant-differing (*sa* vs. *sha*). One might expect younger children to perform less well than the 3-5-year-olds in the current study. However, Holt and Lalonde (2012) and Lalonde and Holt (2014) found *d’* values similar to or higher than the older children in the current work (see Table 2). This suggests that change/no-change may be easier for children to understand or may provide better bottom-up perceptual support than same/different. It is also possible that children in Holt and Lalonde (2012; Lalonde & Holt, 2014) benefited from a quieter lab environment, stronger practice effects in the multi-day test procedure, or from a vowel difference that was more prominent than the one used in the current studies. The advantage of the Holt and Lalonde work (2012; Lalonde & Holt, 2014) is its greater sensitivity in younger children, while the advantage of the current work is in testing more stimuli in a shorter (single-session) time frame. Future work should explore whether the tasks might be combined to test more stimuli in a shorter period of time with greater sensitivity across a wide age range.

**Acknowledgments**. Thanks to lab managers Alicia Escobedo and Kristie McCrary Kambourakis, and graduate student Reina Mizrahi, who all collected data and provided many insights into the work. Thanks also to Mark Appelbaum for spirited discussion of test procedures. Finally, thanks to all the childcare facilities, teachers, parents, and children who gave us the gift of their time to allow this work to take place! Research was supported by NSF CAREER Award BCS-1057080.

# References

Banai, K., & Yifat, R. (2011). Perceptual anchoring in preschool children: Not adultlike, but there. *PLoS ONE, 6*(5). https://doi.org/10.1371/journal.pone.0019769

Barton, D. (1976). Phonemic discrimination and the knowledge of words in children under 3 years. *Papers and Reports on Child Language Development*, *11*, 1–9.

Bernstein, L. E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. *Journal of Phonetics, 11*(4), 383–393. https://doi.org/10.1016/s0095-4470(19)30837-x

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. *Psychological Science*, *16*(6), 451–459.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Boersma, P., &Weenink, D. (2014). *Praat: Doing phonetics by computer* (Version 5.4.01) [Computer program]. Retrieved November 9, 2014, from www.praat.org

Buss, E., Taylor, C. N., & Leibold, L. J. (2014). Factors Affecting Sensitivity to Frequency Change in School-Age Children and Adults. *Journal of Speech Language and Hearing Research*, *57*, 1972–1982. https://doi.org/10.1044/2014

Creel, S. C. (2012). Phonological similarity and mutual exclusivity: on-line recognition of atypical pronunciations in 3-5-year-olds. *Developmental Science*, *15*(5), 697–713. https://doi.org/10.1111/j.1467-7687.2012.01173.x

Creel, S. C. (2016). Ups and downs in auditory development: Preschoolers' sensitivity to pitch contour and timbre. *Cognitive Science*, *40*, 373–403. https://doi.org/10.1111/cogs.12237

Creel, S. C. (2018). Accent detection and social cognition: Evidence of protracted learning. *Developmental Science*, *21*(2), e12524. https://doi.org/10.1111/desc.12524

Creel, S. C. (2019). The familiar-melody advantage in auditory perceptual development: Parallels between spoken language acquisition and general auditory perception. *Attention, Perception, & Psychophysics, 81*(4), 948–957.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language*, *54*, 1–19.

Creel, S. C., & Dahan, D. (2010). The effect of the temporal structure of spoken words on paired-associate learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 110–122. https://doi.org/10.1037/a0017527

Creel, S. C., & Quam, C. (2015). Apples and oranges: Developmental discontinuities in processing? *Trends in Cognitive Sciences*, *19*(12), 713–716. https://doi.org/10.1016/j.tics.2015.09.006

Droit-Volet, S., & Izaute, M. (2009). Improving time discrimination in children and adults in a temporal bisection task: The effects of feedback and no forced choice on decision and memory processes. *Quarterly Journal of Experimental Psychology, 62*(6), 1173–1188. https://doi.org/10.1080/17470210802384180

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Floccia, C., Nazzi, T., Delle Luche, C., Poltrock, S., & Goslin, J. (2014). English-learning one- to two-year-olds do not show a consonant bias in word learning. *Journal of Child Language*, *41*(5), 1085–1114. https://doi.org/10.1017/S0305000913000287

Galle, M. E., & McMurray, B. (2014). The development of voicing categories: A quantitative review of over 40 years of infant speech perception research. *Psychonomic Bulletin & Review, 21*(4), 884–906. https://doi.org/10.3758/s13423-013-0569-y

Garnica, O. K. (1971). The Development of the Perception of Phonemic Differences in Initial Consonants by English-Speaking Children: A Pilot Study. *Papers and Reports on Child Language Development*, *3*, 1–31.

Gerken, L., Murphy, W. D., & Aslin, R. N. (1995). Three- and four-year-olds' perceptual confusions for spoken words. *Perception & Psychophysics*, *57*(4), 475–486.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Gomes, H., Sussman, E., Ritter, W., Kurtzberg, D., Cowan, N., & Vaughan, H. G. (1999). Electrophysiological evidence of developmental changes in the duration of auditory sensory memory. *Developmental Psychology*, *35*(1), 294–302. http://www.ncbi.nlm.nih.gov/pubmed/9923483

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*(4), 377–396. https://doi.org/10.1006/jpho.2000.0121

Holt, R. F., & Carney, A. E. (2005). Multiple looks in speech sound discrimination in adults. *Journal of Speech, Language, and Hearing Research*, *48*(4), 922–943. https://doi.org/10.1044/1092-4388(2005/064)

Holt, R. F., & Carney, A. E. (2007). Developmental Effects of Multiple Looks in Speech Sound Discrimination. *Journal of Speech Language and Hearing Research*, *50*(6), 1404. https://doi.org/10.1044/1092-4388(2007/098)

Holt, R. F., & Lalonde, K. (2012). Assessing toddlers’ speech-sound discrimination. *International Journal of Pediatric Otorhinolaryngology*, *76*(5), 680–692. https://doi.org/10.1016/j.ijporl.2012.02.020

Idemaru, K., & Holt, L. L. (2013). The developmental trajectory of children’s perception and production of English /r/-/l/. *Journal of the Acoustical Society of America*, *4232*. https://doi.org/10.1121/1.4802905

Jensen, J. K., & Neff, D. L. (1993). Development of Basic Auditory Discrimination in Preschool Children. *Psychological Science*, *4*(2), 104–107. https://doi.org/10.1111/j.1467-9280.1993.tb00469.x

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). Academic Press.

Keen, R. (2003). Representation of objects and events: Why do infants look so smart and toddlers look so dumb? *Current Directions in Psychological Science*, *12*(3), 79–83. https://doi.org/10.1111/1467-8721.01234

Keller, T. A., & Cowan, N. (1994). Developmental increase in the duration of memory for tone pitch. *Developmental Psychology*, *30*(6), 855–863. https://doi.org/10.1037//0012-1649.30.6.855

Krueger, B. I., & Storkel, H. L. (2020). Children’s Response Bias and Identification of Misarticulated Words. *Journal of Speech Language and Hearing Research*, *63*(January), 259–273.

Krueger, B. I., Storkel, H. L., & Minai, U. (2018). The Influence of Misarticulations on Children's Word Identification and Processing. *Journal of Speech Language and Hearing Research*, *61*(April), 820–836. https://doi.org/10.1044/2017_JSLHR-S-16-0379

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, *363*, 979–1000. https://doi.org/10.1098/rstb.2007.2154

Lalonde, K., & Holt, R. F. (2014). Cognitive and linguistic sources of variance in 2-year-olds' speech-sound discrimination: A preliminary investigation. *Journal of Speech, Language, and Hearing Research*, *57*(1), 308–326. https://doi.org/10.1044/1092-4388(2013/12-0227)

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*(3), 384–422.

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*(2), 252–272. https://doi.org/10.1016/j.jml.2007.03.005

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, *54*(8), 1472–1491.

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science*, *13*(3), 407–420. https://doi.org/10.1111/j.1467-7687.2009.00898.x

Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: differences between consonants and vowels. *Cognition*, *98*(1), 13–30. https://doi.org/10.1016/j.cognition.2004.10.005

Nazzi, T., Floccia, C., Moquet, B., & Butler, J. (2009). Bias for consonantal information over vocalic information in 30-month-olds: cross-linguistic evidence from French and English. *Journal of Experimental Child Psychology*, *102*(4), 522–537. https://doi.org/10.1016/j.jecp.2008.05.003

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. https://doi.org/10.1177/0963721416655786

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio*, *2*, 221–247.

Newton, C., Chiat, S., & Hald, L. (2008). Evaluation of a novel technique for assessing speech discrimination in children. *Clinical Linguistics & Phonetics*, *22*(4–5), 325–333. https://doi.org/10.1080/02699200801919117

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *Journal of the Acoustical Society of America*, *110*(3), 1598–1605. https://doi.org/10.1121/1.1379078

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, *112*(2), 711. https://doi.org/10.1121/1.1496082

Ohde, R. N., & Haley, K. L. (1997). Stop-consonant and vowel perception in 3- and 4-year-old children. *Journal of the Acoustical Society of America*, *102*(6), 3711–3722. http://www.ncbi.nlm.nih.gov/pubmed/9407663

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421–435. http://www.ncbi.nlm.nih.gov/pubmed/8189202

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/–/ð/ perception: Evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, *109*(5), 2190–2201. https://doi.org/10.1121/1.1362689

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 182–206). Blackwell. https://doi.org/10.1002/9780470757024

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation Norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*(November 1990), 779–798.

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: evidence from second language learning. *Child Development*, *49*(4), 1114–1128.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382.

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language*, *29*(2), 251–274. https://doi.org/10.1017/S0305000902005032

Sussman, J. E. (1991). Stimulus ratio effects on speech discrimination by children and adults. *Journal of Speech and Hearing Research, 34*, 671–678.

Swingley, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words. *Developmental Psychology*, *52*(7), 1011–1023.

Swingley, D., & van der Feest, S. V. H. (2019). A Cross-linguistic Examination of Toddlers' Interpretation of Vowel Duration. *Infancy, 24*(3), 300–317. https://doi.org/10.1111/infa.12280

Templin, M. C. (1957). Appendix IV. Data on articulation of specific speech sounds. In *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press. https://www.jstor.org/stable/10.5749/j.ctttv2st

Treiman, R., Broderick, V., Tincoff, R., & Rodriguez, K. (1998). Children's Phonological Awareness: Confusions between Phonemes that Differ Only in Voicing. *Journal of Experimental Child Psychology*, *68*, 3–21.

Van Ooijen, B. (1996). Vowel mutability and lexical selection in English: evidence from a word reconstruction task. *Memory & Cognition*, *24*(5), 573–583. http://www.ncbi.nlm.nih.gov/pubmed/8870528

Viemeister, N. F., & Wakefield, H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, *90*(1), 858–865.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers, 36*(3), 481–487. https://doi.org/10.3758/BF03195594

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289–1309

Werker, J. F., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: toward a new synthesis. *Annual Review of Psychology, 50*, 509–535. https://doi.org/10.1146/annurev.psych.50.1.509

Zélanti, P. S., & Droit-Volet, S. (2012). Auditory and visual differences in time perception? An investigation from a developmental perspective with neuropsychological tests. Journal of Experimental Child Psychology, 112(3), 296–311. https://doi.org/10.1016/j.jecp.2012.01.003

## Figure Captions

Figure 1. D-prime scores (±standard errors) by experiment and by similarity type (different consonants, such as deev-teev; different vowels, deev-div; or dissimilar words, deev-vush). Maximum *d'* was 4.65, zero represents chance. Jittered points are individual participants.

Figure 2. D-prime values in Experiment 3 by Similarity Type, with standard errors. Maximum *d'* was 4.65, zero represents chance. Jittered points are individual participants.

Figure 3. Estimates of the increased difficulty of similar-sounding over dissimilar-sounding word pairs, with standard errors over participants. Higher values = greater difficulty. Left three bars in each panel are discrimination studies from the current paper, while right two bars after the vertical line (labeled "WrdLrn") are word-learning studies from Creel and Frye (under review). For Experiments 1-2, all participants contributed data to each contrast; for Experiment 3, 14-16 contributed to each; for word learning 1, 8; for word learning 2, 12.

Figure 4. Age effects collapsed across studies, with linear fits in each condition. Maximum *d'* was 4.65, zero represents chance. Shaded regions are 95% confidence intervals.