2023. *Proc Ling Soc Amer* 8(1). 5530. https://doi.org/10.3765/plsa.v8i1.5530.

Modeling harmony biases in learning exceptions to vowel harmony

Sara Finley[*]

Abstract. Artificial language learning experiments typically show non-categorical results after training on categorical data. This is generally due to incomplete learning, but these results can also reveal biases. One example is that participants trained on a vowel harmony language with alternating and non-alternating affixes prefer the non-alternating affix in harmonic contexts (Finley 2021). In this paper, I show that (i) the preference for harmonic items in non-alternating affixes replicates for remote (online) data collection, and (ii) that this effect can be modeled with MaxEnt Harmonic Grammar. In Harmonic Grammar, the harmony score of each candidate determines its grammaticality, and the probability of surfacing. Because non-alternating affixes that satisfy vowel harmony have higher harmony scores than non-alternating affixes that violate harmony, harmonic candidates will be more likely to surface than disharmonic candidates, even when both types of items surface at levels greater than expected by chance. The theoretical and methodological implications for these results are discussed.

Keywords. phonology; vowel harmony; artificial language learning; exceptions; replication; MaxEnt learning

1. Introduction. Artificial language learning experiments have been a popular and fruitful method for probing the nature of linguistic representations, and how they might be learned (Culbertson 2012; Moreton & Pater 2012). In these experiments, participants (often adults) are presented with words from a novel, made-up language that exemplifies some property found in natural languages. Following exposure, participants are tested on their learning and generalization of the language. In most cases, even when participants are presented with a categorical pattern (e.g., a pattern without variation or exception), results are rarely, if ever, categorical. This is likely because the relatively short exposure phase (a few minutes) is enough time to learn the pattern generally, but not enough time to respond categorically for all trials. In addition, individual differences in effort, attention, and ability yield variability in the data (Ettlinger et al. 2016).

While this variability may seem problematic, it is possible that this variation can lead to important insights into learners' inferences about linguistic representations, and their biases. For example, Finley (2021) trained adult, English speakers on a categorical vowel harmony language with one affix that alternated in response to harmony, and another non-alternating, 'exceptional' affix. In this artificial language, CVCV stem vowels were harmonic for backness and rounding, and affixes either alternated with respect to vowel harmony (e.g., [me] for stems with front/unround vowels, and [mo] for stems with back/round vowels), or did not alternate (e.g., was always [go]). Participants correctly applied harmony to the affixes at rates significantly greater than chance (compared to 50% chance, and a no-training control condition), in a two-alternative forced-choice task comparing an affixed form ending in [e] vs. an affixed form end-

[*] The author would like to thank Tiarra Lewis, Anqi (Stella) Wang, Saara Charania, and Barbara Millward for their assistance in developing this project, as well as Cassandra Hunt for assistance in proofreading this manuscript and the LSA poster. I would also like to thank the audience at the LSA 2023 Annual Meeting for their participation and questions. Partial funding for this project was received from a PLU Benson-Starkovic grant to Sara Finley. Sara Finley, Pacific Lutheran University (finleysr@plu.edu).

ing in [o]. Despite having learned the rule generally, participants were more likely to select the correct non-alternating affix [go] when the affix vacuously satisfied harmony (e.g., when the stem vowels were also back/round).

Finley (2021) suggests that this harmony bias can be modeled with Harmonic Grammar (Legendre, Miyata & Smolensky 1990), where candidates are given a harmony score based on constraint violations. The higher the harmony score, the more likely the candidate will emerge. In the case of vowel harmony with exceptional non-alternating affixes (Finley 2010), harmony can be modeled with a high-ranked morpheme-specific constraint, a mid-ranked Agree constraint (Baković 2000), and a lower-ranked general faithfulness constraint (note that F is used as a placeholder for Back/Round features). Table 1 shows a basic Harmonic Grammar applied to the stimuli in Finley (2021). Note that violations of constraints are negative, meaning that the candidates that violate the most, highest ranked constraints will have the lowest harmony, and will not be selected. For each input, there are two possible outputs, based on the two alternative, forced-choice test in Finley (2021), between [e] and [o].

Violations of ID[F]-go occur when affix /go/ changes to [ge]. Because the morpheme-specific constraint is high ranked (a weight of 100), penalties are large. This allows for the possibility of disharmony when the non-alternating affix is present. Violations of the general ID[F] constraint occur for both the alternating and non-alternating affixes. Because this constraint is lower ranked (weight of 10), harmony applies even if the underlying form changes. In this example /e/ was arbitrarily selected as the underlying form for /me/, but the general results still hold for either /e/ or /o/. The issue of the underlying form for the alternating affix is discussed in more detail further in this paper.

	Candidates	ID[F]-go	Agree[F]	ID[F]	Total
		-100	-50	-10	Harmony Score
Front Vowel Stem	/beme+me/				
Alternating Affix	☞ [bememe]				0
	[bememo]		-50	-10	-60
Back Vowel Stem	/bomo+me/				
Alternating Affix	bomome		-50		-50
	☞ bomomo			-10	-10
Front Vowel Stem	/beme+go/				
Non-Alternating Affix	bemege	-100		-10	-110
	☞ bemego		-50		-50
Back Vowel Stem	/bomo+go/				
Non-Alternating Affix	bomoge	-100	-50	-10	-160
	☞ bomogo				0

Table 1. Harmonic Grammar analysis of training data

The disharmonic, non-alternating form [bemego] has the lowest harmony score of all the winners (-50). The candidates that are both faithful and harmonic have scores of 0 (the highest possible), and the unfaithful winning candidate /bomo-me/ → [bomomo] has a harmony score close to 0. In this respect, the non-alternating form in a disharmonic context is the 'worst' possible winner. This could potentially explain why participants were less likely to select the correct item when the non-alternating affix was in a disharmonic context (even if at a rate greater than chance). However, this approach cannot work if participant selects the winning, highest harmony candidate every time. What is needed is a probabilistic approach to selecting candidates that can also incorporate learning.

Such an approach is possible with MaxEnt Harmonic Grammar (Hayes & Wilson 2008; Goldwater & Johnson 2003), where constraint weights are learned from the training data, and candidates are selected probabilistically based on the learned weights. The goal of the present paper is to extend the analysis of the harmony bias found in Finley (2021) using MaxEnt Harmonic Grammar. Because such an extension relies on trusting that the data from Finley (2021) is robust enough to withstand replication under multiple conditions, the paper also includes a replication of the original effect. Replication research has become increasingly important in psychology and cognitive science (Ebersole et al. 2016; Open Science Collaboration 2015), including experimental linguistics (Roettger & Baer-Henney 2019). Knowing that a given experimental finding replicates across different populations, modalities and stimuli is important for confirming the robustness of a reported effect. Having two sets of data to compare to the simulated learning results can also be useful in understanding and generalizing such results to human cognition.

This paper is organized as follows. Sections 2 and 3 present the Method and Results of the replication study. Section 4 presents the MaxEnt Harmonic Grammar simulations. Sections 5 and 6 provide discussion and conclusions.

2. Experimental method. The purpose of the experiment presented in this paper was to replicate the harmony bias shown in Finley (2021). Participants who were trained on a vowel harmony language where one affix alternates in accordance with vowel harmony, but another affix does not, learned the behavior of both types of affixes, but preferred the non-alternating affixes in harmonic contexts.

2.1. PARTICIPANTS. Forty-one participants were recruited to participate in this study, but 39 participants were included for analysis. One participant was dropped because they reported that their native language was Samoan, which is reported to show some vowel harmony patterns (Alderete & Finley 2016). All other participants reported native-level competence in American English, and no experience with a vowel harmony language, natural or artificial. The most common second language reported was Spanish. A second participant was excluded because they indicated in a post-completion survey that they did not wish their data to be included in the final analysis.

Twelve participants indicated that they were female; all other participants indicated that they were male. Seven participants were recruited from the Pacific Lutheran Psychology Familiarization (PsycRes) participant pool and were compensated with course credit. All other participants were recruited through Amazon's Mechanical Turk (location set to USA) and were paid $4 for their participation. Any participant who took longer than 45 minutes to complete the study were excluded from analysis, but all participants completed the task in under 40 minutes.

2.2. DESIGN AND STIMULI. In general, the design and stimuli were the same as the design from Finley (2021); participants were trained on a back/round vowel harmony pattern with one affix that alternated between [me] and [mo] in accordance with vowel harmony, and another affix that was always [go] regardless of the stem vowels. Stems containing the vowels [i] and [e] always ended in [me], while stems containing the vowels [o] and [u] ended in [mo]. Following training, a two-alternative forced choice test (described in more detail below) assessed participants' learning; participants chose between two words that were identical except the ending (either [e] or [o]) that either obeyed harmony or disobeyed harmony.

All training and test items were presented auditorily without any orthographic representation, and participants were asked to wear headphones. The stimuli were recorded in a sound-attenuated booth by an adult female American English speaker (different from the speaker used in Finley 2021). All sound editing was conducted in Praat (Boersma & Weenink 2017). Sound files were normalized to 70Hz, but participants could adjust the volume of the sounds as needed.

Participants were trained on the harmony pattern via 24 sets of triads: stem, stem+me/mo, and stem+go. In Finley (2021), the order of the affixed form was counterbalanced across participants, where half of participants always heard the alternating affix first, and the other half heard the non-alternating affix first. However, because the order of presentation did not appear to affect results, participants were presented with the same order (alternating affix first). The consonants were chosen from the set [p, t, k, b, d, g, m, n], and the vowels were from the set [o, u, e, i]. Half of the 24 items contained front vowels in the stems, while the other half contained back vowels. All harmonic combinations of vowels were included, but no stems were disharmonic; disharmony only appeared when front vowel stems were combined with the non-alternating affix.

The set of triads was presented five times in a different randomized order for each cycle. This number was reduced from Finley (2021) (where items were repeated eight times) to better accommodate online data collection. Examples of training items are shown in Table 2. Full lists of stimuli, analysis code, anonymized data, and stimuli files can be found at: https://osf.io/jda32/.

	Stem	Stem+me	Stem+go
Front Vowel Stem	degi	degime	degigo
	kete	keteme	ketego
	tipe	tipeme	tipego
	niki	nikime	nikigo
Back Vowel Stem	bono	bonomo	bonogo
	doku	dokumo	dokugo
	tunu	tunumo	tunugo
	kupo	kupomo	kupogo

Table 2. Examples of training items

There were 50 items in the two alternative forced choice task, 10 each from Old, New_me, New_go, OldAgglut, and NewAgglut. Within these, New_me and New_go items were classified in terms of their stem vowels (front vs. back vowel), creating five of each of these items. Old items came from the training set, New items were different from the training set, but contained the same vowels and consonants as the items in the training set, meaning they were very similar to those items. Agglut (short for Agglutinative) items were items that contained both the non-alternating ([go]) affix followed by the alternating ([me/mo]) affix. Old_go, and Agglut items always had front vowel (disharmonic) stems. Table 3 shows examples of test items used in the experiment. Note that the ‘correct’ item could be disharmonic in the case of the non-alternating affix.

Item Type	Harmonic	Disharmonic
Old_go	degige	degigo
	ketege	ketego
Old_me	kupomo	kupome
	degime	degimo
New_meF	dimime	dimimo
	kipeme	kipemo
New_meB	nodomo	nodome
	gutomo	gutome
New_goF	pemime	pemimo
	nepeme	nepemo
New_goB	gutogo	gutoge
	modugo	moduge
Old_Agglut	bemigomo	bemigome
	kinegomo	kinegome
New_Agglut	kipegomo	kipegome
	minegomo	minegome

Table 3. Examples of test items

2.3. PROCEDURE. The online study was run using the FindingFive experiment design platform (FindingFive Team 2019). The experiment started with a sound-check stimuli file ([udvu]) in a different voice than the experimental stimuli. Participants were asked to adjust the volume on the computer and headphones as needed to properly hear the sound files. After participants completed a basic demographic survey, they were given instructions for the training phase of the study. Participants were told to listen to each sound file once, and then click the 'Continue' button in the browser to hear the next set of words.

After training was complete, participants were advised to take a short break, and then return to take the test when they were ready. The instructions in the test phase informed participants that on each trial, they would hear two words and their job was to select the word that most likely belonged to the language in the study. Participants could choose to click on two buttons 'First' or 'Second' to indicate whether they believed the first item, or the second item was most likely to be correct. Upon completion of the test, participants were given a written debriefing, and a short completion survey that asked for feedback. Participants were also given the opportunity to opt out of having their data included in the analysis. A status bar at the top of the browser indicated participants' progress in the study.

3. Data analysis. To simplify the statistical analysis, the Old and the Agglut items were not included in the inferential statistical models, but the means and standard deviations are included in Table 4, below.

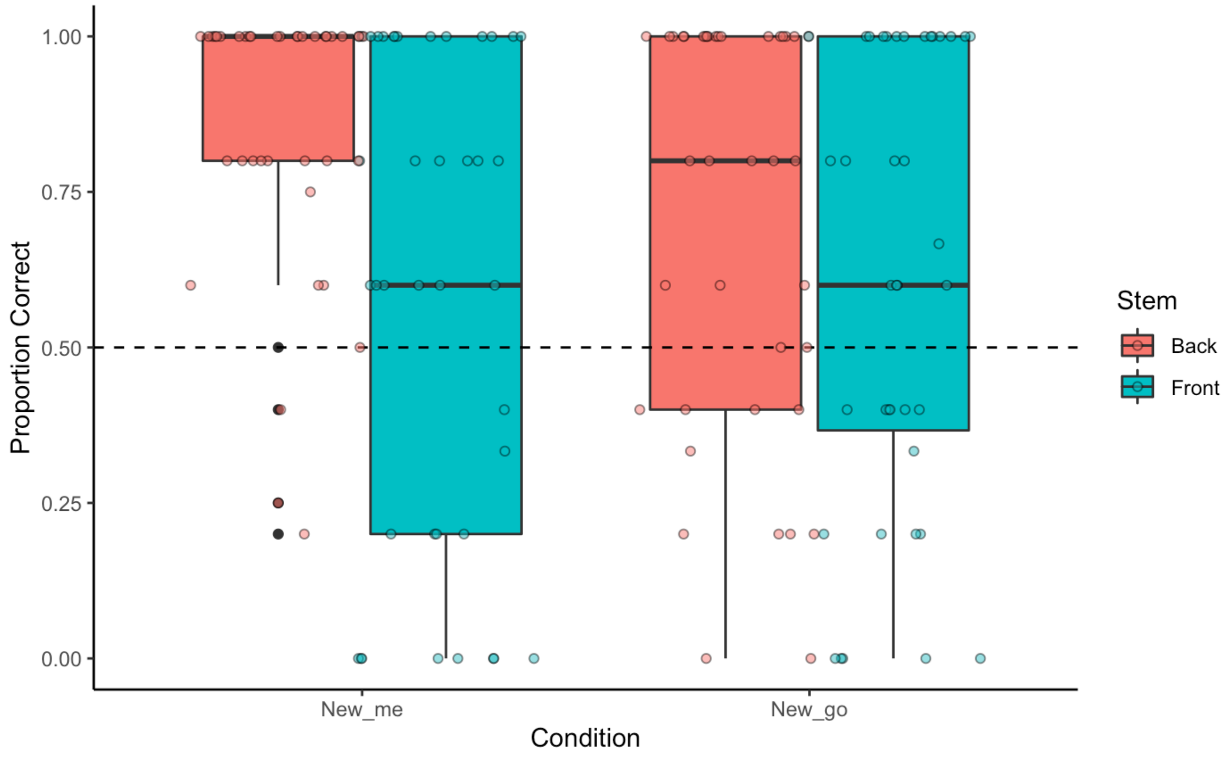

Figure 1. Boxplots for New_me and New_go items by stem vowel

Participant means and aggregated results for New_Me and New_go items are displayed alongside the box plot in Figure 1. Data from all eligible participants was included, except for items with a response time longer than 10s. This occurred for 42 items across thirteen participants.

Item Type	Mean	Standard Deviation
Old_go	0.65	0.48
Old_me	0.69	0.46
New_meF	0.62	0.49
New_meB	0.72	0.45
New_goF	0.59	0.49
New_goB	0.86	0.35
Old_Agglut	0.49	0.50
New_Agglut	0.45	0.50

Table 4. Overall means and standard deviation by test item type

The data were analyzed in R (R Core Team 2019) and R-Studio (RStudio Team 2020) using binomial mixed effects logistic regression models in the lme4 package (Bates et al. 2015). The emmeans (Lenth et al. 2019) package was used to conduct pairwise comparisons between different test items, using Tukey corrections for multiple comparisons. The model was dummy coded, with New_goF items set as the baseline, to ensure that even if participants showed a harmony bias for the non-alternating affix, that they still selected the non-alternating affix at a rate greater than chance (as indicated by a significant intercept). The model included random intercepts for subject and items, as well as random slopes for item type by individual item. The results of the model showed a significant intercept (β =0.45, *SE* = 0.22, *z* = 2.05, *p* = 0.040), indicating that participants were able to correctly select the non-alternating affix in disharmonic contexts. The theoretically relevant contrasts (shown in Table 5) showed significant differences between New-goF and New_goB items, replicating the harmony bias for the non-alternating affix. There was also a numerical difference between New_meF and New_meB items, but this was not significant after corrections for multiple comparisons. This is essentially the same pattern of results found in Finley (2021).

Contrast	Estimate	SE	z ratio	p value
New_goF – New_goB	-1.63	0.27	-6.01	<0.0001
New_goF – New_meF	-0.10	0.23	-0.45	0.97
New_meF –New_meB	-0.54	0.24	-2.29	0.10
New_goB –New_meB	0.99	0.28	3.58	0.0020

Table 5. Summary of contrast comparisons

To rule out the possibility that participants did not learn the harmony pattern, but simply selected the back vowel for all trials, a second model was run with New_meF as the baseline. This model, with the same random effects structure reported above, yielded a significant intercept (β =0.55, *SE* = 0.22, *z* = 2.49, *p* = 0.013), indicating that participants did learn the harmony pattern. While the pattern of results of the replication generally replicates those of Finley (2021), it is interesting to note that the rate of correct responses to New_goF items was relatively low (0.59) compared to Finley (2021) (0.71). This could be due to the online data collection procedure, which had a shorter exposure phase, and no way to monitor participant attention.

4. MaxEnt model. The harmony bias shown in Finley (2021) and the replication provided in this paper can be explained in terms of Harmonic Grammar. As discussed above, the non-alternating affix should surface as [go] regardless of the stem vowels because faithfulness to the non-alternating affix is ranked higher than constraints governing vowel harmony. The faithful form of the non-alternating affix will receive higher harmony than its unfaithful form, even if it triggers disharmony. However, the harmony score of the non-alternating affix should be higher in a harmonic context than a disharmonic context, even if both are 'winners'. One issue with this explanation is that the grammar selects a single output, and generally does not compare grammaticality of different winning candidates. This means that if the grammar always selects the candidate with the highest harmony, the non-alternating affix should be selected categorically, since the training data are also categorical. However, learning models using MaxEnt Harmonic Grammar (Hayes & Wilson 2008; Goldwater & Johnson 2003) can show non-categorical behavior from categorical input, particularly if the learning weight is set low. Previous artificial language learning studies in phonology (Strütjen et al. 2018; White 2017; Wilson 2006; Finley 2022) have been simulated with MaxEnt Harmonic Grammar. While the specific elements of how the MaxEnt Harmonic Grammar learning algorithm works are beyond the scope of this paper, readers are invited to see Hayes and Wilson (2008) for a review, as well as White (2017) for more detailed descriptions of applications of the MaxEnt Grammar Tool (Hayes, Wilson & George 2009) applied to artificial language learning.

I simulated Finley (2021) and its replication using the MaxEnt Grammar Tool (Hayes, Wilson & George 2009). I created a table with the different types of training and test items, and the constraints needed to generate the grammar. The candidates were always the two alternatives for the forced choice task (words ending in [e] vs. [o]). Because the violation profiles were identical for items with the same vowel structure (e.g., *pikeme* and *gedime* both satisfy the same set of constraints in this model), I only listed each type of item once, but increased the n for each to reflect the number of types of items (rather than list each individual training and test item). I assumed that the stem vowels were faithful to the input (e.g., did not alternate to create harmony), and that the underlying form of the non-alternating affix was /-go/. However, because it was not clear whether the form for the alternating affix [me] should be unspecified (e.g., [mE], where no faithfulness constraints are violated for the alternating affix), a back vowel (/mo/, where the faithfulness constraint is violated in front vowel contexts) to match the non-alternating affix, or a front vowel (/me/, where the faithfulness constraint is violated in back vowel contexts), I ran the simulation three times, each with a different underlying form. A sample table assuming /me/ as the underlying form is given in Table 6. Note that the order of the constraints in the table does not matter, because the algorithm learns the weights based on the violation profile (violations are indicated with a *).

Description	Candidates	n	ID[F]	Agree[F]	ID[F]-go
Front Vowel Stem	/beme+me/				
Alternating Affix	bememe	60			
	bememo		*	*	
Back Vowel Stem	/bomo+me/				
Alternating Affix	bomome			*	
	bomomo	60	*		
Front Vowel Stem	/beme+go/				
Non-Alternating Affix	bemege		*		*
	bemego	60			
Back Vowel Stem	/bomo+go/				
Non-Alternating Affix	bomoge		*	*	*
	bomogo	60			

Table 6. Examples of training data inputted into the MaxEnt Learner

The MaxEnt Grammar Tool also requires the user to specify parameters for μ (a proxy for prior constraint weight bias), and σ^2 (a proxy for a learning weight, where large values allow for greater changes). I set μ to 0, under the assumption that participants did not know vowel harmony, and the preference for non-alternating items in harmonic items emerges after participants have learned that the language has vowel harmony. I set σ^2 to 0.1. This value was somewhat arbitrary, but previous models of artificial language learning have ranged from 0.01 (Wilson 2006) to 0.6 (White 2017). Setting this value too high yields categorical learning. Table 7 shows the learned weights of the constraints, while Table 8 shows the probability of selecting the 'correct' output, along with the mean responses for the human data from Finley (2021), and the replication from the present experiment.

	Agree[F]	ID[F]-go	ID[F]
Unspecified /mE/	0.97	0.97	0.97
Underlyingly /mo/	0.96	1.17	0.53
Underlyingly /me/	0.96	1.17	0.53

Table 7. Learned weights for each constraint

The model with an unspecified input learned a weight of 0.97 for all three constraints. This weight simulated the harmony bias for the non-alternating affix (e.g., 0.73 for New_GoF and 0.95 for New_GoB); there were no differences between New_MeF and New_meB items. The weights for both underlying forms (/me/ and /mo/) were identical; the morpheme-specific constraint was weighted highest, followed by the harmony-inducing constraint, and the general ID constraint was ranked lowest. This is the general analysis that was suggested in Finley (2010). While both assumptions for the underlying form produced the same constraint weights, the predicted responses were different, based on the stem vowel features. The correct response for [me/mo] items was selected more often for front vowel stems when the affix was underlyingly front, and the correct response for [me/mo] items was selected more often for back vowel stems when the affix was underlyingly back. Importantly, all three underlying forms simulated the overall difference between New_GoF and New_GoB items.

	New_goF	New-GoB	New_meF	New_meB
Finley 2021	0.70	0.83	0.67	0.71
Replication	0.59	0.86	0.62	0.72
Unspecified /mE/	0.73	0.95	0.73	0.73
Underlyingly /mo/	0.68	0.93	0.61	0.82
Underlyingly /me/	0.68	0.93	0.82	0.61

Table 8. Probability of selecting correct candidate

The model with an underlyingly back vowel for the alternating affix appears to best replicate the human data, as both the replication and Finley (2021) showed a numerical trend for a preference for New_meB items over New_meF items. While this may indicate that learners assumed that the underlying form of the alternating affix was back, this interpretation should be taken with caution. First, the predicted trend was not significant. Second, it is possible that the bias for correct responses in back vowel contexts in the replication could be driven by a bias to select the back vowel overall, since 75% of the affixed items in the training set ended in [o] (50% of the alternating forms, and 100% of the non-alternating forms). It is possible that this trend was more pronounced in the replication due to the online format of the replication study.

5. Discussion. This study presented a replication of Finley (2021), where participants were trained on a vowel harmony language with one alternating affix and another non-alternating affix. In both studies, participants learned the behavior of both the alternating and non-alternating affixes but showed a bias for the non-alternating affix in harmonic contexts. This result was simulated with MaxEnt Harmonic Grammar (Hayes and Wilson, 2008).

This paper has two major contributions. First, it replicates a previous finding using an online format, with a slightly different set of stimuli (e.g., a different talker). Replications are important for science to demonstrate the robustness of specific effects (Ebersole et al. 2016; Nosek et al. 2015), and there is a need for more replication projects in linguistics (Kobrock & Roettger 2022). One issue with the present replication is that even though the stimuli were re-recorded, and the modality was different, the author of the study and the replication was the same, which could potentially introduce some unintended bias into the process.

The artificial language learning results were successfully simulated using MaxEnt Harmonic Grammar. This is important because the harmony bias emerged despite training with categorical data. It also emerged without any exposure bias for harmony, as μ was set to 0. Previous modeled learning biases using perceptual data, by manipulating either μ (Finley 2022, White 2017) or σ^2 (Wilson 2006). In the present study, there were no set biases, and the result emerged regardless of the assumption about the underlying form of the alternating affixes. This supports the view that some learning biases can emerge from the formal properties of the grammar, as opposed to substantive grounding. This research opens the avenue for several lines of investigation about learning biases. The present study (as well as previous studies) used MaxEnt to simulate human learning after participant data had already been collected. While this can be useful for proof of concept and to better understand the biases, it has limitations. The desired results of the model are already known, and the researchers could 'tweak' the model to find the desired result. Ideally, researchers should use the MaxEnt grammar tool to make predictions about participant behavior. Knowing that artificial language learning can be successfully simulated with MaxEnt grammars leaves open the possibility for using the tool to better understand the mechanisms of human learning, and the representational biases therein. For example, the results of the simulation may be useful for understanding how learners make inferences about underlying forms. The simulation in the present study was run three times, each with a different underlying form for the non-alternating affix, with a slightly different result. Future research could explore different possibilities of assumptions about underlying representations in second or artificial language learning. Such research could be used to test assumptions about constraints, representations, and biases.

While this study has important contributions to the cognitive science of language, there are some limitations that must be acknowledged. First, the comparison between front and back vowel stems was made over a relatively small number of items (five in each category). While Finley (2021) increased that number in subsequent experiments, the small number of items overall is still a limitation. Second, while the MaxEnt model was successful at simulating human learners in both the replication and original study, the MaxEnt model did not account for additional factors that might have influenced learnability. For example, all the stems satisfied harmony, which could have helped learners discover vowel harmony in the language, and bias learners towards harmony. Second, the small vowel inventory invariably meant that vowel identity was created across syllables, which could have also influenced learnability of the harmony bias.

Finally, the fact that the affixed items ended in [o] on 75 percent of trials may have influenced participant responses. If some participants blindly 'probability matched', where they selected [o] 75% of the time, this would lead to a bias towards back vowels, and therefore more correct responses for items where the stem vowel was back. Looking at individual patterns of results, there appear to be six patterns of behavior overall. The first (n=8) showed responses at or above 60% for all test items and appeared to learn both affixes. The second type (n=9) showed a harmony bias, where they scored at or above 60% on all items, but lower than 25% on New-GoF items, suggesting they selected the harmonic response for all items. The third pattern (n=9) showed an [o] bias, selecting [o] on the majority of trials, thereby getting the New-go items correct, but getting the New_meF items wrong. The fourth pattern (n=3) appeared to associate [o] with the [go] affix, but [e] with the [me]/[mo] affix, scoring high on the New-go and New_meF items, but poorly on New_meB items. The fifth pattern (n=2) appeared to only learn the non-alternating affix, and then was around chance for the New_me items. The sixth pattern (n=8) showed no clear pattern and had average scores below 60% suggesting that they did not learn the patterns or were merely guessing throughout. This high number of different patterns of individual learning makes it difficult to draw strong conclusions from the MaxEnt model because the MaxEnt learning tool only presents a single solution, thoguh this could be seen as the average of all types of learners. It also raises the possibility that the harmony bias emerges because high responses to New-goF items is consistent with many possible (though incorrect) interpretations of the training data. Future research should work to better integrate different learning strategies into understanding participant behavior and interpreting the results in a broader cognitive context.

6. Conclusion. I have presented a replication of the harmony bias for non-alternating affixes first reported in Finley (2021) using online, remote data collection. The results were generally similar to the in-person results from the original study, supporting the general robustness of this effect. In addition, these results were simulated using MaxEnt Harmonic Grammar. The results were relatively similar for all representations of the underlying form of the alternating affix, further demonstrating that categorical training data can yield non-categorical results in learning, and that learning biases can emerge without any specific bias built into the constraints or learning weights.

References

Alderete, John & Sara Finley. 2016. Gradient vowel harmony in Oceanic. *Language and Linguistics* 17(6). 769–796. https://doi.org/10.1177/1606822X16660960.

Baković, Eric. 2000. *Harmony, dominance and control*. New Brunswick, NJ: Rutgers University dissertation.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using {lme4}. *Journal of Statistical Software* 67. 1–48. https://doi.org/10.18637/jss.v067.i01.

Boersma, Paul & David Weenink. 2017. Praat: Doing phonetics by computer. https://www.fon.hum.uva.nl/praat/

Culbertson, Jennifer. 2012. Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass* 6(5). 310–329. https://doi.org/10.1002/lnc3.338.

Ebersole, Charles R., Olivia E. Atherton, Aimee L. Belanger, Hayley M. Skulborstad, Jill M. Allen, Jonathan B. Banks, Erica Baranski, et al. 2016. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* 67. 68–82. https://doi.org/10.1016/j.jesp.2015.10.012.

Ettlinger, Marc, Kara Morgan-Short, Mandy Faretta-Stutenberg & Patrick C. M. Wong. 2016. The relationship between artificial and second language learning. *Cognitive Science* 40(4). 822–847. https://doi.org/10.1111/cogs.12257.

FindingFive Team. 2019. *FindingFive: A web platform for creating, running, and managing your studies in one place.* NJ, USA: FindingFive Corporation. https://www.findingfive.com.

Finley, Sara. 2010. Exceptions in vowel harmony are local. *Lingua* 120(6). 1549–1566.

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.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory*, 111–120. Stockholm: Stockholm University.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440. https://doi.org/10.1162/ling.2008.39.3.379.

Hayes, Bruce, Colin Wilson & Benjamin George. 2009. Manual for Maxent grammar tool. Manuscript. UCLA.

Kobrock, Kristina & Timo B. Roettger. 2022. Assessing the replication landscape in experimental linguistics. Manuscript. University of Oslo. https://osf.io/9ceas/.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. Harmonic Grammar–A formal multi-level connectionist theory of linguistic well-formedness. *Proceedings of the Annual Meeting of the Cognitive Science Society* 12. 884–891.

Lenth, Russell, H. Singmann, J. Love, P. Buerkner & M. Herve. 2019. Package 'emmeans.' https://cran.r-project.org/web/packages/emmeans/index.html

Moreton, Elliott & Joe Pater. 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.

Nosek, B. A., G. Alter, G. C. Banks, S. D. Borsboom D. Bowman, [...] & T. Yarkoni. 2015. Promoting an open research culture. *Science* 348(6242). 1422–1425. https://doi.org/10.1126/science.aab2374.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*. 349(6251). https://doi.org/10.1126/science.aac4716.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/.

Roettger, Timo B. & Dinah Baer-Henney. 2019. Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis* 1(4). 1–23. https://doi.org/10.3765/pda.v1art4.13

RStudio Team. 2020. RStudio: Integrated development for R. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Strütjen, Kim, Dinah Baer-Henney, Peter Indefrey & Ruben van de Vijver. 2018. Perceptual bias in learning a vowel nasalization pattern. Manuscript. Heinrich Heine University.

White, James Clifford. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language* 93(1). 1–36. https://doi:10.1353/lan.2017.0001.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30. 945–982. https://doi.org/10.1207/s15516709cog0000_89.