2017. *Proc Ling Soc Amer* 2, 48:1-15. https://doi.org/10.3765/plsa.v2i0.4105.

Distributional learning on Mechanical Turk and effects of attentional shifts

Emily Moeng[*]

Abstract. This study seeks to determine whether distributional learning can be replicated on an online platform like Mechanical Turk. In doing so, factors that may affect distributional learning, such as level of attention, participant age, and stimuli, are explored. It is found that even distributional learning, which requires making fine phonetic distinctions, can be replicated on Mechanical Turk, and that attention may nullify the effect of distributional learning.

Keywords. distributional learning; Mechanical Turk; replicability, attention

Introduction. This study describes a set of five experiments conducted on Mechanical Turk, an online participant pool run through Amazon. Two methodological and two theoretical contributions are made: (1) experiments requiring fine phonetic distinctions can be replicated on an online platform like Mechanical Turk, (2) changes in methodology to adapt an experiment to one which is not conducted face-to-face in a lab should be kept to a minimum, and (3) attention plays a role in what is known as distributional learning.

1. Background. When language learners are acquiring meaningful sound distinctions in their language, they must decide which acoustic variations cause a change in meaning in their language, and which acoustic variations do not. For example, although an English learner and a Mandarin learner may both hear [i] (***ea**ts*) and [y] (*Lou **ea**ts* in fast speech), the English learner must learn that [i] and [y] are variant pronunciations of a single phonetic category, whereas the Mandarin learner must learn that they belong to two different phonetic categories. The acquisition of phonetic categories has been noted to occur in infants anywhere between the age of 6 months (Kuhl et al., 1992) to 10 months of age (Werker and Tees, 1984; Eilers et al., 1979), and has also been found in adults learning an artificial language (Maye and Gerken, 2000). This section will give a background on artificial language experiments which show that adults make use of statistical cues when determining phonetic categories. In particular, this section describes what is known as **distributional learning**.

1.1. Distributional learning. One of the most widely-cited accounts for how language learners acquire phonetic categories is that language learners make use of distributional learning (for example, Werker et al., 2012). According to this account, language learners map tokens into some phonetic space and make use of the relative frequencies at which tokens cluster in regions of this space to infer the number of phonetic categories in the language he or she is being exposed to. Learners exposed to a bimodal distribution of tokens along some phonetic dimension(s) will infer that there are two phonetic categories, whereas learners exposed to a monomodal distribution will infer that there is only one phonetic category.

Artificial language learning tasks show that learners are capable of the computations necessary to utilize this proposed distributional learning. Maye and Gerken (2000) find that

[*] This material is based upon work supported by the National Science Foundation under Grant No. 1451792. Special thanks to the P-side Research Group at the University of North Carolina, as well as the audience at LSA 91 for their helpful comments and suggestions. I am also grateful to Rachel Hayes-Harb, LouAnn Gerken, and Jessica Maye for sending me and allowing me to use their stimuli. All errors are my own. Author: Emily Moeng, University of North Carolina at Chapel Hill (e.moeng@gmail.com).

participants exposed to an artificial language with a bimodal distribution of tokens ranging between a voiceless unaspirated stop [t] (as in *s**t**eam*, not ***t**eam*) and a pre-voiced stop [d] (***d**eem*) are more likely to say that a pair of syllables differing only by [t] or [d] (both of which sound "d"-like to a naïve English speaker) are "different" syllables in the language they had heard.

Maye and colleague's findings are widely cited (for example, there are 1062 Google Scholar citations of Maye et al. (2002), as of this writing), and have been replicated a number of times. Experimental support has been found for adults (Maye and Gerken, 2000; Maye and Gerken, 2001; Hayes-Harb, 2007; Escudero et al., 2011) and infants (Maye et al., 2002). Attempts to replicate Maye and Gerken's (2000) findings to other stimuli have shown mixed success. Stimuli successfully used in replications include the stop pairs [t] vs. [d], and [k] vs. [ɡ] (Maye and Gerken, 2000; Maye and Gerken, 2001; Maye et al., 2002; Hayes-Harb, 2007); the vowel pairs [a] vs. [ɑ], and [i] vs. [ɪ] (Gulian et al., 2007; Escudero et al., 2011); and the Thai tone pairs [33] and [241] (Ong et al., 2016). However, Peperkamp et al. (2003) failed to replicate these findings when testing fricatives ranging from [ʁ] to [χ] with French-speaking adult participants.

1.2. Possible Effect of Attention on Distributional Learning. Another method that learners may use to form phonetic categories, which I will refer to as **lexicon-based learning**, has also been suggested. Feldman et al. (2011) find evidence that participants make use of word-level phonetic environment to form lexical categories. Specifically, they find that learners exposed to a training phase containing [ɡut**ɑ**], [ɡut**ɔ**], [lit**ɑ**], and [lit**ɔ**] (with no accompanying semantic information) are less likely to claim that [tɑ] and [tɔ] are "different" words in a later test phase, compared to learners who heard only [ɡut**ɑ**] and [lit**ɔ**], but never [lit**ɑ**] and [ɡut**ɔ**] (or vice versa). That is, the identity of the contextual syllable (in this case, [ɡu-] or [li-]) guided the listener in determining how many phonetic categories belonged to the language. If the two sounds [ɑ] and [ɔ] were heard in different lexical contexts, participants placed the two sounds into two separate categories. On the other hand, if the two sounds were heard in the same lexical contexts, speakers placed them into a single category. This type of **lexicon-based learning** can be thought of as an initial assumption that minimal pairs do not exist in the language being learned. (Also see Thiessen, 2007.)

It is possible that participants paying more attention than normal during a distributional learning task actually exhibit two stages of learning: first, through distributional learning, a learner exposed to a bimodal distribution of phones forms two initial phonetic categories. Then in a second step, participants may become hyper-aware that these two phonetic categories are embedded in the same lexical context, possibly collapsing their initial phonetic categories into one phonetic category through lexicon-based learning. The experiments described below suggest that attention does have some type of effect on distributional learning.

1.3. Mechanical Turk. One of the main purposes of these experiments is to determine whether Mechanical Turk is an appropriate platform for conducting further distributional learning experiments. Mechanical Turk ("MTurk") is an online participant pool hosted by Amazon (see Crump et al. (2013) for a discussion concerning the legitimacy of drawing participants for psychological experiments from MTurk). Kleinschmidt and Jaeger (2012) find that MTurk is suitable at least for some speech perception experiments, in an experiment involving stimuli taken from a 9-point continuum between [aba] and [ada]. These experiments will be testing the continuum from voiceless unaspirated [k] (as in *s**k**ill* rather than ***k**ill*) to prevoiced [ɡ] (as in ***g**ill*), a continuum which has been used successfully in distributional learning experiments such as Maye and Gerken (2001) as well as Hayes-Harb (2007).

In attempting to adapt previous distributional learning experiments to this online platform (in which the experimenter cannot be sure that the participant is wearing headphones or even listening), a few methodological considerations arose. Methodological suggestions will be given in the discussion section for those wishing to conduct phonetic experiments on MTurk.

2. Experiment 1. The goal of Experiment 1 is to determine whether distributional learning can be replicated on MTurk. Participants were asked to participate only if they (1) had no known history of speech or hearing impairments, (2) were over the age of 18, (3) were a native speaker of English, (4) had regular access to a computer with an internet connection, and (5) were using a computer able to play audio. Because this experiment was run online rather than face-to-face, only participants using a computer in the United States were allowed to participate to increase the chance that the participant would be a native English speaker. This can be done through a MTurk "qualification," attributes that participants ("Workers," to use the MTurk terminology) on MTurk can obtain. Qualifications used to screen participants in Experiment 1 are as follows:

- Only Workers using a computer in the United States were allowed to participate
- Only Workers who had an approval rating of equal or greater to 90% on all tasks they had completed on MTurk ("HITs") were allowed to participate
- Only Workers who had at least 50 tasks approved by those putting forth tasks ("Requesters") were allowed to participate

2.1. STIMULI. Stimuli consisted of experimental syllables and filler syllables. Experimental syllables were drawn from three 8-point continua ranging between voiceless unaspirated [k] (*s**k**ill*), and [g] (***g**ill*). Continuum points will be referred to as G_1-G_8, where G_1 indicates the most [g]-end of the continuum, and G_8 indicates the most [k] end. Following Maye and Gerken (2000), each of the three continua differed in following vowel: [kɑ]-[gɑ], [kæ]-[gæ], and [kɚ]-[gɚ]. Stimuli were recorded by the experimenter, a native speaker of English.

Recordings were made in a soundproof booth on an Acer netbook at 44100 Hz using a Logitech H390 USB Headset. Recordings were done in Praat (Boersma, 2002), software for speech analysis, synthesis, and manipulation. Before any manipulations were performed, all stimuli (experimental and filler) were scaled to an intensity of 72 Hz in Praat. The experimenter recorded tokens of [sk-] and [g-] followed by each of the three context vowels [ɑ æ ɚ] and removed the [s] portion from the [sk]-initial syllables. These formed the end points of each of the [k]-[g] continua. Prevoicing was then removed from the [g-] syllables. All cuts were made where the waveform crossed 0 Hz to avoid clicks and other unnatural non-speech sounds when splicing sounds together. All splicing was done in Praat. Each of the three pairs of endpoints ([kɑ kæ kɚ] from [sk-] syllables with the [s] portion removed, and [gɑ gæ gɚ] with the pre-voicing removed) were then input into TANDEM-STRAIGHT (Kawahara, 2008), which is a piece of software which creates natural-sounding continua between two sounds. TANDEM-STRAIGHT allows the user to mark any number of landmarks on one spectrogram (for example, the beginning of the steady state of the vowel, the onset of voicing, etc.) that corresponds with a similar landmark on another spectrogram, so that durations between landmarks can be stretched or compressed in the intervening continuum points. TANDEM-STRAIGHT returned 6 continuum points, for a total of 8 continuum points including the endpoints. Following this, the prevoicing which was removed from the [g-] portion of the [gæ] token, which was 140 ms in length was shortened into 8 prevoicing portions ranging from 0-140 ms in length (0, 20, 40… 140). These prevoicing portions were then spliced back onto each of the continuum points (for all three continua), with the 140 ms prevoicing portion being spliced onto the [g]-most end (G_1), the 20 ms prevoicing portion being spliced onto the penultimate of the [k]-most end (G_7), and the [k]-most end (G_8) having no prevoicing spliced on.

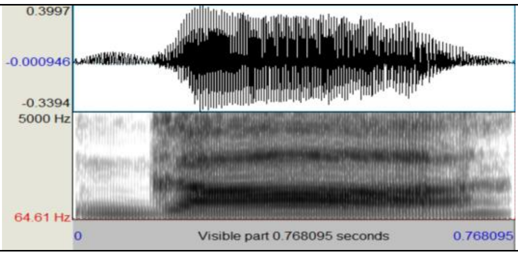
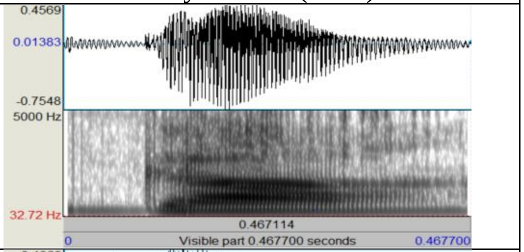
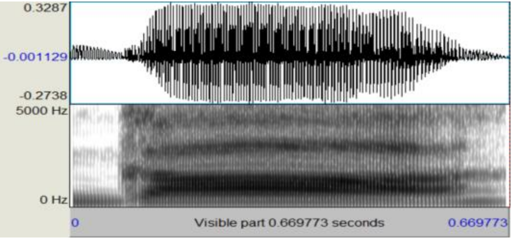
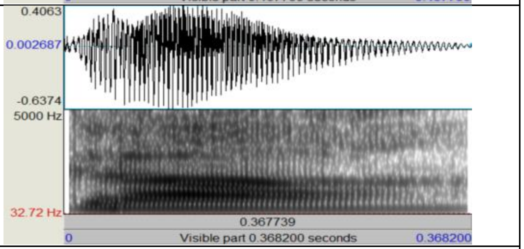
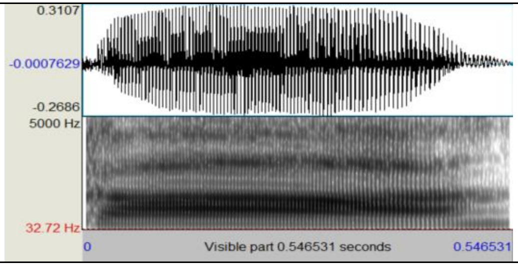
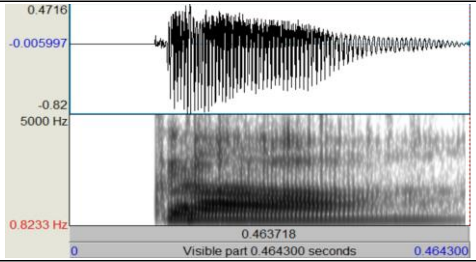
	My stimuli	Stimuli used by Maye and Gerken (2001) and Hayes-Harb (2007)
G_1ɑ		
G_4ɑ		
G_8ɑ		

Figure 1. Waveforms and spectrograms of G_1ɑ, G_4ɑ, and G_8ɑ for stimuli that I created (left), and for stimuli created by Jessica Maye and LouAnn Gerken (right).

All stimuli were judged by a native speaker to sound natural. For visual reference, Figure 1 shows waveforms and spectrograms of the continuum points G_1ɑ, G_4ɑ, and G_8ɑ. For comparison, the G_1ɑ, G_4ɑ, and G_8ɑ stimuli originally used by Maye and Gerken (2001) and also used by Hayes-Harb (2007) are shown on the right. One of the more notable differences between the stimuli on the left and the stimuli on the right is that the stimuli on the left are longer in duration.

2.2. PROCEDURE. Experiment 1 consisted of 4 parts, listed below:

(1) Practice Test
(2) Train, and concurrent Train Catch task
(3) Test, and concurrent Test Catch task
Non-modal Test phase
(4) Questionnaire

During the Practice Test phase, participants were given pairs of English words produced by the same speaker that were either Same Pairs, or Different Pairs. Same Pairs consisted of repetitions of the same word that were different enough to be distinguished as different tokens (e.g. *lock*$_1$ vs. *lock*$_2$). Different Pairs consisted of minimal pairs (e.g. *lock* vs. *rock*, *desk* vs. *disk*). Participants were asked to press the “S” key if the pairs of words that they heard were the “same” word, or the “D” key of they were “different” words. Pairs were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played. Answers were judged as "correct" if participants answered "different" on Different Pairs and "same" on Same Pairs. Participants who answered fewer than 5/8 correct on the Practice Test were excluded. No participant in Experiment 1 failed to meet this criterion.

During the Train phase, participants heard a monomodal or bimodal distribution of phones, depending on which condition they were in. The Bimodal group heard a bimodal frequency of phones of the frequencies shown in the dotted line in Figure 2, and the Monomodal group heard a monomodal frequency of phones of the frequencies shown in the solid line in Figure 2. These frequency values follow those used by Maye and Gerken (2000). This resulted in 16 experimental tokens from each of the three continua (1+1+2+4+4+2+1+1, or 1+4+2+1+1+2+4+1). In addition, three recordings of 8 filler syllables ([fɑ], [fæ], [tɛ], [tej], [mæ], [næ], [sɛ], and [zɛ]) were made. Each of these 24 filler tokens were repeated twice during each Train repetition.

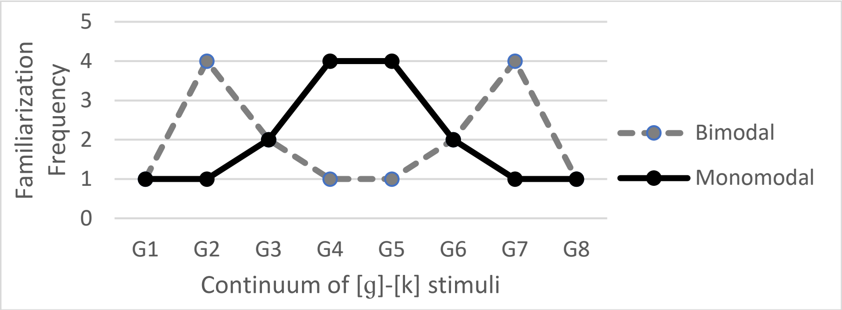
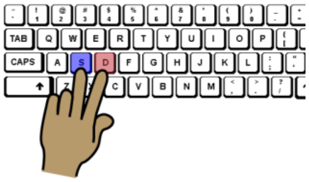


Figure 2. Familiarization frequency of experimental stimuli for the Bimodal (dashed line) and Monomodal (solid line) groups during the Train phase.

In addition, a concurrent non-linguistic Train Catch task was included during the Train phase. The concurrent Train Catch task had the goal of ensuring that participants were wearing headphones and paying attention. To do this, each Train repetition contained 6 randomly-interspersed catches -- 3 one-beep tokens and 3 two-beep tokens. Participants were instructed to press the "1" or "2" keys if they heard one of these beep tokens, to indicate how many beeps they had heard. Beeps were chosen to be at a low enough frequency (50 Hz) that most computer speakers would not pick up on the sound, thereby testing whether participants were wearing headphones or not. Beeps were 340 ms long, and were 140 ms apart for the two-beep tokens. Each Train repetition was repeated 4 times, resulting in a total of 24 Train Catch beeps, 192 fillers, and 192 experimental tokens. If participants answered fewer than 18 out of the 24 Train Catch beeps correctly, their results were not included in the analysis. 13 participants in Experiment 1 did not meet this criterion.

The Test phase was similar to the Practice Test phase, except participants were given pairs of words they had heard in the artificial language they had heard during the Train phase. Again, participants were given pairs of syllables that were either Same Pairs, or Different Pairs. Same Pairs consisted of repetitions of the same exact token for experimental tokens (i.e. G_1a vs. G_1a), or different tokens for control tokens ([fɑ]$_1$ vs. [fɑ]$_2$). Control Same Pairs were judged by the experimenter to sound different enough to be distinguished as separate tokens. Experimental Different Pairs consisted of pairs that occurred on opposite ends of the 8-point continuum, and were equidistant from the midpoint, for experimental tokens (i.e. G_1 vs. G_8, G_2 vs. G_7, G_3 vs. G_6, and G_4 vs. G_5). This differs from previous studies mentioned in the background section in that previous studies only tested the endpoints G_1 vs. G_8. The reason this study tested all points on the 8-point continuum was to keep the Test phase non-modal, so that no further training would occur during the Test phase. This was done because further training and testing had been planned as a follow-up (but will not be reported on here). Experimental Same Pairs consisted of identical tokens on the continuum (i.e. G_1 vs. G_1, G_2 vs. G_2, etc.) All members of the 8-point continuum were equally represented, again, to keep the Test phase non-modal.

In the Test phase, participants were given the following instructions:

> *This next part will be similar to the practice testing you did earlier in English, **but this time it will ask you about the made-up language that you just heard**.*
>
> *Like before, please place one finger over the "S" key and another key over the "D" keys on your keyboard, as shown below.*
>
> 
>
> *Like before, if you think they are repetitions of the same word, press the "**S**" key for "**Same**".*
>
> *If you think they are different words, press the "**D**" key for "**Different**".*

Words in a pair were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played.

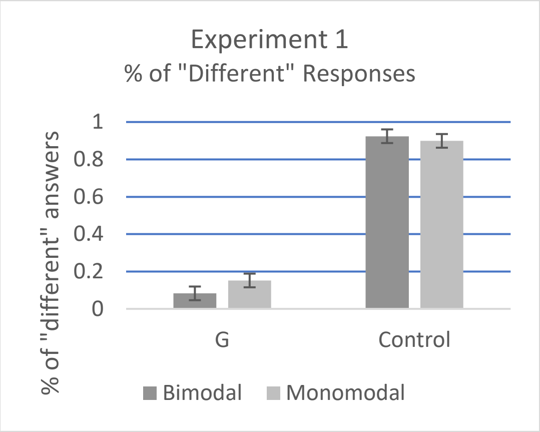
In addition, there was a concurrent Test Catch that occurred during the Test phase. The goal of the Test Catch was to ensure participants were not answering without listening to the test stimuli, since it was possible to answer before the sound had completed playing. The Test Catch consisted of 6 buzzer sounds, 750 ms in length. Test Catch trials were randomly interspersed in the Test phase. During a Test Catch trial, participants would hear a buzzer rather than a word. Participants were asked to not press any keys if they heard a buzzer. If participants pressed a key for more than one of these Test Catch buzzer trials, their results were not included in the analysis. No participants failed to meet this criterion.

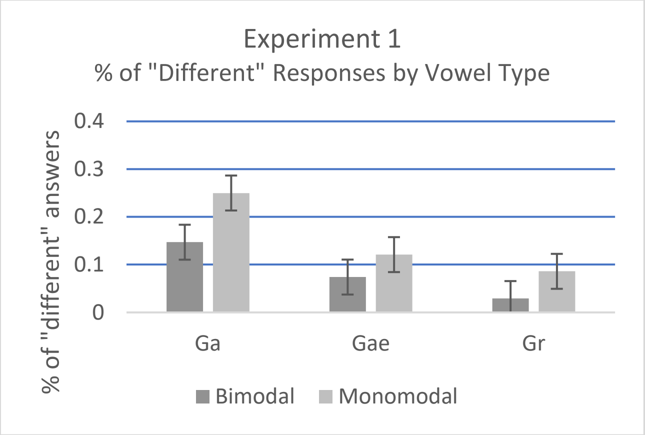
Participants were placed randomly into one of two conditions: a Monomodal group or a Bimodal group. In total, 13 participants were rejected from analysis (some for multiple reasons), leaving 34 in the Bimodal group and 29 in the Monomodal group.

Following experiments will alter the general design of this experiment, including stimuli, the inclusion of the Train Catch beep task, and the age group of participants allowed to participate.

2.3. RESULTS. The percentage of “different” responses that participants gave for Different Pairs was calculated as the total number of trials for which participants answered that the pair of syllables were “different” for experimental Different Pairs encountered during the Test phase, divided by the total number of experimental Different Pairs in the Test phase. All statistical tests were done in R (R Core Team, 2015), using the *aov* method. Significance was set at a level of $p< 0.05$.

A one-way ANOVA was conducted to compare the effect of training condition on the percentage of “different” responses in Bimodal and Monomodal conditions. Participants in the Bimodal group were numerically less likely to respond that experimental Different Pairs were “different” (8.3%) than the Monomodal group (15.2%), although this difference was not significant [F(1)=3.55, p=0.064]. Results for the percentage of “different” responses for all experimental Different Pairs (“G”) and all control Different Pairs (“Control”) can be seen in Figure 3. The Bimodal group responded that control Different Pairs were “different” 92.4% of the time, while the Monomodal group responded “different” 89.9% of the time [F(1)=0.70, *p*=0.407]. Percentage of “different” responses for experimental Different Pairs were also broken down by following vowel, and are plotted in Figure 4. As can be seen, all three context vowel types numerically show the same trend of the Bimodal group answering “different” less often than the Monomodal group, indicating that one vowel context alone was not responsible for the cumulated trend reported above. Differences were not significant.





Figures 3 and 4. Exp 1: Percentage of “different” responses. Error bars indicate 1 standard error.

2.4. DISCUSSION. For all experimental pairs, regardless of following vowel, the Bimodal group was numerically responded that pairs were different *less* often than the Monomodal group. This is surprising given the results of Maye and Gerken (2000), Maye and Gerken (2001), and Hayes-Harb (2007), who find that the group trained on a bimodal distribution of experimental phones is significantly *more* likely to answer that experimental pairs are different compared to the group trained on a monomodal distribution of experimental phones. These differences are not significant though, so it is unclear whether this is a true difference between groups or not, and if so, what this difference might be caused by. Regardless, there could be several reasons for why the results of Experiment 1 do not replicate those of previous studies.

- **Stimuli**: The stimuli used here were created by me, and so are different from stimuli used in previous studies. This will be explored further in Experiments 3a and 3b.
- **Train Catch task**: The Train Catch task may have had the effect of making participants pay more attention than they normally would have to the Train phase, as this task required them to listen actively for beeps, rather than listen passively for the duration of the Train phase. This could cause the Bimodal group to behave more like the Monomodal group in the theoretical two-step learning process described in Section 1.2.
- **Participants**: The population used in my study may be different from the population used in previous studies. While I assume that most participants in previous distributional learning experiments were undergraduate students, this study conducted on MTurk did not specifically draw from the undergraduate population. In particular, only 7 participants reported being between 18-25 years old in the after-experiment questionnaire. 22 reported being 26-35 years old, 14 reported being 36-45 years old, and 19 reported being 46-65 years old. (One participant chose not to answer.)
- **Test Catch task**: The buzzer during the Test phase may have affected responses.
- **Non-Modality of Test phase:** The non-modal nature of the Test phase may also have played some role in how participants were answering.

To see whether participant age was the reason behind the unusual result, Experiment 2a was conducted using only participants who fell within the age range of a typical undergraduate student.

3. Experiment 2a. Experiment 2a was identical to Experiment 1, with the exception that only 18-25 year old Workers were allowed to participate.

3.1. PROCEDURE. Experiment 2a and all remaining experiments used the same qualifications used in Experiment 1, with the following additional qualification:

- Only Workers who were ages 18-25 were allowed to participate

The same exclusion criteria used in Experiment 1 were used in Experiment 2a. The number of participants excluded is detailed below:

- Fewer than 5/8 correct on the Practice Test (9 excluded)
- Fewer than 18/24 correct on the Train Catch task (28 excluded)
- Clicked through more than 1/6 of the Test Catch buzzers (0 excluded)
- Reported not being a native speaker of English (0 excluded)
- Reported a history of a speech or hearing disorder (2 excluded)

In total, 36 participants were rejected from analysis (with some being excluded for multiple reasons), leaving 25 in the Bimodal group and 31 in the Monomodal group. Other than the additional qualification, the procedure and stimuli were identical to that of Experiment 1.

3.2. RESULTS. As with Experiment 1, the percentage of "different" answers was calculated. For reasons of space, graphs summarizing the results of all five experiments will be given in Section 6. A one-way ANOVA revealed that there was no significant difference between the percentage of "different" answers in the Bimodal group (13.7%) and the Monomodal group (15.1%) for experimental Different Pairs [F(1)=0.09, *p*=0.771]. There was also no significant difference between the percentage of "different" answers in the Bimodal group (85.7%) and the Monomodal group (86.0%) for control Different Pairs [F(1)=0.01, *p*=0.937].

3.3. DISCUSSION. There was no significant difference between any of the pairs compared. It may be the case that age of participant plays a factor in distributional learning, as Experiment 1 found the unusual result that the Bimodal group was nearly half as likely to answer "different" than the Monomodal group (although not significantly so), whereas Experiment 2a found little difference between the Bimodal and Monomodal groups. It is still unclear whether distributional learning can be replicated on MTurk as Experiment 2a did not yield any significant differences between groups. All remaining experiments will minimize the number of methodological changes that were made for the purposes of these two experiments, Experiments 1 and 2a, and the methodology of Maye and Gerken (2000).

4. Experiment 2b. The above results may indicate that the inclusion of the Train Check beeps and the Test Check buzzer, or the non-modal nature of the Test phase, somehow negates the effects of distributional learning. To see if this is the case, Experiment 2b removes the Train Check beeps. Because these beeps were removed, a Sound Check was included at the beginning, as it was still desired that participants be encouraged to wear headphones for the duration of the experiment. To make up for the lack of Train Check beeps, participants were also asked in the Questionnaire whether or not they were wearing headphones, and how much attention they were paying. Questionnaire responses will be discussed in Section 7.

4.1. PROCEDURE. Experiment 2b consisted of 5 parts, listed below:

(1) Sound Check
(2) Practice Test
(3) Train (NO concurrent Train Catch task)
(4) Test (NO concurrent Test Catch task)
Only continuum points 1 vs. 8 were tested
(5) Questionnaire

The Sound Check consisted of 3 one-beep tokens and 3 two-beep tokens, randomly interspersed. Participants were instructed to press the "1" or "2" keys if they heard one of these beep tokens, to indicate how many beeps they had heard. The beeps used were the same low-frequency beeps as those described in Experiments 1 and 2a, thereby testing whether participants were wearing headphones or not for the duration of the Sound Check.

The Practice Test and Train phase were identical to that of Experiments 1 and 2a, with the exception that the Train phase did not contain the Train Catch task. As before, each Train repetition was repeated four times, resulting in a total of 192 experimental tokens and 192 filler tokens.

Two alterations were made to the Test phase. First, the concurrent Test Catch buzzer was removed. Second, in Experiments 1 and 2a, the Test phase was kept non-modal, in a departure from Maye and Gerken (2000). In this and all remaining experiments, the Test phase followed the methodology of Maye and Gerken, and only tested the endpoints, G_1 and G_8. Therefore, experimental Same Pairs consisted of G_1 vs. G_1 or G_8 vs. G_8, while experimental Different Pairs consisted of G_1 vs. G_8. Each Test repetition consisted of 4 experimental pairs (2 Same Pairs and 2 Different Pairs), for each of the three vowel contexts, resulting in 12 experimental pairs. Each Test repetition also contained 4 control Same Pairs and 4 control Different Pairs. There were two repetitions of each Test phase resulting in a total of 24 experimental pairs and 16 control pairs. Again, participants were instructed to press the "S" key if the pairs of words that they heard were the "same" word, or the "D" key of they were "different" words. Pairs were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played.

The following exclusion criteria were used in Experiment 2b:

- Fewer than 5/6 on the pre-experiment Sound Check (12 excluded)
- Fewer than 5/8 correct on the Practice Test (0 excluded)
- Reported not being a native speaker of English (0 excluded)
- Reported a history of a speech or hearing disorder (1 excluded)

In total, 13 participants were rejected from analysis, leaving 27 in the Bimodal group and 34 in the Monomodal group.

4.2. RESULTS. Again, the percentage of "different" answers that participants gave for Different Pairs was calculated. Graphical results are given in Section 6. A one-way ANOVA revealed that the Bimodal group responded "different" significantly more often (25%) than the Monomodal group (15.2%) for experimental Different Pairs [F(1)=4.00, p=0.050]. There was no significant difference between the percentage of "different" answers in the Bimodal group (88.4%) and the Monomodal group (93.4%) for control Different Pairs [F(1)=1.78, p=0.187].

4.3. DISCUSSION. Experiment 2b is the first indication that distributional learning *can* be replicated on an online platform such as MTurk. Experiments 3a and 3b seek to determine whether the addition of the beeps during the Train phase was (at least partially) responsible for the lack of significance in Experiments 1 and 2a.

5. Experiments 3a and 3b. The goal of Experiments 3a and 3b is to determine whether the above results, that the distributional effect only becomes significant when there is no Train Check beep task, are particular to the stimuli that I used. Experiments 3a and 3b follow the methodology of Maye and Gerken (2001) and Hayes-Harb (2007) as closely as possible, using their exact same stimuli and procedure, with the small addition of the Sound Check task (used in Experiment 2b) preceding both experiments.[1] In addition, Experiment 3a included the Train Check beeps described in Experiments 1 and 2a. The following exclusion criteria were used:

- Fewer than 5/6 on the pre-experiment Sound Check (Exp 3a: 14 excl, Exp 3b: 7 excl)
- Fewer than 5/8 correct on the Practice Test (Exp 3a: 0 excl, Exp 3b: 1 excl)
- Fewer than 18/24 correct on the Train Check task (Exp 3a: 9 excl, Exp 3b: N/A)
- Reported not being a native speaker of English (Exp 3a: 0 excl, Exp 3b: 0 excl)
- Reported a history of a speech or hearing disorder (Exp 3a: 0 excl, Exp 3b: 1 excl)

In total, 14 participants were rejected from analysis from Experiment 3a, leaving 28 in the Bimodal group and 31 in the Monomodal group. 7 participants were rejected from analysis from Experiment 3b, leaving 21 in the Bimodal group and 27 in the Monomodal group.

5.1. PROCEDURE. The procedure of Experiment 3b was identical to that followed by both Maye and Gerken (2000) and the phonetic learning part of the experiment run by Hayes-Harb (2007), but was preceded by the Sound Check task described in Experiment 2b. As was the case for Experiment 2b, this experiment consisted of a Sound Check, Practice phase, a Train phase, and a Test phase, followed by a Questionnaire. The procedure and stimuli of Experiments 3a and 3b were identical, except the concurrent Train Check beep-monitoring task was included in Experiment 3a. Each Train repetition consisted of 16 experimental tokens for each of the three vowel contexts, two repetitions of four separate tokens of 6 fillers [mɑ mæ mɚ lɑ læ lɚ], and, for Experiment 3a, 3 one-beep tokens and 3 2-beep tokens. Each Train repetition was repeated 4 times, resulting in a total of 192 fillers, 192 experimental tokens, and, for Experiment 3a, 24 beep tokens. As noted earlier, stimuli come from those used by Maye and Gerken (2001) as well as Hayes-Harb (2007). Examples of select continuum points can be seen in Figure 1.

5.2. RESULTS. For Experiment 3a, there was no significant difference between the percentage of “different” answers in the Bimodal group (14.3%) and the Monomodal group (12.6%) for experimental Different Pairs [F(1)=0.12, *p*=0.730]. There was also no significant difference between the percentage of “different” answers in the Bimodal group (94.3%) and the Monomodal group (94.6%) for control Different Pairs [F(1)=0.00, *p*=0.943]. For Experiment 3b, there was a significant difference between the percentage of “different” answers in the Bimodal group (13.5%) and the Monomodal group (4.3%) for experimental Different Pairs [F(1)=7.86, *p=0*.007], with the Bimodal group responding “different” more often than the Monomodal group. There was no significant difference between the percentage of “different” answers in the Bimodal group (90.9%) and the Monomodal group (95.1%) for control Different Pairs [F(1)=1.71, *p*=0.198].

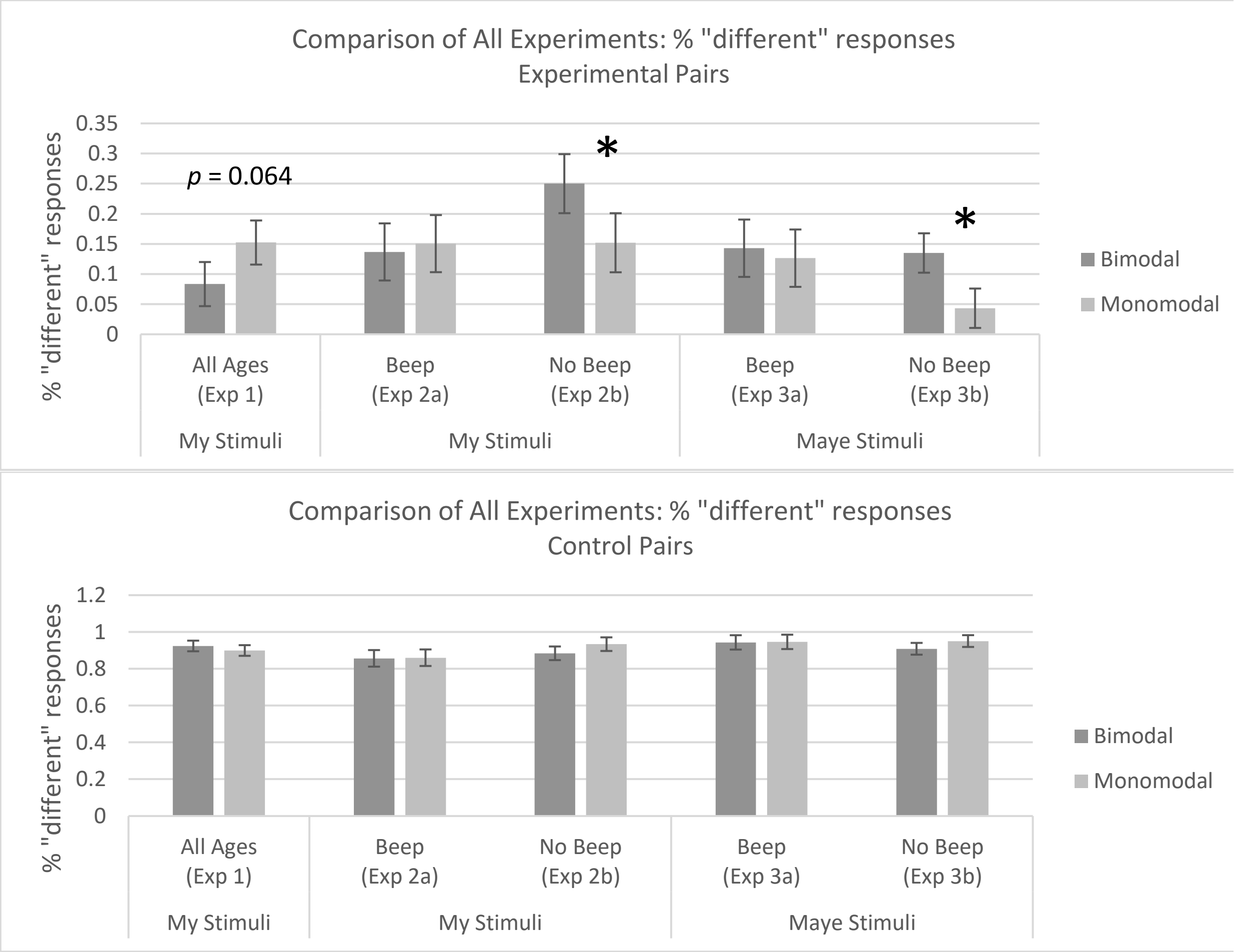
5.3. DISCUSSION. The addition of beep tokens during the Train phase seems to have a negating effect on distributional learning as Experiment 3a, which only differed from Experiment 3b in the inclusion of the beep tokens, failed to show a significant effect of training condition.

[1] Many thanks to Rachel Hayes-Harb, LouAnn Gerken, and Jessica Maye for sending me and allowing me to use their stimuli.

6. Results. This section will briefly summarize the procedure and results of the five experiments described above. Table 1 summarizes the differences between procedures of all five experiments. As can be seen in Figure 5, only “No Beep” experiments found that the Bimodal condition was significantly more likely to answer “different” than the Monomodal condition.

	Experiment 1	*Experiment 2a*	*Experiment 2b*	*Experiment 3a*	*Experiment 3b*
Stimuli	Created by author	Created by author	Created by author	Originally used by Maye and Gerken (2001)	Originally used by Maye and Gerken (2001)
Procedure			Sound Check	Sound Check	Sound Check
	Practice Test	Practice Test	Practice Test	Practice Test	Practice Test
	Train phase + Train Catch beeps	Train phase + Train Catch beeps	Train phase	Train phase + Train Catch beeps	Train phase
	Test phase (non-modal) + Test Catch buzzer	Test phase (non-modal) + Test Catch buzzer	Test phase (1 v. 8)	Test phase (1 v. 8)	Test phase (1 v. 8)
	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire

Table 1: Summary of procedures for Experiments 1, 2a, 2b, 3a, and 3b.



Figures 5 and 6. Comparison of all experiments. Summary of percentage of “different” responses made by participants for Different Experimental Pairs and for Different Control Pairs.

7. Discussion. This study makes two methodological contributions, and one theoretical conclusion, which will be discussed in this section.

7.1. METHODOLOGICAL CONSIDERATIONS WHEN TESTING ON MTURK. This study concludes that (1) it is possible to replicate results of studies that require fine phonetic distinctions on MTurk, and (2) changes made to adapt an experiment to MTurk should be kept to a minimum.

Regarding (1), it is believed that the inclusion of a short task confined to the beginning of an MTurk experiment, particularly a task which requires participants to listen for sounds which most computer speakers cannot pick up (50 Hz beeps in this case), is sufficient encouragement to participants to wear headphones for the duration of the experiment. Questions were included in the after-experiment questionnaire for Experiments 2b, 3a, and 3b to determine whether participants were actually wearing headphones. Participants were specifically told that their answers would not affect their payment. Participants were asked to answer whether they were (a) wearing headphones the entire time, (b) wearing headphones most of the time, (c) wearing headphones some of the time, or (d) not wearing headphones at all. Most of the participants reported that they were wearing headphones the entire time, with only a few reporting wearing headphones only "most of the time," (one each in Experiments 2b, 3a, and 3b), and no participants reporting wearing headphones "some of the time" or not wearing headphones at all.

In addition, the questionnaire in Experiments 2b, 3a, and 3b asked participants how much attention they were paying to various phases of the experiment, and whether or not they would pay more, the same, or less attention if this same experiment were being conducted in a lab setting. Of those three experiments, Experiments 2b and 3b did NOT contain the Train Catch beeps, whereas Experiment 3a did. Participants were given the following options, for both the Train phase and the Test phase: (a) I focused all of my attention on this portion of the experiment, (b) I mostly paid attention, (c) I was not paying very much attention, or (d) I paid very little attention. Very few participants reported "not paying very much" attention, or reported paying "very little" attention. A glance at how participants answered to options (a) and (b) suggests that the experiments which did contain the Train Catch beeps increased overall attention for both the Train and Test phases. A breakdown of answers is shown below:

		Did not contain Train Catch beeps		Contained Train Catch beeps
		Experiment 2b	Experiment 3b	Experiment 3a
Train phase	*I focused all of my attention on this portion of the experiment*	58%	52%	68%
	I mostly paid attention	38%	47%	30%
Test phase	*I focused all of my attention on this portion of the experiment*	84%	85%	95%
	I mostly paid attention	15%	15%	5%

Table 2: Questionnaire responses regarding participants' attention.

It appears as though participants in all experiments reported paying more attention during the more-active Test phase, with a greater percentage of participants reporting focusing "all" of their attention to that portion (specifically, 84-95%). In addition, a smaller percentage of participants in Experiments 2b and 3b, which did not contain the Train Catch beeps, reported focusing "all" of their attention on the Train phase (specifically, 52-58%), compared to participants in Experiment 3a, which *did* contain the Train Catch beeps (68%). The increased attention seemed to carry over from the Train phase to the Test phase -- a greater percentage of non-beep participants (95%) reported focusing all of their attention than beep participants (84-85%), even though the Test phase was identical in all three experiments.

In attempting to replicate distributional learning on MTurk, it was initially believed that certain changes needed to be made to ensure that participants were paying attention and wearing headphones. However, results of this study suggest that forcing participants to pay *too* much attention may have had unintended effects, at least for distributional learning. Therefore, it is suggested that only minimal changes, such as including a short sound check at the beginning of the experiment, should be made when attempting to replicate studies on MTurk.

7.2. EFFECT OF ATTENTION ON DISTRIBUTIONAL LEARNING. This study found that increased attention likely affects distributional learning. The results of the experiment pairs Experiments 3a and 3b, as well as those of Experiments 2a and 2b, seem to indicate that the effect of distributional learning is either a weakened effect, is not in effect at all, or is cancelled out by some other effect. While Experiments 2a and 2b differed on a few other fronts, Experiments 3a and 3b only differed in one aspect: Experiment 3a contained an extra Train Catch beep-monitoring task during the Train phase – everything else was kept the same. And yet the “No Beep” versions of each of these pairs of experiments (Experiments 2b and 3b) yielded a significant difference between the Bimodal and Monomodal groups, whereas the “No Beep” versions did not.

As suggested in the previous section, the inclusion of this extra beep-monitoring task seemed to have the effect of increasing participants’ overall attention to the entire experiment. It was suggested in the background section that greater attention may result in a two-step learning process of phonetic categories: distributional learning, followed by lexical learning. That is, the Bimodal group initially creates two proto-phonetic categories via distributional learning, which is then followed by the awareness that each of those proto-phonetic categories always occur in the same lexical context (i.e. only ever preceding [ɑ æ ɚ]). It is possible that this is the cause for the difference between the “Beep” and “No Beep” experiments. On the other hand, if this were the case we would expect the Bimodal group in the “Beep” version to answer that experimental pairs were different less often than in the “No Beep” version. This seems to be the case for Experiments 2a and 2b, but not for Experiments 3a and 3b, where the Bimodal group stays about constant between the “Beep” and “No Beep” experiments (rather, it is the Monomodal group that answers “different” more often in the “Beep” experiment than in the “No Beep” experiment).

Another possibility is that the addition of the beep-monitoring task actually decreased participants’ attention to the syllables they were being exposed to, and that the *lack* of attention to the training data is responsible for the absence of distributional learning. A similar finding has been made for speech segmentation. Learners are able to make use of transitional probabilities to segment a stream of speech (Saffran et al. 1996), but if their attention is diverted, they exhibit less learning. Toro et al. (2005) and Saffran et al. (1997) both exposed learners to a speech stream in a segmentation experiment. Both found that learners were able to successfully make use of transitional probabilities if attention was diverted to a task with little demand that did not make use of the same sensory modality (like drawing while listening to the speech stream). However, Toro et al. found that more demanding tasks or tasks that made use of the same sensory modality (that is, a concurrent auditory task) negatively affected participants’ abilities to segment speech using transitional probabilities. More research is needed to determine whether the lack of distributional learning found in the “Beep” experiments reported here is due to more or less attention being paid to the training data.

8. Conclusion. This study finds that attention plays a role in distributional learning, and makes several methodological suggestions for those wishing to run experiments on MTurk. There may also be some “anti”-distributional learning occurring in Experiment 1, even though differences between conditions were not significant. This study did not delve into what factors may have contributed to this effect, but tentatively suggests that it may be some effect of age.

9. Appendix

		Bimodal % “different” answers (N=34)	Monomodal % “different” answers (N=29)	*p*-value	F-value
Experiment 1	Experimental pairs (All G pairs combined)	8.3%	15.2%	0.064	3.55
	Gɑ	14.7%	25.0%	0.140	2.24
	Gæ	7.4%	12.1%	0.242	1.40
	Gr	2.9%	8.6%	0.104	2.73
	Control	92.4%	89.9%	0.407	0.70
		Bimodal % “different” answers (N=25)	Monomodal % “different” answers (N=31)	*p*-value	F-value
Experiment 2a	Experimental pairs (All G pairs combined)	13.7%	15.1%	0.771	0.09
	Gɑ	18.0%	18.5%	0.944	0.01
	Gæ	15.0%	14.5%	0.939	0.01
	Gr	8.0%	12.1%	0.426	0.64
	Control	85.7%	86.0%	0.937	0.01
		Bimodal % “different” answers (N=27)	Monomodal % “different” answers (N=34)	*p*-value	F-value
Experiment 2b	Experimental pairs (All G pairs combined)	25.0%	15.2%	0.050*	4.00
	Gɑ	42.6%	27.9%	0.094	2.90
	Gæ	19.4%	12.5%	0.233	1.45
	Gr	13.0%	5.1%	0.096	2.86
	Control	88.4%	93.4%	0.187	1.78
		Bimodal % “different” answers (N=28)	Monomodal % “different” answers (N=31)	*p*-value	F-value
Experiment 3a	Experimental pairs (All G pairs combined)	14.3%	12.6%	0.730	0.12
	Gɑ	15.2%	13.7%	0.83	0.05
	Gæ	10.7%	8.9%	0.723	0.13
	Gr	17.0%	15.3%	0.836	0.04
	Control	94.3%	94.6%	0.943	0.01
		Bimodal % “different” answers (N=21)	Monomodal % “different” answers (N=27)	*p*-value	F-value
Experiment 3b	Experimental pairs (All G pairs combined)	13.5%	4.3%	0.007**	7.86
	Gɑ	14.3%	3.7%	0.062	3.67
	Gæ	13.1%	2.8%	0.008**	7.77
	Gr	13.1%	6.5%	0.236	1.44
	Control	90.9%	95.1%	0.198	1.71

References

Boersma, Paul. 2002. Praat, a system for doing phonetics by computer. *Glot International* 5. 341-345.

Cristià, Alejandrina, Grant L. McGuire, Amanda Seidl & Alexander L. Francis. 2011. Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics* 39(3). 388-402.

Crump, Matthew, John McDonnell, and Todd Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8(3). e57410.

Gulian, Margarita, Paola Escudero, & Paul Boersma. 2007. Supervision hampers distributional learning of vowel contrasts. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*. 1893-1896. Saarbrücken: University of Saarbrucken.

Eilers, Rebecca E., William Gavin, & Wesley R. Wilson. 1979. Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development* 50(1). 14-18.

Escudero, Paola, Titia Benders, & Karin Wanrooij. 2011. Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America* 130(4). EL206-EL212.

Feldman, Naomi, Myers, Emily, White, Katherine, Griffiths, Thomas, & Morgan, James. 2011. Learners use word-level statistics in phonetic category acquisition. *Proceedings of the 35th Boston University Conference on Language Development*.

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

Kawahara, Hideki, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno. 2008. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *ICASSP 2008 IEEE*. 3933-3936.

Kleinschmidt, Dave, and T. Florian Jaeger. 2012. A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. *CogSci*.

Kuhl, Patricia K., Karen A. Williams, Francisco Lacerda, Kenneth N. Stevens & Björn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255(5044). 606-608.

Maye, Jessica, and LouAnn Gerken. 2000. Learning phonemes without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development, Vol. 2*. 522-533.

Maye, Jessica, and LouAnn Gerken. 2001. Learning phonemes: How far can the input take us. *Proceedings of the 25th Annual Boston University Conference on Language Development, Vol. 1*. 480. Somerville, MA: Cascadilla Press.

Maye, Jessica, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3): B101-B111.

Ong, Jia Hoong, Denis Burnham, Catherine J. Stevens, and Paola Escudero. 2016. Naïve Learners Show Cross-Domain Transfer after Distributional Learning: The Case of Lexical and Musical Pitch. *Frontiers in Psychology* 7.

Peperkamp, Sharon, Michéle Pettinato & Emmanuel Dupoux. 2003. Allophonic variation and the acquisition of phoneme categories. In B. Beachley, A. Brown, & F. Conlin (eds.), *Proceedings of the 27th Boston University Conference on Language Development*. 650-661. Sommerville, MA: Cascadilla Press.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org.

Saffran, Jenny, Elissa Newport, and Richard Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35 (4). 606-621.

Saffran, Jenny, Elissa Newport, Richard Aslin, Rachel Tunick, & Sandra Barrueco. 1997. Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science* (8). 101–105.

Thiessen, Erik D. 2007. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language* 56(1). 16-34.

Toro, Juan M., Scott Sinnett, and Salvador Soto-Faraco. 2005. Speech segmentation by statistical learning depends on attention. *Cognition* 97(2). B25-B34.

Werker, Janet F. & Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1).