Learners' Generalization of Alternation Patterns from Ambiguous Data

Bingzi Yu[1], Shuang Zheng[2], Youngah Do[2]
[1]Massachusetts Institute of Technology, [2]The University of Hong Kong

1 Introduction

A single set of language data could be compatible with multiple generalizations, varying from the ones based on specific segments to the ones based on abstract representations. To successfully acquire a lexicon, beyond memorizing segmental sequences, learners need to find regularities, including absolute or optional patterns and their categorical or gradient distributions, and need to extend them to account for novel data (Aslin & Newport, 2012; Marcus et al., 1999). Previous studies have explored how learners tend to generalize phonological regularities among logically possible ones (Chambers et al., 2003; Moreton, 2002). For example, it was found that learners prefer to generalize patterns based on subsegmental features (Finley & Badecker, 2009).

Making generalizations is an essential part of language acquisition, and multiple theoretical frameworks have been developed to predict how learners choose the optimal grammar throughout the acquisition process. First, early in SPE (Chomsky & Halle, Morris, 1968), the evaluation metric posited that learners would prefer a grammar with the smallest size. The size of a grammar depends on the number of underlying representations in the lexicon and the set of generalizations used to describe them. For the latter, they claimed that learners acquire the simplest generalizations, defined by the fewest phonological features. However, as Braine (1971) and Baker (1979) pointed out, solely guided by grammar economy, children might fail to acquire the context for certain optional rules[1]. Second, in contrast to the evaluation metric, the Subset Principle (Dell, 1981; Hale & Reiss, 2008) proposed that when presented with competing grammars, learners will always prefer the one that generates a subset of the language produced by the other. Consequently, this strategy leads learners to favor more restrictive grammars involving more complex generalizations. Third, another evaluation approach, originated from the Minimum Description Length (Rissanen, 1978), considers both economy and restrictiveness, stating that learners accept a new generalization only when it adds the minimal necessary stipulation to the existing grammar (Rasin et al., 2021). All these proposals depict the broader picture of grammar formation during phonological acquisition, and the predictions about learners' preference for selecting specific generalizations follow accordingly, giving rise to two major hypotheses in terms of simplicity.

The first hypothesis is in line with the Subset Principle, suggesting that learners might favor the most specific, thus complex generalization. Gerken (2006) found that eight-month-old infants preferred the more specific generalization over the simpler one on some occasions. After being trained with trisyllabic A*di*A words, infants showed different looking time for A*di*A but not ABA words against unfamiliar items in ABB shape during the testing phase[2]. This result suggested that infants failed to generalize the more abstract ABA pattern when the more specific A*di*A generalization was available. However, the preference for the complex

* Bingzi Yu and Shuang Zheng contributed equally to this work and share first authorship. We would like to thank the AMP reviewers, participants at AMP 2024, and audiences at HKU Linguistics Seminar and MIT Phonology Circle, for all the invaluable feedback.

[1] The argument assumes that a grammar without specified contexts for optional rules is smaller and therefore simpler than one that includes them. However, the hypothesis based on the evaluation metric wrongly predicts that children would prefer the grammar without optional rules.

[2] During training, A*di*A words were presented with a fixed [di] syllable and four different A syllables, e.g., [ledile] and [widiwi]. In the testing phase, ABA words involved varying B syllables and unseen A syllables, e.g., [bakoba] and [gapoga].

Proceedings of *AMP 2023/2024*

generalization was not consistent. For example, in Gerken's first experiment, infants trained with ABA words containing varying B syllables successfully distinguished novel ABA items from unfamiliar ABB items, suggesting they generalized beyond the specific input syllables. Moreover, such preference could be confounded with other factors. For example, as discussed by Aslin and Newport (2012) in relation to similar findings from their work (Reeder et al., 2010, 2009) and Gerken's study, the inconsistency of contextual cues may also inhibit further generalizations. In addition, it has been pointed out that the Subset Principle needs extra mechanisms to restrict the preference for the most complex generalization, otherwise the generalized output would simply mirror the input segments, thereby restraining the acquisitional process (Marcus et al., 1999; Rasin et al., 2021).

The second hypothesis advocates the preference for the simplest generalizations, echoing Chomsky and Halle's discussion based on the evaluation metric. Durvasula and Liter (2020) provided support for this hypothesis by exploring phonotactic learning in laboratory settings. They conducted an Artificial Language Learning (ALL) experiment to test participants' preference for phonotactic generalizations when given ambiguous patterns. During the training stage, participants were exposed to a language (e.g., {[fisu], [vuzi], [pita], [badi]…}) in which the two cooccurring consonants always agreed in voicing as well as in continuancy. The language was compatible with three major feature-based generalizations: (1) Voicing harmony; (2) Continuancy harmony; and (3) Voicing + Continuancy harmony. The first two harmony patterns were defined to be simple as they utilized a single feature, either voicing or continuancy, whereas the third one was more complex with two features. They found that learners were able to keep track of multiple generalizations but only those with the fewest features, i.e., (1) Voicing harmony and (2) Continuancy harmony. The finding that learners could form more than one generalization simultaneously was consistent with some previous work. For instance, learners show gradient acceptability to different phonotactic violations, suggesting that they have formed multiple generalizations based on natural classes (Albright, 2009). In alternation learning, participants were also found to manage multiple rules to account for both the dominant pattern and exceptions (Finley, 2021). With regard to simplicity, the result of Durvasula and Liter was against the Subset Principle as participants did not form the most complex generalizations, suggesting a simplicity bias for phonotactic learning.

Against this background, it is necessary to investigate how learners tend to generalize from other phonological patterns, namely morphophonological alternation. The interaction between phonotactics and alternation has been extensively studied in previous literature, and there is no consensus on whether they share unified or distinctive learning mechanisms. From a theoretical perspective, according to the rule-based theory of phonology, alternation is accounted for by re-write rules while phonotactics is subject to Morpheme Structure Constraints; on the other hand, Optimality Theory (Smolensky & Prince, 1993) aims at unifying both phonotactics and alternation using the same set of constraints. Adopting experimental methods, researchers have found that the two acquisition processes can mutually influence one another (Pater & Tessier, 2005; Pizzo & Pater, 2016). In one direction, phonotactic generalization could aid in alternation learning. For instance, Pater and Tessier (2005) showed that it was easier to learn a language whose alternation patterns were motivated by the phonotactic rules in the participants' native language. A similar facilitation effect was found in Chong (2021), in which vowel harmony across morpheme boundary was learned the best when the stems also exhibited categorical harmony though learners were more conservative without overt alternation evidence. In the other direction, Pizzo and Pater (2016) provided mixed evidence for the hypothesis that alternation knowledge affects phonotactic judgment. Do and Yeung (2021) suggested that whether there exists a clear link between the two learning procedures could be language specific. They found that speakers of Cantonese, a language that does not exhibit a clear connection between its phonotactics and alternation, acquired alternation patterns rather independently from the phonotactic input in experiments.

In addition to the investigations on their connections, generalization preferences of phonotactics and alternation have been independently studied. For example, while experiments testing phonotactic generalizations and alternation generalizations both reported mixed evidence for the substantive bias hypothesis (see a detailed review in Zheng & Do, 2025), specific preferences in the two learning domains diverge. For example, preferences of word-final segments are subject to biases towards naturalness. However, final nasals with higher perceptual salience are favored over final obstruents when generalizing patterns of phonotactics (Greenwood, Experiment 3, 2016), while final devoicing is favored over final nasalization when generalizing patterns of alternation as they are more similar to the underlying voiced segments (Albright & Do, 2017; White, 2014). Considering these mixed hypotheses and observations on the generalization preferences for phonotactics and alternation, we aim to examine how learners generalize when exposed to alternation data and compare the results to patterns of phonotactic generalizations.

The current study adopted the ALL paradigm to study generalization during alternation learning in face of ambiguous data. For consistency, we defined simplicity in this paper by the number of features, identifying the simplest generalization as the one with the minimal feature set. Particularly, we raised three research questions: First, do learners prefer the simple generalization with fewer features or the complex generalization? Second, do learners focus on only a single generalization or keep track of multiple generalizations? Third, answers to the first two questions will shed light on the comparison between phonotactic and alternation learning: do learners show a similar preference for the two learning processes? In brief, we observed that, partially consistent with Durvasula and Liter's (2020) finding regarding phonotactic learning, learners were only able to make a single simplest generalization for alternation patterns.

2 Method

2.1 *Participants* The experiment was released via Prolific (*Prolific*, 2014). 250 adult participants who self-identified as native English speakers were recruited. Upon completion, participants received £3, as recommended by Prolific. Excluding 50 participants who failed to complete the experiment, data from 200 participants (86 females, 111 males, and 3 non-binary genders) entered the analysis. Most participants were aged above 30 years old (N = 134).

2.2 *Stimuli* The experiment included two phases: the training phase and the testing phase. The stimuli used in both phases followed the design in Durvasula and Liter's work (2020). In the training phase, participants were trained on a language comprising 72 nonce words, 36 VCV stems, and 36 pairs of stems and VCV-CV suffixal forms. Vowels were drawn from the set of {a, i, u}, and the consonants were selected from {b, p, d, t, v, f, z, s}. For a training stem, the suffixal form was created by adding a CV suffix, in which the vowel was always /o/, and the consonant was selected from {g, k, ʒ, ʃ}. The suffixal form always exhibited an agreement in voicing and continuancy between the two consonants across the morpheme boundary. For example, [adi] & [adi-go] and [ifu] & [ifu-ʃo] are legal pairs in the language, as /d/ and /g/ are both voiced and non-continuant obstruent, and /f/ and /ʃ/ are both voiceless and continuant obstruent. We tried to ensure that voicing and continuancy were the only two features available for generalizations, but there still existed other perceivable features such as stridency and spread glottis[3]. However, it turned out that these features did not contribute to participants' generalizing preferences in a systematic manner (see Footnote 6 in the Result section for more details). Every stem and its plural form were accompanied by a picture of a non-existing object or a group of them to assist participants in understanding the concept of singular and plural. The object images were downloaded from the Novel Object and Unusual Name (NOUN) database (Horst & Hout, 2016).

For the testing phase, we adopted both implicit and explicit modes and each participant was randomly directed to either of the modes. In the implicit test, participants were asked to decide whether a suffixal form belonged to the language; In the explicit test, they were asked to decide whether the suffixal form was correct given a stem form. The second mode was more explicit in the sense that participants' attention was particularly guided towards the suffix alternation, which exhibited voicing or continuancy harmony. For both modes, the stimuli were separated into the following five categories, each containing 12 items, resulting in 60 testing trials in total:

1. HarmonyOld: training items, e.g., [adigo].
2. HarmonyNew: items exhibiting voicing and continuancy harmony with unseen consonant sequences, e.g., [apiko].
3. OnlyVoi: items agreed in voicing but not continuancy, e.g., [atiʃo].
4. OnlyCont: items agreed in continuancy but not voicing, e.g., [afiʒo].
5. Disharmony: items disagreed in voicing and continuancy, e.g., [isugo].

[3] With an accent of British English, voiceless stops {p, t} were produced with slight aspiration as {p^h, t^h} in synthesized speech.

To make sure that participants do not memorize consonant sequences without utilizing any featural generalizations, we withheld certain consonant combinations during the training phase and used them as HarmonyNew stimuli in the testing phase. To ensure exposure to all segments, half of the consonants from each natural class (e.g., {b, t, f, z}) were shown in singular forms only, and the other half (e.g., {d, p, s, v}) were presented in singular-plural pairs. This design resulted in four ways to group consonants in the training stimuli. Participants were randomly assigned to one of the four training sets, and testing stimuli were generated with by-participant randomization.

All audio stimuli were generated by Amazon Polly text-to-speech synthesizer (*Amazon Polly*, 2023), with a British English accent in a female voice (voice ID: Amy). The quality of the audio was assessed by two phonetically trained graduate students. The output file was in MP3 format with a 24 Hz sampling rate.

2.3 *Procedure* The experiment was built with PsychoPy2 (Peirce et al., 2019) and ran online on Pavlovia (Bridges et al., 2020). After agreeing to participate in the experiment, participants first completed a Qualtrics survey (*Qualtrics*, 2023) to answer demographic questions and provide their linguistic background. At the end of the survey, they were redirected to the experiment page.

At the beginning of the training session, participants were informed that their task was to familiarize themselves with the names of unfamiliar objects in an alien language. Aligned with the concept of exposure withholding, the training session involved two parts. The first part contained 36 trials for singular stimuli without repetition. When a trial started, the audio of the stem form was played, and the image of a single object occurred simultaneously, lasting for two seconds (see Figure 1(i)). The second part contained 36 trials for singular-plural pairs with three repetitions. In between the two parts, participants were explicitly informed that they would proceed to learn how to call a group of objects in the language. The training for the singular form was displayed the same way as in the first part. After two seconds, the audio of the plural form was played, and three objects occurred in a triangular formation (see Figure 1(ii)) for two seconds. All training trials proceeded automatically, and a one-second eye-fixing page with a cross in the middle of the screen was inserted between every trial. Throughout the entire training session, no orthographical form was provided.

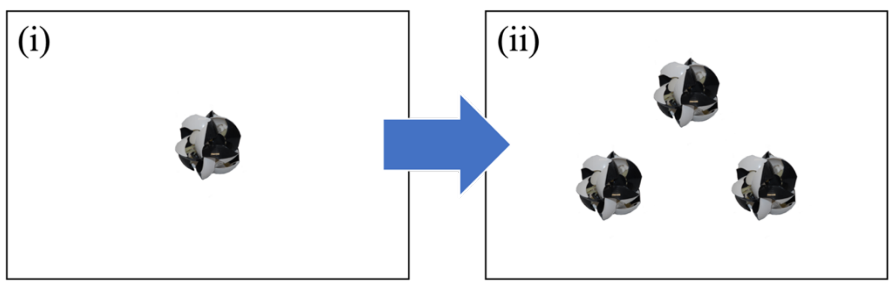


Figure 1: (i) shows the layout of a singular trial; from (i) to (ii) presents the procedure in a singular-plural trial.

Both the implicit and explicit testing modes consisted of 60 two-way forced-choice questions, and participants were randomly assigned to one of them. For every trial, after the audio of the singular form (or the singular-plural pair in the explicit mode) was played, two buttons appeared with the text "yes" or "no" on top of them, and participants were asked to press "left" or "right" key on the keyboard to indicate their judgment on the target items (Figure 2). Upon making a selection, the corresponding button turned blue, and they were able to move forward to the next trial. The procedure for selecting, replaying the audio, and advancing using the keyboard was demonstrated during an instructional session before the formal testing phase. Similar to the training phase, an eye-fixing page appeared between every trial.

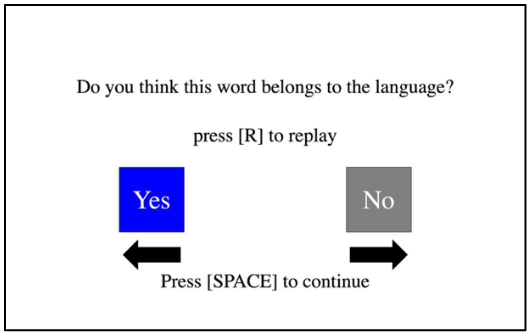


Figure 2: The layout of a testing trial in the implicit mode. As shown in the image, when the participants pressed the "left" key on the keyboard, the left button, indicating a "yes" answer, was highlighted in blue. In the explicit mode, the question was formed as "Is the second word correct?", where the second word referred to the suffixal form.

On top of the online experiment, we ran 12 extra offline sessions with the explicit testing mode in a sound-proof booth to check whether a clearer audio perception and theoretically increased participant attention in a lab environment would lead to any different learning performance. The offline participants included 6 students of different majors from the University of Hong Kong (HKU) and 6 linguistic students from Massachusetts Institute of Technology (MIT). See the Discussion section for more details about the rationale.

3 Results[4]

Participants' responses to the testing items (i.e., "Yes" or "No") were analyzed using mixed-effects logistic regression with the *lm4* package (Bates et al., 2018). Individual participants and testing items were included as the maximal random effect structure that converged (Barr et al., 2013). We checked whether the item type and the testing mode affected participants' judgments of stimuli items by fitting a regression model with the two as fixed effects using the *afex* package (Singmann et al., 2016). The package computed the *p*-value for each fixed effect using likelihood ratio tests. Since the testing mode did not yield a significant difference in responses ($\chi^2 = 0.06$, $p = 0.808$), we merged the data from implicit and explicit testing modes in the following report. Item type was a statistically significant predictor of responses ($\chi^2 = 266.23$, $p < 0.001$). In other words, whether participants were willing to accept an item as part of the training language largely depended on whether the consonant combination in the item exhibited (1) Voicing harmony, (2) Continuancy harmony, or (3) Voicing + Continuancy harmony. We assessed the acquisition of each harmony generalization in the following analysis.

The acceptance rate for each item type was calculated from the proportion of "Yes" responses and presented in Figure 3. The average acceptance rates of HarmonyOld (75.95%)[5], HarmonyNew (60.72%), and OnlyCont items (59.75%) were above chance, while the average acceptance rates of OnlyVoi (50.91%) and Disharmony items (48.67%) were only at chance level. We computed pairwise comparisons between different item types with the "mvt" adjustment method in the *emmeans* package (Lenth, 2025), and the statistical analysis was in line with the visual inspection. First, HarmonyOld items outperformed all other item types ($p < 0.001$). The fact that participants found exposure items more acceptable than novel items confirms an overall successful acquisition of the exposure language. Crucially, we were interested in participants' judgments of OnlyCont and OnlyVoi items, as they directly corresponded to participants' generalizations of the simple Continuancy harmony and the simple Voicing harmony. We found that the acceptance rate of OnlyCont items was significantly higher than that of Disharmony items ($p < 0.001$), while the acceptance

[4] All data and code plus audio stimuli are available at https://osf.io/y3f54/.

[5] Note that 19 out of the 200 participants from two testing modes incorrectly rejected HarmonyOld items at least half the time, which may indicate a failure in learning. We conducted separate analyses excluding the 19 participants, but the trend of all statistical tests remained the same. For the sake of keeping our analyses comparable with those of Durvasula & Liter (2020), all statistical tests reported in the main text included responses from the 19 participants.

rate of OnlyVoi items was not ($p = 0.716$). This provided evidence for the Continuancy generalization but suggested null evidence for the Voicing generalization. In addition, OnlyCont items were considered acceptable to a similar extent as the HarmonyNew items ($p = 0.992$), which agreed with both the two simple generalizations and the complex generalization. This again hints that the addition of Voicing harmony did not help with participants' generalizations.

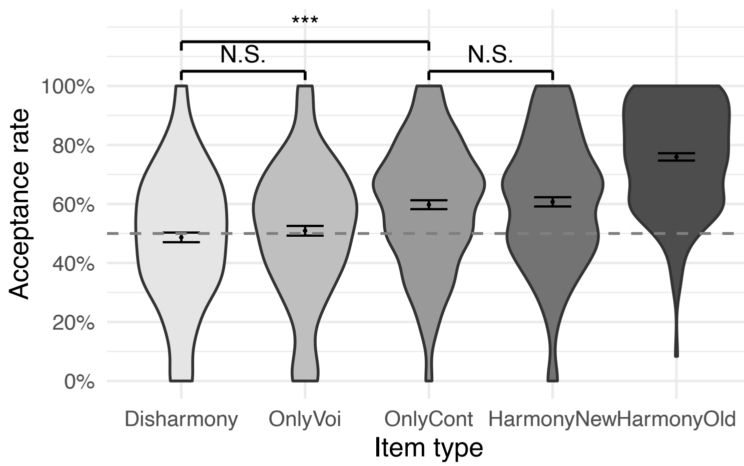


Figure 3: Acceptance rates of five item types. Error bars represent standard errors.

Two conceivable scenarios could have led to the results above. On the one hand, there could have been a preference for Continuancy harmony over Voicing harmony among all learners. On the other hand, individual learners could have preferred to generalize either Continuancy or Voicing harmony, with a larger number of learners leaning towards Continuancy harmony. If that were the case, those individuals would have been less willing to accept OnlyVoi items as they were more willing to accept OnlyCont items. To determine which possibility is true to our data, we examined individuals' judgments of OnlyVoi and OnlyCont items. As shown in Figure 4, there was a positive correlation between the acceptance rates of the two item types (Kendall's $\tau = 0.274$, $p < 0.001$), invalidating the second possibility. This suggests that, out of the two simple harmony generalizations, Continuancy harmony was successfully acquired while Voicing harmony was not.

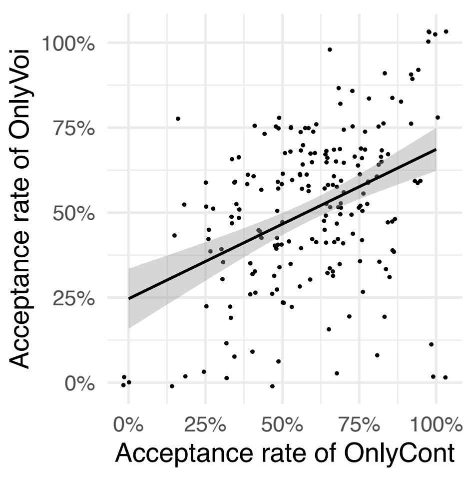


Figure 4: Correlation between OnlyCont items' acceptance rate and OnlyVoi items' acceptance rate. Individual data points are jittered to avoid overlapping and improve visibility.

Nevertheless, it is still possible that Voicing harmony played a role when combined with Continuancy harmony in the complex generalization. We further investigated the acquisition of the complex generalization by comparing a regression model that only included the two simple generalizations as two predictors (Table 1) with a regression model that additionally included the complex generalization as the interaction of the two predictors (Table 2)[6]. The acceptance of OnlyVoi items was determined by the Voicing harmony generalization and OnlyCont items by the Continuancy harmony generalization. HarmonyNew items were coded as an additive effect of the two simple generalizations in Model 1 while all three generalizations (two simple generalizations + one complex generalization) in Model 2. The results of both regression models showed that only Continuancy harmony had a significant effect on participants' responses but not Voicing harmony or the interaction term, aligning with the results of post hoc tests. According to the results of a Chi-squared test, the interaction term included in Model 2 did not significantly improve the model fit compared to Model 1 (χ^2 = 0.38, p = 0.539), indicating that the complex generalization did not contribute to the acceptance of a testing item.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.05	0.08	-0.64	0.524
Voicing harmony	0.07	0.06	1.18	0.238
Continuancy harmony	0.48	0.06	7.82	<0.001

Table 1: Regression model 1.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.07	0.08	-0.82	0.414
Voicing harmony	0.11	0.09	1.27	0.203
Continuancy harmony	0.52	0.09	5.97	<0.001
Voicing harmony:Continuancy harmony	-0.08	0.12	-0.62	0.539

Table 2: Regression model 2.

To confirm that the non-significant interaction term in Model 2 is representative of the learning outcomes without factoring in the complex generalization, we conducted a Monte Carlo simulation following the two probabilistic models in Durvasula & Liter (2020). Both models described possible ways for learners to employ the two simple harmony generalizations in the acceptability judgments without incorporating the complex generalization. The first model assumed that learners recruited both simple generalizations simultaneously for each testing item, calculating the rejection probability as the product of the probabilities of their associated generalizations. In the second model, acceptability judgments were based on a randomly selected one of the two simple generalizations, and the rejection probability was computed as the average of the probabilities linked to each generalization. Data from 1,000 simulated experiments using two probability models with various randomization parameters were analyzed. Results invariably exhibited no statistically significant difference between the simulated data without the complex generalization and the real experimental data, supporting the hypothesis that participants did not generalize with the complex harmony.

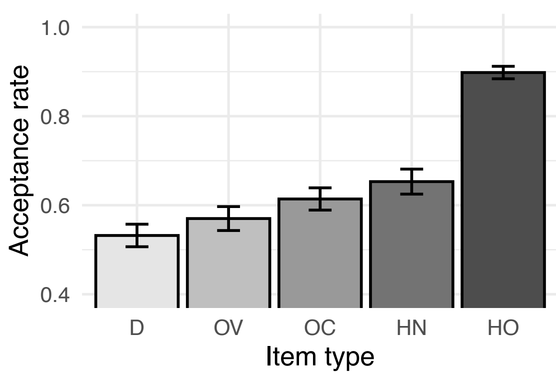
Probability Model	First Simulation	Second Simulation
Both generalizations evaluated together	0.461	0.475
One generalization evaluated at a time	0.327	0.336

Table 3: The proportion of interaction coefficients from the simulated data that was further away from the mean of the interaction coefficients than the interaction coefficient from the actual experimental data.

[6] As noted in the Method section, two stem consonants, {p^h, t^h}, were aspirated, and another two, {s, z}, were strident. We ran two additional regression models incorporating stem aspiration and stem stridency, respectively, into model 2 (i.e., including interactions of all three terms) and found no statistically significant main effect of the two (Aspiration: β = -0.98, z = -0.76, p = 0.445; Stridency: β = -0.12, z = -0.95, p = 0.344). Therefore, we assumed that there was no systematic influence from the additional features to our results.

4 Discussion

The current study tested learners' generalization preference in the acquisition of an alternation pattern that was simultaneously compatible with (1) simple Voicing harmony, (2) simple Continuancy harmony, and (3) complex Voicing + Continuancy harmony. We found strong evidence of the Continuancy harmony generalization while no indication of the Voicing harmony generalization or the complex harmony generalization. Our results demonstrate learners' inclination to focus on one out of multiple available generalizations and prioritize the simple generalization with fewer featural specifications over the complex generalization in learning alternation. This supports the single simplest generalization account which was laid out in the Introduction section. However, the current findings partly diverge from the phonotactic learning results in Durvasula & Liter (2020). When learners were exposed to a phonotactic pattern designed with the same three harmony generalizations in Durvasula & Liter (Experiment 2 and 3, 2020), their acceptance judgments of testing stimuli were affected by both Voicing harmony and Continuancy harmony as described in the best fitting logistic regression model (Experiment 2, Voicing: $\beta = 0.18$, $z = 1.94$, $p = 0.05$; Experiment 2, Continuancy: $\beta = 0.40$, $z = 4.27$, $p < 0.001$; Experiment 3, Voicing: $\beta = 0.31$, $z = 2.95$, $p < 0.01$; Experiment 3, Continuancy: $\beta = 0.37$, $z = 3.65$, $p < 0.001$). This signifies that learners were able to keep track of multiple simple generalizations regarding the phonotactic pattern. In contrast, only Continuancy harmony was a significant predictor of participants' responses to testing stimuli in the logistic regression model that most effectively captured the alternation learning results in our experiment (Table 1). Furthermore, the discrepancy between the two studies was revealed by pairwise comparisons among item types, as shown in Figure 5 (replicated from Durvasular & Liter, Experiment 2, 2020) and Figure 6 (rescaled from Figure 3). The difference in the acceptance rates of OnlyVoi items and Disharmony items reached statistical significance (Experiment 3) or at least marginal significance (Experiment 2) in the phonotactic learning experiments but not in the alternation learning experiment, although OnlyCont items were accepted significantly more than Disharmony items in both studies.



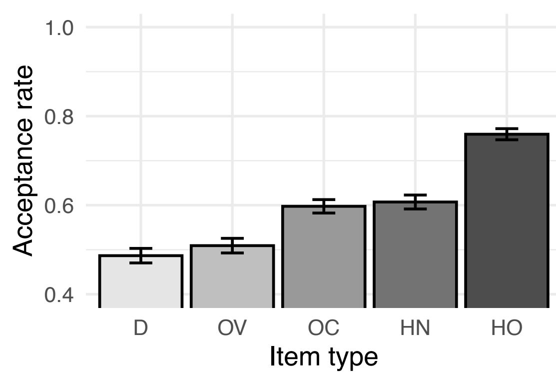


Figure 5 (left): Phonotactics learning results (Durvasula & Liter, Experiment 2, 2020). **Figure 6 (right)**: Alternation learning results (rescaled). Error bars represent standard errors. Item type from left to right: Disharmony (D), OnlyVoi (OV), OnlyCont (OC), HarmonyNew (HN), HarmonyOld (HO).

Another divergence in the collected data, immediately available on the bar plots, is the poorer performance of HarmonyOld items in our study than in Durvasula & Liter (2020). This is not unexpected considering that their study was conducted offline, while ours was online, which inevitably left us with less control over the audio quality, environmental noise, and participants' attentiveness more generally. In order to make sure that the single simplest generalization in alternation learning was not simply a consequence of the less-controlled-for factors in online experimentation, we conducted an offline sanity check with students from HKU and MIT (see the Method section for more details). As presented in Figure 7, there was no systematic preference for OnlyVoi items over Disharmony items. Notably, there was a clear preference for OnlyCont items in the offline alternation learning results, replicating the online results. Therefore, we expect a discrepancy in phonotactic and alternation learning even in the absence of the potential influence from the experimentation mode.

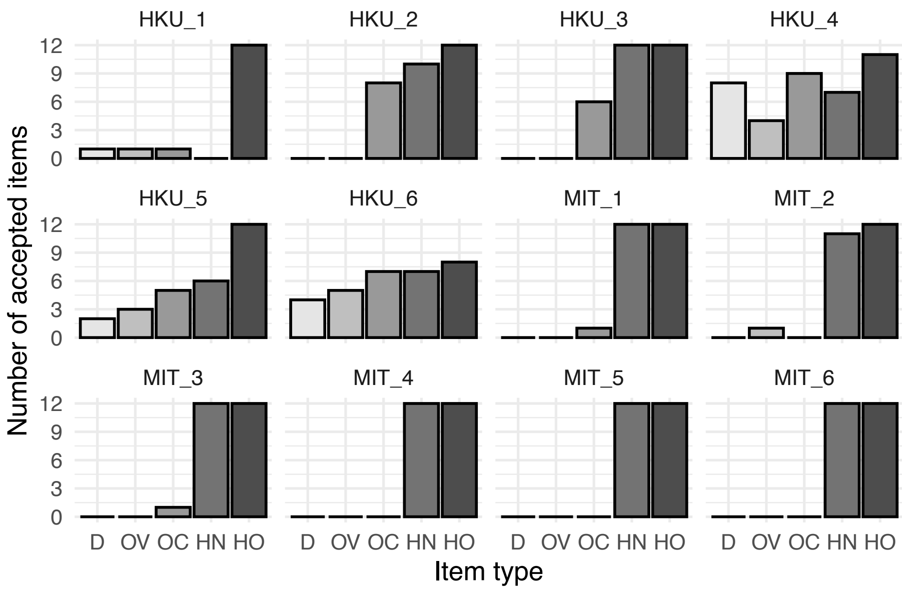


Figure 7: Offline alternation learning results. Item type from left to right: Disharmony (D), OnlyVoi (OV), OnlyCont (OC), HarmonyNew (HN), HarmonyOld (HO).

This discrepancy could be attributed to a fundamental difference in the two learning domains. While phonotactics is static regularities pre-existing in the lexicon, alternation arises from an active process that changes forms across morpheme boundaries. As discussed in the Introduction section, phonotactics and alternation are encoded separately as Morpheme Structure Constraints and re-write rules in rule-based phonological theories (Chomsky & Halle, 1968). Granted that the distinct formalizations of the two lead to the 'duplication problem' in the grammatical description of a phonological generalization (Kenstowicz & Kisseberth, 2014), they provide an accurate account of how parallel phonotactic and alternation patterns might have developed independently from one source and may evolve into mismatches as a result of more sound changes (Paster, 2013). Moreover, experimental investigations suggest that phonotactic learning does not necessarily facilitate alternation learning, particularly when the training language lacks overt evidence of the alternation pattern (Chong, 2021) or when the participants' first language does not provide phonotactic support for alternation (Do & Yeung, 2021). This alludes to separate learning mechanisms behind phonotactics and alternation. On top of that, alternation but not phonotactic patterns are subject to a paradigm uniformity bias (Hayes, 2004; McCarthy, 1998), which favors correspondences between output forms in a morphological paradigm and, in a sense, further complicates alternation learning compared to phonotactic learning. Our study and its comparison against Durvasula & Liter (2020) contribute to the discourse by revealing that the number of generalizations, i.e., single or multiple, recruited for learning ambiguous data varies with the domain of learning, i.e., alternation or phonotactics.

In spite of differences in the learning domains, participants in both studies generalized Continuancy harmony more readily than Voicing harmony, as shown in Figure 5 and Figure 6. Parallel patterns with different features typically elicit different results in ALL experiments (Durvasula & Liter, 2020). In the examination of constraint cumulativity in phonotactics, learners did not tolerate the violation of single consonant harmony (nasal harmony or sibilant harmony) more than the violation of both consonant and vowel harmony, while they did consider the violation of single vowel harmony (backness harmony) more acceptable (Breiss, 2020). When exposed to a potentially saltatory alternation pattern that changed the voiceless stops to the voiced fricatives, learners displayed more willingness to extend the alternation to intermediate voiced stops than intermediate voiceless fricatives (White, 2014). This is because featural distance, i.e., the number of participating features, is not the only metric to describe similarity from the original set to the extension set in a phonological generalization. Cristia et al. (2013) reported that the generalization of sound patterns could also be predicted by language-specific phonetic distance (Mielke, 2012). With regard to alternation, Albright and Do (2017) and White (2014) found that the learning data they collected was regulated by the perceptual similarity between input-output pairs (Wang & Bilger, 1973). Specifically, alternations between perceptually more similar sounds were preferred per the P-map hypothesis (Steriade, 2001). In contrast, the acquisition of spirantization patterns in Finley (2022) was better modeled by phonological similarity (Frisch, 1996) as opposed to perceptual similarity (Wang & Bilger, 1973).

However, calculating input-output similarity is not feasible in our study because we (1) did not specify the underlying form of the suffix in the training language and (2) could not determine the underlying form of the suffix in participants' mental representation after the limited experimental exposure of the artificial language. Moreover, the input-output similarity does not explain the comparable learning results in the phonotactic experiments (Durvasula & Liter, 2020). Under these considerations, we computed the similarity between pairs of output consonants (Table 5 and Table 6) in our study and in Durvasula & Liter (2020), following the measures of perceptual similarity and phonological similarity in Finley (2022)[7]. We found that Continuancy pairs were not perceptually more similar than Voicing pairs in our study (Voicing pairs = 0.037, Continuancy pairs = 0.016) or in Durvasula & Liter (2020) (Voicing pairs = 0.058, Continuancy pairs = 0.027). On the other hand, the phonological similarity hypothesis was supported by the alternation learning data (Voicing pairs = 0.141, Continuancy pairs = 0.175) and the phonotactic learning data (Voicing pairs = 0.201, Continuancy pairs = 0.210). There are of course alternative similarity-related hypotheses, such as the phonetic similarity (Mielke, 2012). We simply want to demonstrate that some kind of similarity calculation, in addition to feature counting, can be helpful to the report on phonological learnability.

	pʰ	b	tʰ	d	f	v	s	z
kʰ		0.029		0.018	0.069		0.033	
g	0.020		0.019			0.052		0.041
ʃ	0.017		0.040			0.010		0.018
ʒ		0.014		0.027	0.008		0.009	

Table 5: Perceptual similarity between pairs of consonants in our study. Light grey cells are Voicing pairs and dark grey cells are Continuancy pairs.

	pʰ	b	tʰ	d	f	v	s	z
kʰ		0.19		0.16	0.14		0.11	
g	0.21		0.17			0.16		0.13
ʃ	0.10		0.18			0.10		0.24
ʒ		0.11		0.20	0.09		0.24	

Table 6: Phonological similarity between pairs of consonants in our study. Light grey cells are Voicing pairs and dark grey cells are Continuancy pairs.

Finally, both our study and Durvasula & Liter (2020) found that learners prefer to employ the smallest number of features required, i.e., one as opposed to two, in phonological generalizations. The results were incompatible with the Subset Principle (Dell, 1981) but conformed to the SPE evaluation metric (Chomsky & Halle, 1968). In their seminal review work, Moreton and Pater (2012) noted that experimental research using the ALL paradigm has consistently reported better acquisition of a phonological pattern involving fewer features over a pattern involving more. For example, exposure to one pattern that targets {p, t, k} and {b, d, g} gave rise to better performance in testing compared to exposure to another pattern that targets {p, d, k} and {b, t, g} (Saffran & Thiessen, 2003). Moreton and Pater (2012) termed this preference for simpler patterns over complex ones "the structural bias." Diverging from the previous research, learners in our study and Durvasula & Liter (2020) were trained on one pattern that could be generalized with fewer features or more and provided sufficient positive evidence to acquire the multiple relevant features in the pattern. Nevertheless, they still exhibited a preference for the simpler generalization of the pattern over the complex generalization. The experimental evidence together points to an overarching simplicity bias in phonological learning, which should be accounted for in the formal description of phonological grammar.

[7] Perceptual similarity was derived from the mutual confusability score of pairs of sounds in the perceptual identification tasks with VC and CV syllables (Wang & Bilger, 1973). Phonological similarity assumed that the English participants recruited their native segment inventory knowledge and measured natural class structure and featural redundancy in English (Frisch, 1996).

References

Albright, Adam. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41. https://doi.org/10.1017/S0952675709001705

Albright, Adam, & Do, Youngah. (2017). A substantive bias for perceptually minimal alternations in Artificial Grammar learning. *Old World Conference on Phonology (OCP), 2017*. https://hub.hku.hk/bitstream/10722/274779/1/Abstract.pdf?accept=1

Amazon Polly. (n.d.). [Computer software]. https://docs.aws.amazon.com/polly/index.html

Aslin, Richard N., & Newport, Elissa L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Baker, Carl L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581.

Barr, Dale J., Levy, Roger, Scheepers, Christoph, & Tily, Harry J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, Douglas, Kliegl, Reinhold, Vasishth, Shravan, & Baayen, Harald. (2018). *Parsimonious Mixed Models* (arXiv:1506.04967). arXiv. http://arxiv.org/abs/1506.04967

Braine, Martin DS. (1971). On two types of models of the internalization of grammars. *The Ontogenesis of Grammar*, *1971*, 153–186.

Breiss, Canaan. (2020). Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies. *Phonology*, *37*(4), 551–576. https://doi.org/10.1017/S0952675720000275

Bridges, David, Pitiot, Alain, MacAskill, Michael R., & Peirce, Jonathan W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414.

Chambers, Kyle E., Onishi, Kristine H., & Fisher, Cynthia. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chomsky, Noam & Halle, Morris. (1968). The sound pattern of English. *Haper and Row*.

Chong, Adam J. (2021). The effect of phonotactics on alternation learning. *Language*, *97*(2), 213–244. https://doi.org/10.1353/lan.2021.0017

Cristia, Alejandrina, Mielke, Jeff, Daland, Robert, & Peperkamp, Sharon. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2). https://doi.org/10.1515/lp-2013-0010

Dell, François. (1981). On the learnability of optional phonological rules. *Linguistic Inquiry*, *12*(1), 31–37.

Do, Youngah, & Yeung, Ping Hei. (2021). Evidence against a link between learning phonotactics and learning phonological alternations. *Linguistics Vanguard*, *7*(1), 20200127. https://doi.org/10.1515/lingvan-2020-0127

Durvasula, Karthik, & Liter, Adam. (2020). There is a simplicity bias when generalising from ambiguous data. *Phonology*, *37*(2), 177–213.

Finley, Sara. (2021). Learning Exceptions in Phonological Alternations. *Language and Speech*, *64*(4), 991–1017. https://doi.org/10.1177/0023830920978679

Finley, Sara. (2022). Generalization to Novel Consonants: Place Versus Voice. *Journal of Psycholinguistic Research*, *51*(6), 1283–1309. https://doi.org/10.1007/s10936-022-09897-1

Finley, Sara, & Badecker, William. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*(3), 423–437.

Frisch, Stefan. (1996). *Similarity and frequency in phonology* [Northwestern University].

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

Greenwood, Anna. (2016). *An experimental investigation of phonetic naturalness*. University of California, Santa Cruz.

Hale, Mark, & Reiss, Charles. (2008). *The phonological enterprise*. OUP Oxford. https://books.google.com/books?hl=en&lr=&id=ThNREAAAQBAJ&oi=fnd&pg=PR5&dq=hale+reiss+2008&ots=wz1sTilZOR&sig=xw2JWyxjuHLWTPvUQbqP_Ox4GIY

Horst, Jessica S., & Hout, Michael C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Kenstowicz, Michael, & Kisseberth, Charles. (2014). *Topics in phonological theory*. Elsevier.

Lenth, Russell V. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://rvlenth.github.io/emmeans/

Lysvik, Julian Kirkeby. (2020). *Where does naturalness in phonology come from? - Insights from artificial language learning* [PhD Thesis]. University of Oslo.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77–80. https://doi.org/10.1126/science.283.5398.77

Mielke, Jeff. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006

Moreton, Elliott. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition*, *84*(1), 55–71.

Moreton, Elliott, & Pater, Joe. (2012). Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass*, *6*(11), 686–701. https://doi.org/10.1002/lnc3.363

Paster, Mary. (2013). Rethinking the 'duplication problem.' *Lingua*, *126*, 78–91.

Pater, Joe, & Tessier, Anne-Michelle. (2005). Phonotactics and alternations: Testing the connection with artificial language learning. *University of Massachusetts Occasional Papers in Linguisitcs [UMOP]*, *31*, 1–16.

Peirce, Jonathan, Gray, Jeremy R., Simpson, Sol, MacAskill, Michael, Höchenberger, Richard, Sogo, Hiroyuki, Kastman, Erik, & Lindeløv, Jonas Kristoffer. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Pizzo, Presley, & Pater, Joe. (2016). Does Learning Alternations Affect Phonotactic Judgments? *Proceedings of the Annual Meetings on Phonology*, *3*. https://doi.org/10.3765/amp.v3i0.3694

Prolific (Version 2023-09). (2014). [Computer software]. Prolific. https://www.prolific.co

Qualtrics. (2023). [Computer software]. Qualtrics. https://www.qualtrics.com

Rasin, Ezer, Berger, Iddo, Lan, Nur, Shefi, Itamar, & Katzir, Roni. (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of Language Modelling*, *9*(1), Article 1. https://doi.org/10.15398/jlm.v9i1.266

Reeder, Patricia A., Newport, Elissa L., & Aslin, Richard N. (2009). *The role of distributional information in linguistic category formation*. https://philpapers.org/rec/REETRO-9

Reeder, Patricia, Newport, Elissa, & Aslin, Richard. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32). https://escholarship.org/content/qt23m494t4/qt23m494t4.pdf

Rissanen, Jorma. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Saffran, Jenny R., & Thiessen, Erik D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494. https://doi.org/10.1037/0012-1649.39.3.484

Singmann, Henrik, Bolker, Ben, Westfall, Jake, Aust, Frederik, & Ben-Shachar, Mattan S. (2016). afex: Analysis of Factorial Experiments. R package version 0.16-1. *R Package Version 0.16*, *1*.

Smolensky, Paul, & Prince, Alan S. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. https://doi.org/10.7282/T34M92MV

Steriade, Donca. (2001). *The phonology of perceptibility effects: The P-map and its consequences for constraint organization*. https://direct.mit.edu/books/edited-volume/chapter-pdf/2282897/9780262274890_cag.pdf

Wang, Marilyn D., & Bilger, Robert C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248–1266.

White, James. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition*, *130*(1), 96–115. https://doi.org/10.1016/j.cognition.2013.09.008

Zheng, Shuang, & Do, Youngah. (2025). Substantive Bias in Artificial Phonology Learning. *Language and Linguistics Compass*, *19*(1), e70005. https://doi.org/10.1111/lnc3.70005