An Investigation of the Tolerance Principle with Two Productive Rules in Child Language Acquisition

Shengqi (Iris) Zhong[1]

[1] Smith College

Author Note

Submitted to the Department of Psychology of Smith College in partial fulfillment of the requirements for the degree of Bachelor of Arts with Honors.

Acknowledgment

First, I want to thank my thesis committee, Professor Jill de Villiers, Professor Brianna McMillan, and Professor Kathryn Schuler, for generously taking the time and effort to review this work.

Next, I want to express my sincere gratuitude to my thesis advisor Professor Jill de Villiers. This thesis would not have been possible without the training and support from her. I have had the greatest time in her classes and her lab, and I would not have studied language acquisition if I had not met her at Smith. I also want to thank her for bringing me to the Boston University Conference on Language Development, when I was first exposed to the Tolerance Principle. Lastly, she has provided enormous support for my thesis, from helping in designing the experiment, to drafting proposals, to giving feedback.

I am grateful to Professor Kathryn Schuler from the University of Pennsylvania for offering me the opportunity to work in her lab, and sending helpful comments when I was designing the experiment.

I also want to thank my thesis companions, Jenna Croteau and Lydia Quevedo, for helping me with coding and publishing the experiment on Lookit.

I would like to thank Rongzhi Liu from the University of California, Berkeley, who mentored me during Summer 2020 and gave useful suggestions on this research.

I want to thank Jill's grandson, Graham, and Jenna's cousin, Jackson, for spending their time piloting my experiment, and providing feedback for me.

Lastly, I would like to thank all of my friends and family members who have supported me throughout the year. Especially, I want to thank my housemates Rachel Yan, Elaine Ye, and Yanwan Zhu, and my best friend Yena Li for their emotional, technical and academic support in this difficult time.

Contents

Introduction

Over-regularization in Language Acquisition

An amazing characteristic of language acquisition is that children are able to learn from limited input to generate a sophisticated language system. Needless to say, acquiring rules and generalizing them to new instances is an important strategy that children use all the time. However, no rules are perfect, and exceptions always exist, which makes learning more complicated.

Think about learning English past tense. Most of the verbs use a regular past tense marker, adding -ed at the end. The rest of the verbs that do not use the regular inflection are irregulars. Intuitively, children's learning of past tense markers should improve over time: first, they might misuse the marker frequently, but with more exposure in a few years, they should produce fewer mistakes. In fact, children's acquisition of past tense typically goes through three stages: at the first stage, when children have just started using the past tense, they produce irregulars correctly; during the second stage, however, they overregularize – applying the regular past tense marker on irregular verb stems; in the last stage, their production approximates that of adults, and overgeneralization becomes much less frequent (Taatgen & Dijkstra, 2003).

Such a U-shaped pattern of acquisition, characterized by good language performance followed by bad performance, and back to good performance again (Carlucci & Case, 2013), has been studied for a long time. Two major explanations are connectionist theory and dual-processing theory (Elsen, 1998). The connectionist model links phonological and representations of verb stem and past tense together through a complex network (Rumelhart & McClelland, 1985). It essentially states that no explicit rules are acquired, and past tense inflection is gradually learned, rather than suddenly shifted (McClelland & Patterson, 2002). A simulation using the connectionist model does show a U-shaped learning curve.

On the other hand, dual processing theory (i.e., words-and-rules theory) proposes that rule and lexicon learning are both in place (Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995; Pinker & Prince, 1988; Pinker & Ullman, 2002). According to the words-and-rules model, children first store all regulars and irregulars as lexicons, and learn both by rote or through an associative network similar to the connectionist model. Then, as more regular verbs are learned, they generate the -ed inflection rule and generalize to every verb. Finally, they discover the exceptions of the rule and separate them from the regulars; the irregulars are again stored as individual lexical items (Elsen, 1998). The model also gains support at the neurobiological level. Patients with posterior aphasia having word-finding difficulties and patients with Alzheimer's disease made more mistakes in producing irregulars than regulars and novel verbs, indicating that deficiency in vocabulary memory storage and retrieval blocks irregular past tense inflections. Meanwhile, anterior aphasic patients with grammatical impairment failed to produce regular verbs correctly (Ullman et al., 1997). This study showed that there are two separate systems in learning past tense inflection – grammar, which captures the rule for the majority, and lexicon, which stores the exceptions.

Though the connectionist and the dual processing model vary in explaining how the children learn the rules, both agree, to some extent, that U-shaped learning is related to balancing rules and exceptions (Carlucci & Case, 2013). A recent approach to understanding the relationship between rule formation and exceptions is the Tolerance Principle.

The Tolerance Principle

The Tolerance Principle tells whether a rule R is productive, given the total number of items N and the number of exceptions e (Yang, 2016b): R is productive and should be generated if and only if:

$e \leqslant \frac{N}{lnN}$;

Else, R is unproductive if and only if:

$e > \frac{N}{lnN}$,

and all items will be treated as irregulars and are memorized by order of the rank of their frequencies (Yang, 2016b).

The Tolerance Principle is derived from the Elsewhere Condition. The Elsewhere Condition states that in order for one to confirm that the regular marker is the correct inflection for a stem, the learner has to go through all of the exceptions and discover that the stem is not one of them, before reaching the conclusion that the regular marker is right (Kiparsky, 1973).

The Tolerance Principle assumes a Zipfian distribution in the language input. Zipfian distribution adheres to the Zipf's law, such that the frequency of an item is inversely proportional to its frequency rank (Zipf, 2016). For example, if the second frequent word (i.e., Rank #2) is half as frequent as the most frequent word (i.e., Rank #1), and the third frequent word (i.e., Rank #3) is half as frequent as the second frequent word, and so on, then it is a Zipfian distribution (Schuler, Yang, & Newport, 2021). By using this relationship between word rank and frequency, the Tolerance Principle tells us, for any list of items, what the search time is i) if a rule is generated with a certain number of exceptions, or ii) if every instance is memorized. Then, it will evaluate whether option i) or ii) is more cost-efficient.

Note that the Tolerance Principle analyzes type frequency, rather than token frequency. Type frequency refers to the number of distinct items in a set, while token frequency is the count of repetitions of these items (Bybee & Thompson, 1997; Koulaguina & Shi, 2019). For instance, the sing → sang inflection can occur multiple times in the language input, which means it has a high token frequency; however, because it is just one distinct instance, it can be counted only once as type frequency in the Tolerance Principle evaluation. Using type frequency instead of token frequency is reasonable, because children

have to gain enough exposure to a variety of items before they can claim a productive rule (Yang, 2016b).

In terms of explaining children's U-shaped learning of the regular past tense marker, the Tolerance Principle is a step forward from the dual-processing model, as it gives a mathematical calculation with an accurate prediction of when over-regularization starts and ends. At first, young children can produce the irregular past tense inflections correctly, because they have not formed a productive rule with their limited vocabulary, and memorize every instance by rote; but when they have learned a large set of verbs with regular marker -ed with only a few exceptions, they reach the threshold of the Tolerance Principle, generate a productive rule, and generalize it to all of the novel verbs; finally, they recognize the exceptions and stop overgeneralizing the regular marker (Yang & Montrul, 2017).

If we take a closer look at some example thresholds of the Tolerance Principle (see *Table 1*), we can spot an interesting pattern (Gorman & Yang, 2019): if one has a small vocabulary size (i.e., N is small), they can tolerate a higher proportion of exceptions, which might potentially explain children's ability to detect rules fast.

N (total number of items)	θ_N (maximum number of exceptions)	% (exception/total)
10	4	40.00
20	7	35.00
50	13	26.00
100	22	22.00
200	38	19.00
500	80	16.00
1000	145	14.50
5000	587	11.70

Table 1. Examples thresholds of the Tolerance Principle.

Applications of the Tolerance Principle

The Tolerance Principle essentially tells how rules are formed, and because rules can be studied in all subjects, we can see the potential for the application of the Tolerance Principle in a lot of domains. In this section, I will review research on the Tolerance Principle in various fields of linguistics.

First, besides the English past tense discussed in the earlier sections, topics in morphology such as plurals are also studied. For instance, Yang (2016b) explained the reason why English noun plurals are acquired earlier than verb past tense. Plural inflections have much fewer exceptions than past tense, and therefore the plural regular marker rule is productive even when N is as few as 10, and continues to be productive as N increases. Whereas for past tense, because the irregular verbs are highly frequent, they are learned early in the acquisition process, and the -ed regular rule does not formulate until one has learned approximately 1000 verbs. The Tolerance Principle helps explain the different trajectories in the acquisition of plural and past tense.

The Tolerance Principle is applied to syntax research as well. Li, Grohe, Schulz, and Yang (2021) examined the recursive structure in the possessives. In English, there are two kinds of possessives: X's Y, and Y of X. Their corpus study showed that the X's Y structure is recursive because it meets the Tolerance Principle threshold, such that of the nouns that appear in the Y position, most of them also appear in the X position. Similarly, recursion in Y of X also satisfies the Tolerance Principle. A closer look at the examples in Y of X reveals that almost all of the nouns in the Y position are inanimate and represent an internal possession of X (i.e., an inherent property of X). Therefore, the authors speculated that Y of X is not usually seen in recursion, because it is unusual to form a chain of internal possession.

I want to make a special mention of Yang's Tolerance Principle approach in analyzing number acquisition (Yang, 2016a). Of the first a hundred numbers in English, there are 17

irregular numerals that have to be learned by rote (one to fifteen except fourteen, twenty, thirty, fifty), and the others can be deducted from existing numbers. To tolerate 17 exceptions, the Tolerance Principle rules that the smallest number of N is 73, which is exactly the threshold found by Fuson, Richards, and Briars (1982). Yang continued to analyze the case in Mandarin, in which only eleven numerals have to be rote-learned. Therefore, it is predicted that the smallest N that children have to get to to learn the first a hundred numerals is 42, which matches with previous empirical findings (Miller, Kelly, & Zhou, 2005).

The studies above demonstrate that the Tolerance Principle works well in theory and in corpus studies. Next, I will discuss empirical research in statistical learning, and in particular, experiments on the Tolerance Principle.

Empirical Research in Statistical Learning

The Tolerance Principle is not the first attempt in decoding children's abilities to learn from patterns. A lot of experiments have been conducted on infants' and children's statistical learning in language acquisition. Many adopted an artificial language paradigm, because compared to a natural language, artificial language is designed by the researchers and therefore can control for a lot of variables (Gómez & Gerken, 2000).

For instance, research by Saffran, Aslin, and Newport (1996) investigated infants' use of statistical cues to segment words. Eight-month-old infants were exposed to a 2-min speech stream with four three-syllable nonsense words, where the transitional probabilities between the syllable pairs were arranged to be higher within words and lower between words all the time. They found that babies looked longer in the test when the transitional probabilities were different from that from the exposure, suggesting that infants were able to track transitional probabilities within and between words to facilitate word segmentation.

A study by Gerken (2006) explored infants' choice of rule generalization when more than one option is available. In this experiment, nine-month-olds were familiarized with nonsense three-syllable strings. The participants were separated into AAB and ABA groups, and were further divided into Condition 1 and Condition 2, respectively. Infants in Condition 1 were exposed to a series of broader inputs than the infants in Condition 2. In AAB groups, babies from Condition 1 heard words in AAB format (A and B represent different syllables); the babies in Condition 2 heard items in AAdi format, and therefore the third syllable in their input had to be di. Similarly, in ABA groups, babies from Condition 1 were exposed to ABA items, while those from Condition 2 heard AdiA items. In the test trials, infants heard novel words in AAB and ABA format. Those in the AAB group Condition 2 (exposed to AAdi) treated AAB as an unfamiliar word, and so did children in the ABA group Condition 2 (exposed to AdiA) when they heard ABA strings. Thus, infants who only heard AAdi and AdiA made a narrow generalization, and the third syllable in the AAdi condition and the second syllable in the AdiA condition were fixed to be di in their rules. When listening to strings that belonged to the same category in a broader scope, they did not classify them as the same kind as the input.

In a follow-up study, Gerken asked whether infants could switch to the broader generalization if they were exposed to counterexamples to the narrow rule (2010). Infants were randomly divided into AAB or ABA groups. The participants of interest were exposed to the same strings as Condition 2 in the previous study (i.e., AAdi or AdiA), but in addition, three AAB strings (i.e., third syllable was not di) in AAB group and three ABA strings (i.e., second syllable was not di) in ABA group were played at the end of the familiarization phase. With just three exceptions, the infants were able to switch from a narrow (AAdi or AdiA) to a broader (AAB or ABA) generalization.

Though not discussed in the article, Gerken's findings can be explained by the Tolerance Principle. The 2010 study design included 7 items, four of which were AAdi/AdiA strings, and three were AAB/ABA strings. The Tolerance Principle rules that

if two grammars are available at the same time, the learner will choose the one with fewer exceptions (Yang, 2016b). In this case, the AAB/ABA rule apparently wins over the AAdi/AdiA rule, as there are no exceptions to the broad AAB/ABA rule, but three counterexamples to the narrower AAdi/AdiA rule.

Kam and Newport (2005) noticed a surprising difference between adults and children in learning from patterns. They tested the participants' use of determiners in an artificial language, in which unpredictable inconsistencies existed along with the regular rule. Adults were found to retain the inconsistency, and did not over-regularize the language. In comparison, children tended to use the determiner rule everywhere, disregarding the inconsistencies present in the input. This finding was echoed by subsequent research in the Tolerance Principle as well (Newport, 2020; Schuler et al., 2021).

Recent emerging studies have extensively studied and discussed the Tolerance Principle in practice. Schuler et al. (2021) investigated the acquisition of novel noun plurals using an artificial language learning paradigm. Nine novel nouns were created, and a special morpheme was designed to be a regular plural marker. Six more novel morphemes served as the irregulars. In one condition, five out of the nine nouns had the regular marker ($5R/4E$ condition, meaning 5 regulars and 4 exceptions), and the rule should be generated, according to the Tolerance Principle ($4 < \frac{9}{ln9}$). In the other condition, six out of the nine plurals were irregular ($3R/6E$, meaning 3 regulars and 6 exceptions), and therefore the regular rule was unproductive ($6 > \frac{9}{ln9}$). Seventy-two sentences, including 24 with singular and 48 with plural nouns, were generated. Singular sentences were unmarked and paired with an image of one corresponding noun object. Plural sentences were marked with inflection and were paired with 2, 4, or 6 images of the corresponding object. Such combinations gave 18 possible sentences, one singular and one plural sentence for each noun. Then, sentences were repeated a number of times, so that the frequency of the nouns followed the Zipfian distribution, which was required before using the Tolerance Principle. Therefore, the most frequent word in the final stimuli was presented about twice as often

as the second most frequent, and the second most frequent was about twice as often as the third most frequent word, and so on.

Children and adults were randomly assigned to one of the $5R/4E$ and $3R/6E$ conditions. They were shown 72 sentences in random order. A test with twelve trials of novel nouns assessed how they produced plurals given the singulars. Results showed that children obeyed the Tolerance Principle: they over-regularized in the former ($5R/4E$) condition with the novel nouns almost every time, but selected the regular marker no more than chance in the $3R/6E$ condition. On the other hand, adults tracked the token frequency of the markers, known as probability matching. In the $5R/4E$ condition, they produced the regular marker close to the token frequency of the regular marker during exposure, significantly lower than children's production of the regular marker (100%). In the $3R/6E$ condition, adults produced the regular inflection at the rate of which also approximated its token frequency during exposure, significantly different from children's usage.

Several concerns were addressed in the subsequent studies. In the previous experiment, the nouns with the regular marker received higher token frequencies than the exceptions, which is not always the case in a natural language. In Experiment 2, the regulars and the exceptions were evenly distributed, both at the high and the low ends. While adults in both $5R/4E$ and $3R/6E$ conditions, and children in $3R/6E$ behaved similarly as before, the children in the $5R/4E$ condition did not perform as expected by the Tolerance Principle – they answered with the regular inflection significantly less than 100% of the time. The authors proposed that the children could have applied the Tolerance Principle to only part of the items in the exposure, because they only remembered this subset of items. For instance, if children missed a regular item in the $5R/4E$ condition, the input they received would change to $4R/4E$, and they should not form a rule based on the Tolerance Principle. After the researchers calculated the "personal Tolerance Principle" for each child who answered categorically given their performance in rating test trials, most of

the children in the experiment were shown to use the Tolerance Principle to form rules.

To summarize, there is ample evidence to show that infants as young as eight months old are capable of extracting rules from language input. Nevertheless, more empirical research is needed to investigate children's usage of the Tolerance Principle in reality, and the distinction of rule learning in children and adults.

Research Question and Hypothesis

In the first section, I talked about children overusing the regular rule on irregular verbs. On the other hand, we can spot some vague rules among the irregulars as well: for example, verbs that end with -ing (e.g., sing, ring) often inflect like ɪ→æ/_ŋ. However, "over-irregularization," the generalization of irregular markers on the regular stems, is rarely observed in corpora of child language. Why can children usually avoid generalizing irregular rules for English past tense (Xu & Pinker, 1995)?

A convincing explanation is that most of these "rules" generate more exceptions than the threshold of the Tolerance Principle suggests. Let us examine the irregular rule ɪ→æ/_ŋ, one of the few rules that children occasionally use (Xu & Pinker, 1995; Yang, 2016b). Of the first two hundred verbs that children learn, three end with -ing: bring, ring and sing. Two of them follow the irregular rule.

Since $1 < \frac{3}{ln3}$, the rule is productive and it is expected that young children would over-irregularize. The Tolerance Principle is satisfied until the vocabulary size grows to around eight hundred verbs. At this time, most children know eight verbs ending with -ing: bring, fling, ring, sing, spring, sting, swing, wing. Five out of eight do not meet the irregular rule. Five is greater than $\frac{8}{ln8}$, and thus the rule will be abandoned (Yang, 2016b). So it is likely that the over-irregularizaion of ɪ→æ/_ŋ is observed when children are at the transition stage as they expand the vocabulary, but once they have learned enough verbs to override the irregular rule, they stop using it.

What if in another language, the irregular rule is actually productive on its own? When children see a novel item, do they make a subset for the seemingly irregular stems and generate two sub-rules, one for the subset and one for the others, if both rules are productive? Or do they just generate one general rule applicable to all verbs?

In other words, if the children apply the Tolerance Principle, under the circumstance that both the regular rule and the irregular rule are productive, they have two competing choices when they see a novel word:

Rule (a): use the irregular rule when the stem looks like an irregular, and use the regular rule for the rest;

Rule (b): use the regular rule for all of the novel words.

The Tolerance Principle predicts that children's response is categorical. Namely, they should apply the rule to all of the corresponding instances. Therefore, children who form Rule (a) should use the irregular rule on the irregular-looking stem, and the regular rule on the rest 100% of the time; those who apply Rule (b) should use the regular marker on every novel stem they have encountered.

According to Yang (2016b), children always start from a simple grammar. Only if they find out the rule is not productive will they continue to look for subdivisions and subregularities that satisfy the Tolerance Principle. That is to say, Rule (a) proposed in the hypothesis will not be in place, as children will stop at the regular rule because it is already productive. However, there are no known empirical studies to investigate this question, and the current research aims to bridge the gap between theory and practice.

Methods

Design

The experiment adopted an artificial language learning paradigm, consistent with other experiments on statistical learning (Gerken, 2006, 2010; Saffran et al., 1996; Schuler et al., 2021). Instead of audio-only stimuli like in Gerken's study (2006), the novel words in this experiment were associated with actual meanings and images. The main reason was that the target children in the study were 4 to 8 years old, and it was virtually impossible for them to merely listen to nonsense recordings for over ten minutes without losing attention.

The experiment followed a similar design as Schuler's (2021). First, to ensure that where the regular or the irregular markers are placed in the rank does not contribute significantly to children's acquisition of the rule, I created two conditions, in which regulars and irregulars were ranked with different token frequencies. Furthermore, in order to address the possibility that children could forget the least frequent words and memorize exclusively the words with high frequency, I manipulated the frequency rank in both conditions so that children could still possibly derive Rule (a) and Rule (b) if they only catch the six most frequent instances in the exposure.

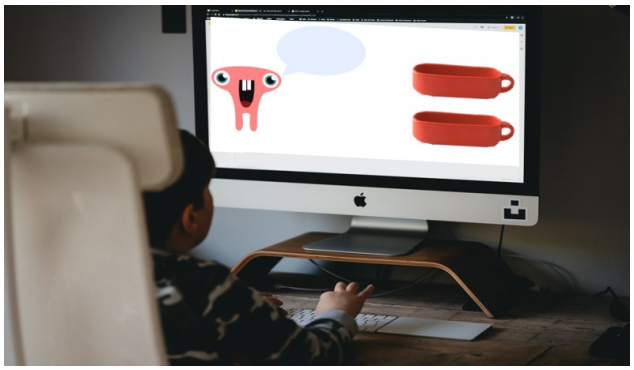


Figure 1. A demonstration of a child participating the experiment.

Exposure Phase.

Type Frequency.

Since the research question is neutral with respect to the kind of linguistic form to test, noun plurals were selected as the study target. Two kinds of novel noun stems were created: AB pattern and AA pattern, where A and B represent syllables. In other words, all noun stems are disyllabic, and the AB pattern noun was constructed by two varying syllables, while the AA pattern noun was two identical syllables combined. Two morphemes were selected as plural markers. ta was the more frequent and therefore the regular plural marker; pi was less frequent and therefore an irregular marker. There were eight noun stems in AB format (kasi, disa, saku, seki, kosa, keso, mabu, kupa) and four in AA format (fofo, keke, meme, susu). All of the noun stems only used vowels from /a/, /ɛ/,/u/, /i/, and /oʊ/. All of the consonants in these novel nouns were English consonants. Each noun had a predetermined plural marker. Of the eight AB format nouns, seven used the ta marker, and one used the pi marker. Of the four AA format nouns, three had pi as the plural marker, and one used the ta marker (see *Table 2*).

	ta	*pi*
AB format	7.00	1.00
AA format	1.00	3.00

Table 2. Type frequency distribution.

Rule (a) is productive in this design, because 7 out of the 8 AB format nouns use ta marker, and 3 out of the 4 AA format nouns use pi marker. Rule (b) is also productive because 8 out of 12 nouns use the ta marker.

Each noun can participate in two types of sentences, one in the singular form, and the other in the plural form. All of the sentences were constructed by a form equivalent to an expletive followed by the noun in singular or plural form. For instance, if the noun stem is pati, its marker is ta, and the expletive is sachi, then the singular and plural sentences are:

Sachi pati.

There (is a) pati.

Sachi patita.

There (are) pati(s).

Token Frequency.

The sentences were replicated various times so that the frequency of the nouns demonstrated a Zipfian distribution. Thus, the most frequent word was about twice as frequent as the second frequent word, and the second frequent word was about twice as frequent as the third frequent word, and so on. There were 69 exposure trials in total. In each trial, the sentence was played twice, and therefore in sum, the children would hear 138 sentences. The token frequency was distributed as in *Table 3*. For each stem, The ratio of the number of singular noun trials to the number of plural noun trials was always 1:2. For instance, if a noun shows up in 15 trials, then 5 of them use the singular form (i.e., the noun stem), and 10 of them use the plural form (i.e., noun stem plus the plural inflection).

Creation of Conditions.

Additionally, the Tolerance Principle accounts for type, instead of token frequency. Thus, the number of times a certain word appears should not affect the performance of the Tolerance Principle, as long as the type frequency is identical and the tokens are distributed in accordance with Zipf's law. To ensure that token frequency did not play a role, two conditions were created: Condition 1 contained fewer ta markers than Condition 2. Specifically, out of the 46 trials that were constructed with plural nouns, 26 trials contained nouns with the ta suffix in Condition 1 (see *Table 3*): disa (n = 6), keso (n = 4), mabu (n = 4), seki (n = 4), keke (n = 2), kupa (n = 2), kasi (n = 2), and kosa (n = 2). Thus, the ta marker showed up approximately 57% of the time in Condition 1. In condition 2, 34 trials had nouns with the ta marker (see *Table 3*): kasi (n = 10), kosa (n = 6), disa (n = 6), seki (n = 4), keso (n = 2), mabu (n = 2), kupa (n = 2), and keke (n = 2).

Frequency Rank	Condition 1 Noun Stem & Plural Marker	Condition 2 Noun Stem & Plural Marker	Number of trials (Singular + Plural)	Number of trials (Singular)	Number of trials (Plural)
1	fofo(pi)	kasi(ta)	15	5	10
2	saku(pi)	kosa(ta)	9	3	6
2	disa(ta)	disa(ta)	9	3	6
4	keso(ta)	saku(pi)	6	2	4
4	mabu(ta)	seki(ta)	6	2	4
4	seki(ta)	susu(pi)	6	2	4
7	keke(ta)	keso(ta)	3	1	2
7	kupa(ta)	mabu(ta)	3	1	2
7	meme(pi)	kupa(ta)	3	1	2
7	kasi(ta)	fofo(pi)	3	1	2
7	kosa(ta)	meme(pi)	3	1	2
7	susu(pi)	keke(ta)	3	1	2
Total			69	23	46

Table 3. Token frequency distribution.

The ta marker turned up approximately 74% of the time in Condition 2. Because it was hypothesized that the children would use the Tolerance Principle to generalize rules, token frequency should not exhibit any interference effect, and no differences were predicted in children's output between the two conditions.

Type Frequency of the Top 6.

Because children would have little exposure to the nouns with the lowest frequencies (number of trials = 3), there is concern that when children acquire the rule, they would forget and overlook these nouns with low frequencies. Thus, the six nouns with the highest frequencies were given extra attention by making sure that these six nouns on their own could also derive the two possible rules in the hypothesis in both Condition 1 and 2 (see *Table 3* and *Table 4*). In Condition 1, the top six nouns consist of four AB pattern nouns with a ta marker (disa, keso, mabu, seki), one AB pattern noun with a pi marker (saku), and one AA pattern noun with a pi marker (fofo). In Condition 2, the top six nouns are composed of four AB pattern nouns with a ta marker (kasi, kosa, disa, seki), one AB pattern noun with a pi marker (saku), and one AA pattern noun with a pi marker (susu).

	ta	*pi*
AB format	4	1
AA format	0	1

Table 4. Top 6 type frequency distribution.

Rule (a) is productive among the top six, because 4 out of the 5 AB format nouns use ta marker, and the only AA format noun uses the pi marker. Rule (b) is also productive because 4 out of 6 nouns use the ta marker.

Test Phase. The test phase was composed of ten trials. In each trial, the children were first exposed to a single novel object with its singular novel expression. Then, they would hear two different versions of plural sentences, and select which one they thought was better. Since the most interesting question was whether the participants would use ta or pi marker in AA format noun, children were asked to select a marker for AA format novel noun stems for 7 times: nene, momo, lolo, sese, fafa, tete, and dede. Then, two questions were asked to choose between ta or pi for AB format noun stems – pabo and beku, and the children were expected to select the ta marker all the time. Finally, there was one attention check question, that asked the children to choose between ta and a marker that had never appeared in the test (ma) for an AB format noun stem – kesa. If the children paid attention to the test, it was predicted that they would prefer the ta marker. The order of the testing items was also randomized.

Predictions

The following outcomes are all possible given the experimental design:

1. Children follow the Tolerance Principle, and generalize either Rule (a) or Rule (b) to the novel nouns in the test phase.

- If children produce Rule (a), then they would choose to use the pi marker 100% of the time when they encountered a AA format noun stem, and use the ta marker 100% of the time when they encountered a AB format noun stem. Their results would not differ between Condition 1 and Condition 2, as token frequency does not influence the performance of the Tolerance Principle.
- If children produced Rule (b), then they would choose to use the ta marker 100% of the time. Their results would not differ between Condition 1 and Condition 2, as token frequency does not influence the performance of the Tolerance Principle.

2. Children adopt the probability matching strategy, using the token frequency in the exposure phase as a clue.

- If the children were in Condition 1, they would use the ta marker approximately 57% of the time.
- If the children were in Condition 2, they would use the ta marker approximately 74% of the time.

3. The children would not grasp any of the rules, and choose their answers at chance. Since each question in the test phase had two available choices, children would use the ta marker around 50% of the time.

Procedure

The experiment was conducted through Lookit, a platform that allows families to participate in online behavioral studies at home via webcam (Scott & Schulz, 2017). After the adults and the children had provided the consent and checked the webcam setup, the children were introduced to the background of this experiment: "A group of aliens from outer space are visiting the Earth. They are describing the things they bring from their planet. But because they speak an alien language, no human beings can understand them! Try your best to guess the rule of their language!" (see *Figure 2*)

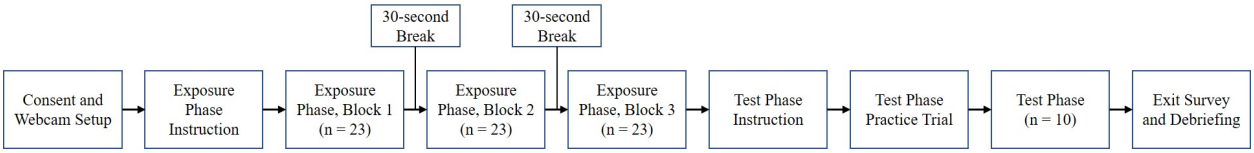


Figure 2. Experiment procedure flow chart.

Then, in the exposure phase, the children were randomly assigned into one of the two conditions, and watched and heard 69 randomized trials of artificial language stimuli described in the previous section. Each trial was 11 seconds long and consisted of an audio

recording done by the researcher utilizing Audacity, and a picture demonstrating an alien describing one novel object or two of the same kind, depending on whether the noun is in singular or plural form (see *Figure 3*). To retain children's attention, the trials were grouped into three blocks, each consisting of 23 trials. They were given a 30-second break after the first block and the second block. In order to minimize the overall primacy and recency effects (i.e., children had a more vivid memory on trials at the beginning and the end), the noun stems were evenly distributed to the three blocks: for example, if a noun was designated to participate in 15 sentences (5 singular and 10 plural sentences combined), it would appear 5 times in each block. However, the number of singular and plural sentences respectively for each noun in a block were randomized. In other words, while the noun in this example should show up 5 times in a block, the number of times it was in a singular sentence versus in a plural sentence was randomly assigned. Nevertheless, the ratio of singular to plural sentences for any noun stem was arranged to be 1:2 overall.

Figure 3. Exposure trial example. The left side shows a singular noun trial, and the right side shows a plural noun trial.

Then, the children were given instructions on the test trials: "After the aliens have talked to you, some human beings are going to try to speak their alien language. We will ask you to decide which human talks like the alien best. But there are no right or wrong answers to these questions. You can click on the picture of the person to make a selection." The parents were allowed to help with clicking if the children were unable to do so, but they were told not to offer any assistance in answering the questions.

The children were first familiarized with the question layout in a practice trial, in which an ordinary object (a clock) was described in English, and the children were asked to select the person who described the location of the clock correctly according to a picture given. Audio feedback was given to inform the children whether their selection was correct or not. When they finished the practice trial, the participants were reminded that unlike the practice, there was no right or wrong answer to the actual test questions.

In each of the ten test trials, the children saw and heard an alien describing a single novel object that was not in the exposure phase (see *Figure 4*). Then, they saw a picture of two objects in the middle of the screen, the same as the one shown earlier. Two images of human characters were on the left and right side of the screen, and these two individuals would try to describe the plural objects in the alien language. The participants heard two recordings representing speech from these two human characters. A blue frame and a speech bubble showed around the human character image to represent that the person was speaking (see *Figure 4*). The children were then required to choose one person who they thought spoke the alien language better by clicking on the character image. They were forced-choice questions, as children had to select one of the images to proceed.

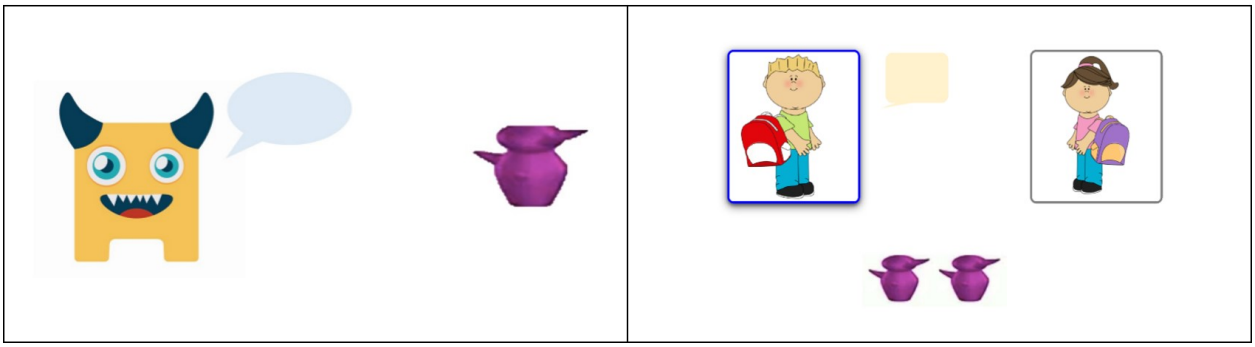


Figure 4. Test trial example. The left side represents an alien describing a single novel object. The right side shows the test when the left character is speaking.

When the children finished the ten test trials, their parent/caregiver was asked to fill out a survey indicating their preferred data privacy level. Lastly, a debriefing section

explained the gist of the experiment (see *Appendix A*). The participants received certificates as a token of appreciation (see *Appendix B*).

Participants

A hundred and eleven responses were collected from English-speaking four-to-eight-year-old children, either through an invitation from Lookit or promotion on social media. Fifty-one responses were excluded from the study either because the children did not finish the study (n = 27), participated more than once (n = 3), had a significant birthdate difference between what was indicated in the experiment and in the Lookit profile (n = 2), consistently chose the character on one side in the test trials (n = 4), or constantly preferred the characters of the same gender in the test phase (n = 15).

The final sample included 28 girls, 31 boys, and one child that identified with some other gender. Fifteen of them were younger than 5 years old, fifteen were 5 to 6 years old, eighteen were 6 to 7 years old, and twelve were older than 7. Approximately 81.7% of the children came from a White family, 18.3% came from an Asian family, 3.3% were from a Black or African American background, 1.7% came from a Middle Eastern or North African background, 10.0% had a Hispanic, Latino, or Spanish origin, and 1.7% identified with another race, ethnicity or origin. Most of the parents/guardians were 30-34 (23.3%), 35-39 (51.7%), or 40-44 years old (20%). The families in this sample on average had a higher education level than the general population, with more than 60% of the parent/guardian holding a graduate or professional degree. Similarly, their annual family income was also higher than the average population, as over 45% of the families earned more than 100,000 dollars each year. All children were English speakers, and 26 of them were bilingual or multilingual. See *Table 5* for details.

	Total Count (%) n = 60
Child gender	
Female	28 (46.7%)
Male	31 (51.7%)
Other	1 (1.7%)
Child age	
Under 5	15 (25%)
5 - 6	15 (25%)
6 - 7	18 (30%)
Over 7	12 (20%)
Race and ethnicity	
White	49 (81.7%)
Asian	11 (18.3%)
Black or African American	2 (3.3%)
Middle Eastern or North African	1 (1.7%)
Hispanic, Latino, or Spanish origin	6 (10.0%)
Another race, ethnicity, or origin	1 (1.7%)
Parent/Guardian age	
30 - 34	14 (23.3%)
35 - 39	31 (51.7%)
40 - 44	12 (20%)
45 - 49	2 (3.3%)
45 - 59	1 (1.7%)
Family annual income	
Under 50,000	7 (11.7%)
50,000 - 100,000	16 (26.7%)
100,000 - 150,000	12 (20.0%)
150,000 - 200,000	8 (13.3%)
Over 200,000	8 (13.3%)
missing	9 (15.0%)
Parent/Guardian education level	
Some or attending college	1 (1.7%)
2-year college degree	2 (3.3%)
4-year college degree	19 (31.7%)
Some or attending graduate or professional school	2 (3.3%)
Graduate or professional degree	36 (60%)

Table 5. Demographic information.

Results

In the final sample, 33 children were assigned to Condition 1, and 27 children were assigned to Condition 2. For each question in the test phase, children were forced to choose one of the two options to proceed. Therefore, there were no missing values, except for one question from two children due to technical issues.

Of the 10 test items, item 1 to item 7 tested children's choice of markers when the novel noun stems were in AA format. These were the questions of interest, as adopting Rule (a) over Rule (b) produces different outcomes, because Rule (a) would predict that the pi marker should be selected, while Rule (b) says children should pick the ta marker. On the other hand, item 8 and 9 asked about the markers for AB format novel nouns, so both rules suggest that children would pick the ta marker instead of the pi marker. Finally, item 10 is a check question, asking children to select between the ta marker and a marker that was not in the exposure (ma) for an AB format novel singular noun, and in theory, they should choose the ta marker.

For each participant, the proportion of answering with the ta marker (ta_proportion) in item 1 to 7, and item 8 to 10 were calculated respectively.

Item 1 to 7

First, the main research question – whether the children adopted Rule (a) or Rule (b) – was examined from their answers to Item 1 to 7. Overall, children selected ta as answers to items 1 to 7 for 51.71% of the time (SD = 0.26). As shown in *Figure 5*, a number of children were choosing the ta marker around 50% of the time, which contradicted with the prediction from the Tolerance Principle.

If most children used Rule (a) to generalize, they would pick the pi marker 100% of the time, therefore the expected distribution should look like *Figure 6a*, with more counts at ta_proportion = 0. Statistically speaking, Rule (a) suggested that ta_proportion should

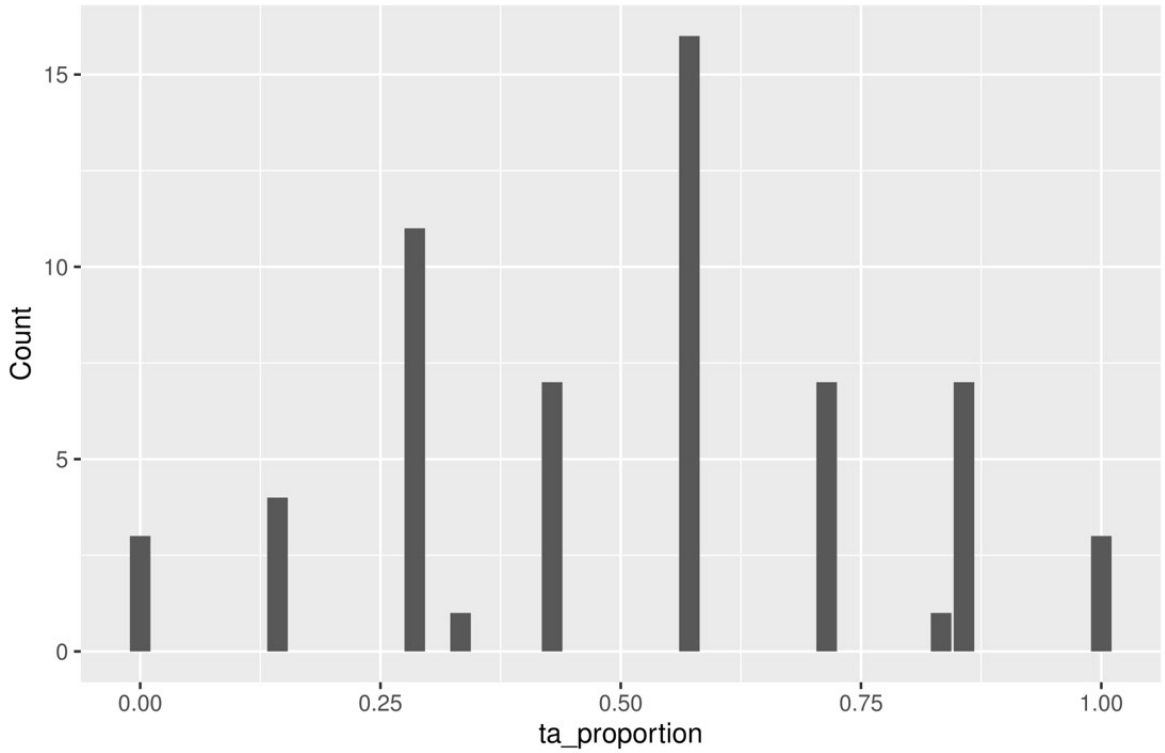


Figure 5. ta_proportion distribution for Item 1-7.

not be significantly different from 0 for item 1 to 7, as children were predicted to consistently choose the pi marker over the ta marker. Since the ta_proportion was not normally distributed, according to the result from the Shapiro-Wilk's normality test ($W = 0.96$, $p = 0.033$), one-sample Wilcoxon tests were adopted to test the hypothesis. The data showed that children's responses were significantly different from 0 ($V = 1653$, $p < 0.001$).

In contrast, if most children chose to use Rule (b), they would select the ta marker 100% of the time, and thus the distribution was expected to look similar to *Figure 6b*, as most children's ta_proportion should be 1. In other words, Rule (b) hypothesized that ta_proportion should not be significantly different from 1 for item 1 to 7, because children would choose ta over pi in all items, but the ta_proportion was also significantly different from 1 ($V = 0$, $p < 0.001$). A follow-up analysis investigated whether the participants' usage of the ta marker was significantly different from choosing at random (0.50), and the result showed no difference from chance ($V = 915$, $p = 1$).

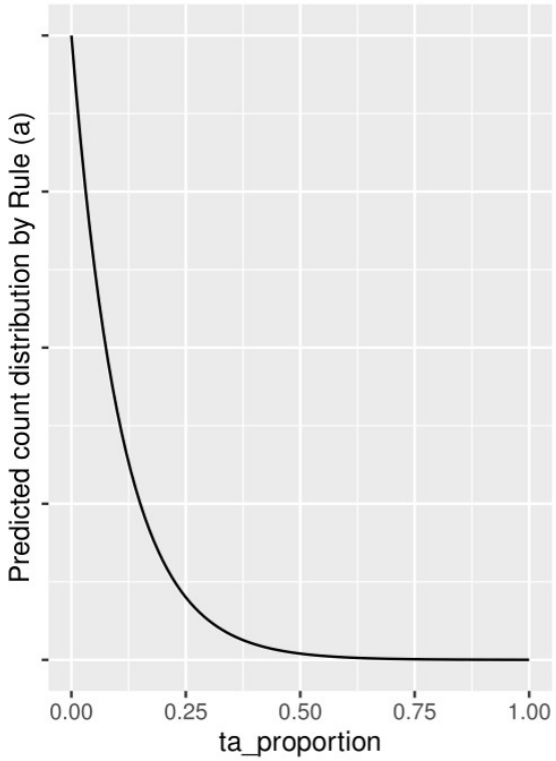


Figure 6a. Expected ta_proportion distribution in Item 1-7 if children use Rule (a).

A possible explanation of these outcomes was that some children adopted Rule (a), while others adopted Rule (b), leading to an average toward 0.50. To test this, the participant responses were divided into two groups, depending on whether ta_proportion was greater or less than 0.50. One-sample Wilcoxon tests were applied on these two groups separately, but the response from the group with the higher ta_proportion ($n = 34$) was still significantly different from 1 ($V = 0$, $p < 0.001$), and the response from the group with the lower ta_proportion ($n = 26$) was significantly different from 0 ($V = 276$, $p < 0.001$).

The next question was: did the children abandon the Tolerance Principle and use probability matching instead? Probability matching means that children produce the same frequency of the ta marker as its frequency in the exposure phase. Of all the exposure trials that contain plural nouns, the ta inflection appears 57% of the time in Condition 1, whereas 74% of the time in Condition 2. Thus, if children adopt the probability matching

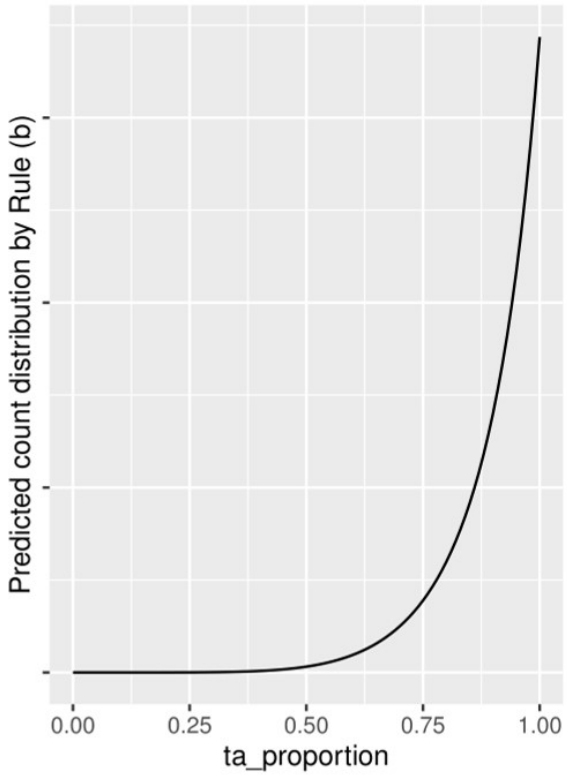


Figure 6b. Expected ta_proportion distribution in Item 1-7 if children use Rule (b).

approach rather than the Tolerance Principle, they should produce the ta marker around 57% of the time if they are in Condition 1, or around 74% if they are in Condition 2. Results from one-sample Wilcoxon tests demonstrated that the percentage usage of the ta marker from children in Condition 1 was not significantly different from 57% ($V = 143$, $p = 0.27$), and the response from children in Condition 2 was significantly different from 74% ($V = 67$, $p = 0.0001$). Though children in Condition 1 responded with the ta marker with a frequency matching the exposure phase, it was likely that they were only choosing by chance (50%), as their ta_proportion was also not significantly different from 0.5 ($V = 153.5$, $p = 0.40$). Similarly, it was also possible that children in Condition 2 were also answering randomly, as their usage of the ta marker was not significantly different from chance either ($V = 329$, $p = 0.39$).

Next, whether age or condition had an effect on children's choice in Item 1 to 7 was

tested. Theoretically, conditions should not influence the ta_proportion, as the Tolerance Principle would predict the same rule even when the token frequencies vary, as long as the type frequencies are equivalent. *Figure 7* explored the relationship between participant age and ta_proportion in different conditions. A linear model was fitted to smooth out the variances, and the colored regions represented the 95% confidence intervals. The figure shows a potential interaction effect, such that older children in Condition 1 tended to use ta less. A multiple linear regression model was used to further investigate the effect of age and condition on the usage of the ta marker ($F(3, 56) = 3.09$, $p = 0.03$, adjusted $R^2 = 0.096$). While age ($t = -0.02$, $p = 0.98$) and condition ($t = 1.776$, $p = 0.08$) did not have significant main effects, their interaction was significant ($t = -2.043$, $p = 0.046$), echoing the finding from *Figure 7* that older children in Condition 1 used fewer ta markers in item 1 to 7.

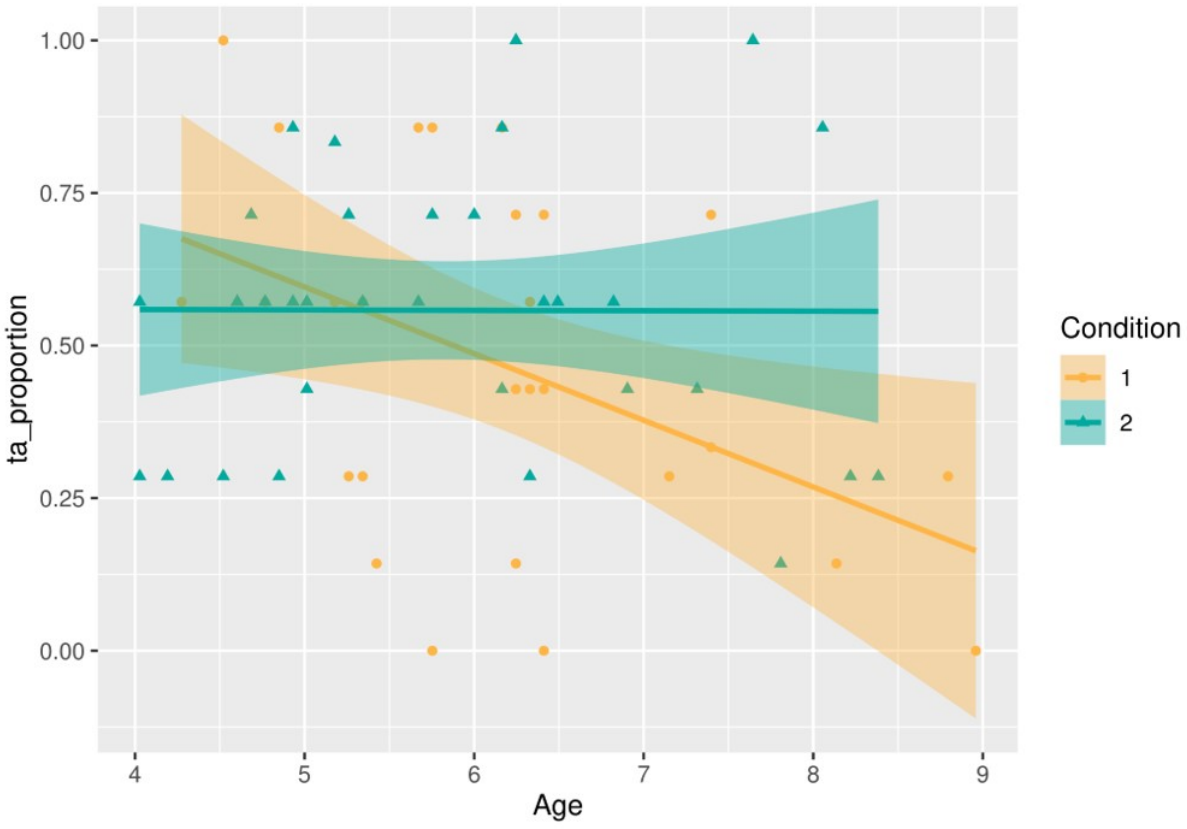


Figure 7. The effect of condition and child age on ta_proportion in Item 1-7.

Child ages were further grouped into four bins: 1 – under 5 years old (n = 15), 2 – 5 - 6 years old (n = 15), 3 – 6 - 7 years old (n = 18), and 4 – over 7 years old (n = 12). *Figure 8* was a boxplot demonstrating the different effects of age and condition on the ta_proportion with respect to conditions. The grey dots represented individual scores from each child, and the red bars were error bars showing ±1 standard error around the mean. If the child answered at random, the ta_proportion should be at 0.50, where the blue dashed line lies. While the variances in different age groups seemed sizable especially in Condition 1, one should keep in mind that the sample size within each bin was fairly small, and therefore the group difference was low in statistical power. A two-way ANOVA proved that neither age group, condition, nor their interaction had a significant effect on the ta_proportion (age group: $F(3, 52) = 1.09$, $p = 0.36$; condition: $F(1,52) = 1.95$, $p = 0.17$; age group * condition: $F(3, 52) = 2.16$, $p = 0.10$). If children were further grouped by whether they were younger or older than 6 years old, the effect of age continued to be insignificant according to the ANOVA test (age group: $F(1,56) = 0.82$, $p = 0.37$; condition: $F(1,56) = 1.83$, $p = 0.18$; age group * condition: $F(3, 56) = 1.28$, $p = 0.26$). The bar graph *Figure 9* also demonstrated that the age and the condition effects were not significant, as the averages were all close to 0.5.

The effect of condition and age group were further assessed by non-parametric tests. The Mann-Whitney U Test showed no significant difference in the ta_proportion between the two conditions ($W = 370$, $p = 0.26$). The median proportion for Condition 1 was 0.43, and the median proportion was 0.57. The Kruskal-Wallis rank sum test was applied to inspect the age group differences, and no significant effect was found ($H(3) = 3.62$, $p = 0.30$). The median proportion for each age group was 0.57, 0.57, 0.57, and 0.29.

Item 8 to 10

If children followed any of the two rules proposed in the hypothesis, they should use the ta marker all the time in Item 8 to 10. Therefore, the predicted ta_proportion was 1.0,

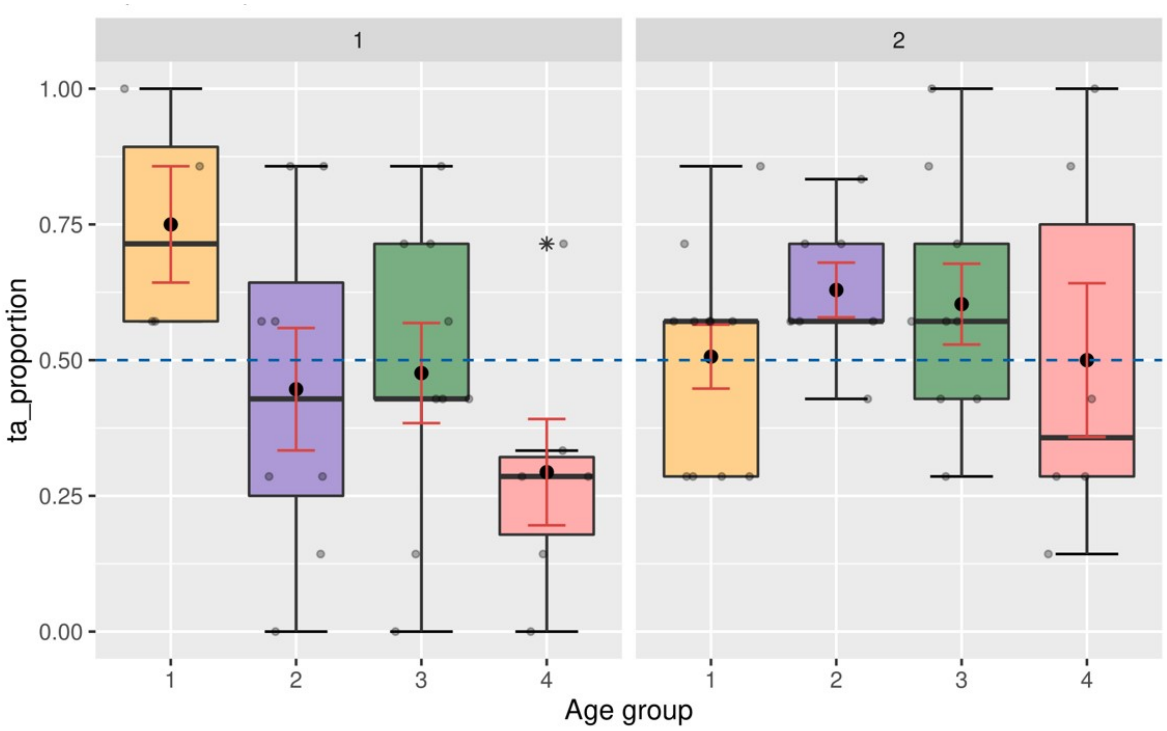


Figure 8. A boxplot of ta_proportion in Item 1-7 grouped by age and condition.

as indicated by the blue dotted line in *Figure 10*. However, only 31 (51.7%), 29 (48.3%), and 38 (63.3%) children answered with the ta marker in these three questions respectively. According to this figure, more children responded with absolute answers (i.e., ta_proportion = 0 or 1) than in the first seven items, likely because fewer items were studied in this section. Regardless, the result from one-sample Wilcoxon test demonstrated a significant difference between ta_proportion and 1 ($V = 0$, $p < 0.001$), showing that children were not using either of the two rules to make generalizations. In addition, one-sample Wilcoxon tests were conducted to see if children were using probability matching in this case, namely approximately 57% of ta_proportion in Condition 1 and 74% of ta_proportion in Condition 2. The findings were similar, such that children's usage of the ta inflection was not significantly different from 57% in Condition 1 ($V = 127$, $p = 0.14$), but it was not significantly different from chance either ($V = 137$, $p = 0.21$). Children in Condition 2 used the ta marker significantly different from 74% ($V = 153$, $p =$

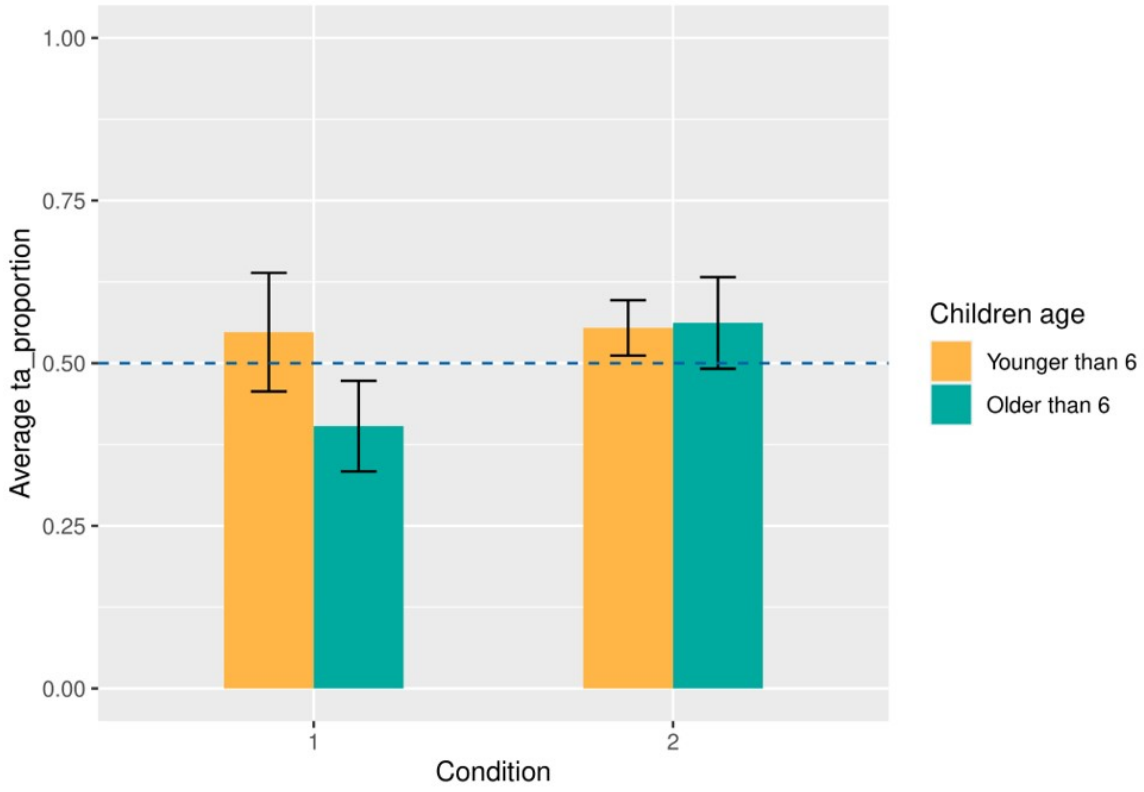


Figure 9. A bar graph of ta_proportion in Item 1-7 grouped by age (young and old) and condition.

0.02). A pairwise Wilcoxon test was conducted to examine whether children treated Item 1 to 7 and Item 8 to 10 differently in terms of ta_proportion, and the result showed no statistically significant difference between children's choices in Item 1 to 7 and Item 8 to 10 ($V = 945.5$, $p = 0.35$).

Revisiting Item 1 to 7 with Selected Children

Because children should choose to use the ta marker in Item 10 no matter if they acquired the proposed rules or not, children who answered this question incorrectly were likely to not attend to the exposure phase enough. Thus, these children were filtered out, and the Wilcoxon tests were applied to the remaining children again to inspect the main

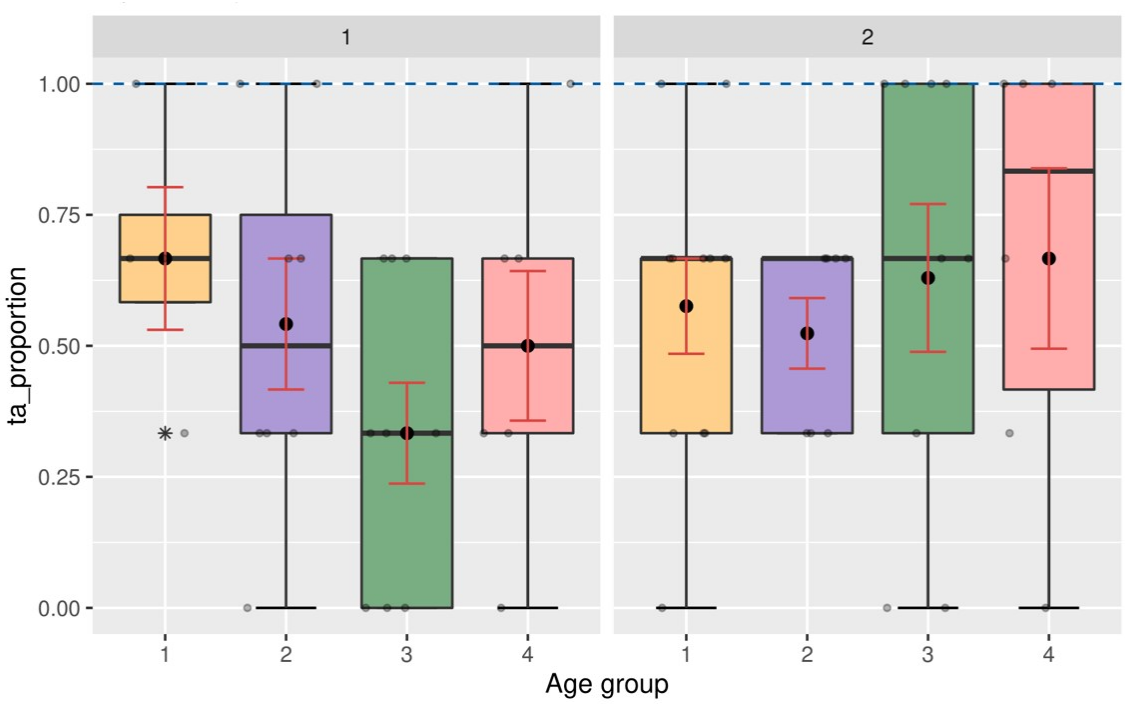


Figure 10. A boxplot of ta_proportion in Item 8-10 grouped by age and condition.

research question. This sample consisted of 38 children, and their percentage usage of the ta marker was compared with 0 and 1. The results indicated that their ta_proportion was still significantly different from 0 ($V = 0$, $p < 0.001$) and 1 ($V = 666$, $p < 0.001$). Additionally, the ta_proportion was not significantly different from 0.5 ($V = 396$, $p = 0.72$), suggesting that those who answered Item 10 correctly might still answer Item 1 to 7 randomly. A set of one-sample Wilcoxon tests were applied again to examine children's performance in Condition 1 and 2 separately. The results showed that the ta_proportion for children in Condition 1 was not significantly different from 0.57 ($V = 72$, $p = 0.85$), but also not significantly different from chance ($V = 74.5$, $p = 0.94$). For participants in Condition 2, their response was significantly different from 0.74 ($V = 22$, $p = 0.001$).

Discussion

Results Interpretation

The experiment was designed to test children's use of the Tolerance Principle when rules of two different scopes are available. Specifically, if the children choose to adopt the Tolerance Principle, there are two possible rules: (a) use the irregular marker for the "irregular-looking" stems, and the regular marker for the others; (b) use the regular marker for all stems. In this experiment, the "irregular-looking" words are those in AA format, and the other stems are in AB format. The irregular inflection is pi, and the regular inflection is ta. Thus, Rule (a) and Rule (b) can be translated as:

- Rule (a): use pi for AA format noun stems, and ta for AB format noun stems;
- Rule (b): use ta for all noun stems.

Additionally, the Tolerance Principle predicts that children should respond in a categorical manner, meaning that they should choose the "correct" marker 100% of the time. Children in Condition 1 and Condition 2, which represent different token frequency distributions, should not differ in their performance, as the Tolerance Principle is not influenced by the token frequency. Alternatively, children could act as adults by applying probability matching. Then, they would choose the ta or pi marker according to their token frequency. On that account, children would choose ta 57% of the time in Condition 1, and 74% in Condition 2.

A third possibility is that children could not learn the patterns in the exposure, and chose the answers by chance. Then, their response would be close to using one of the markers approximately 50% of the time.

There are several signs in the results that suggest that the children leaned toward the third possibility – choosing at random. Overall, for Item 1 to 7, children selected the ta marker versus the pi marker around half of the time. Consequently, their usage of the ta marker was neither close to 0 nor 1, indicating that they were not following either of the

two rules proposed by the Tolerance Principle. A potential explanation explored was that some children used Rule (a), others used Rule (b), and thus when their results were summed together, the average was about 50%. To test this, children were analyzed in two groups by separating those who used the ta marker more than 50% from those who used less than 50%. However, the results again showed that the children did not score on the floor or the ceiling: children did not use the ta marker about 100% when they used the ta marker more than the pi marker, and did not ban the usage of the ta marker when they used the ta marker less than the pi marker.

Children likely did not match with token frequency either. Firstly, the non-parametric test demonstrated no condition effect on children's usage of markers. Additionally, children in Condition 2 selected the ta marker in a way that was significantly different from 74% of the time. Though children in Condition 1 chose the ta marker close to 57%, their response was also not significantly different from chance (50%). Since there was no known reason that one group of children used probability matching while the other group did not, it was more likely that both groups did not use probability matching as a strategy.

Similar results were found for Items 8 to 10. Children were expected always to answer with the ta marker in these questions given the Tolerance Principle, but their usage of the ta marker was significantly different from 1. They did not seem to answer the items matching with token frequency either, because like the findings in Items 1 to 7, children in Condition 1 used the ta marker close to 57% but also close to chance, and those in Condition 2 selected the ta marker significantly differently from token frequency.

These findings were still robust even after removing all children who failed to answer the check question (Item 10) correctly. Like the overall sample, they responded with the ta marker in Item 1 to 7 significantly differently from 0 and 1, implying that the Tolerance Principle was not in place. Probability matching was probably not their strategy either due to similar results as in previous analyses.

The results suggest that children were most likely answering by chance, but an interesting interaction effect between age and condition on the response was spotted. According to *Figure 7* and results from multiple linear regression, older children in Condition 1 tended to use less ta marker and more pi marker. It could be potentially interpreted as a trend in probability matching, as children in Condition 1 were exposed to pi markers more than those in Condition 2. The finding was in line with previous research on children and adults' distinction in learning, such that adults and older children are more inclined to apply probability matching (Kam & Newport, 2005; Newport, 2020; Schuler et al., 2021). However, the effect was probably weak, as the interaction was no longer significant when age was binned and analyzed categorically.

Limitations

There are several pieces of evidence indicating that the experiment was too hard for the participants. For example, almost 25% of the children quit the study halfway. Of those who completed the study, almost 30% of them were excluded from analyses because they exclusively chose the character on the left or the right side of the screen, or preferred the character of one gender over another, demonstrating that they probably had no clue about what the test trials were about. Finally, as discussed above, participants that were kept in analyses still answered the test trials randomly. I have listed the following limitations that may have made the experiment too hard for children to attend to.

First of all, the experiment was conducted online. Online experiments have clear benefits, such as attracting more participants from various locations, and making the experiments more accessible to the public. However, compared to a rigorous lab setting, testing at home without researchers' assistance could lower the quality of the collected data. For instance, many webcam videos have revealed that there were noises or other people around the participants, and most children turned their attention to the surroundings after a few trials. If the experiment was conducted in the lab, the researchers

could redirect children's attention to the stimuli.

Another limitation comes from the study design. In general, artificial language experiments require a large amount of exposure for the participants to acquire the underlying rules. In addition, in order for the Tolerance Principle to work, the frequencies of the tokens have to obey Zipf's law. Thus, with twelve noun stems, it is unavoidable for the exposure phase to be lengthy to such an extent that children failed to follow.

There were deficiencies in the stimuli creation process as well. Occasionally, there were noticeable inconsistencies in audio quality. Also, all of the audio stimuli were recorded by the researcher herself, which could be confusing as aliens, boys, and girls characters all had the same voice.

Aside from experiment difficulty, the sample was also not representative of the general population. In particular, approximately 95% of the participating parents/guardians had a four-year college degree or above, considerably higher than the percentage in the US population (37.52%) (Bureau, 2020). A skewed sample could lead to fallacies in extrapolating the experiment results, though it is difficult to see why children of such a highly educated sample might be more likely to fail.

Future Directions

Given that the experiment was too complicated for the current sample of children, several potential improvements are proposed below.

First, according to adults' feedback, children – particularly on the young end – struggled to concentrate on the study. Researchers may test children older than 4 years old for future experiments on the Tolerance Principle to attain better results.

Secondly, due to limited time, the present experiment was not preceded by a pilot study to probe children's attention span and performance in online experiments. In the future, a simple Tolerance Principle design involving only one productive rule could be

piloted first. After achieving similar results as in in-person studies, researchers can then move on to more complicated research questions like the one investigated in the current study.

Meanwhile, the experimental design can be improved as well. Researchers could make the exposure phase more interactive and entertaining to compensate for the possible interference in online experiments. Example methods include presenting stimuli with simple animations, and asking children to click on certain places to proceed.

On the other hand, recent study by Koulaguina and Shi (2019) has provided an interesting line of research. The token frequency in their experiment did not distribute according to Zipf's law. Instead, out of the 10 rule cases (i.e., number of types is 10), two were repeated 16 times, and eight were repeated 4 times. Nevertheless, their results could still be supported by the equation from the Tolerance Principle. The finding is unanticipated, as the Tolerance Principle is derived under the assumption of the Zipfian distribution in natural language inputs. Possibly, children could tolerate a relatively flexible frequency distribution. Future research can explore the robustness of the Zipfian distribution in empirical child studies. If the cases need not be strictly distributed in a Zipfian manner, the overall number of trials may be greatly reduced, and the exposure phase can be cut short. Then, child participants could complete the study fairly quickly, increasing the likelihood of receiving higher-quality results.

Another topic of interest is the fadeaway of the Tolerance Principle over time. Empirical results have illustrated that contrasted with young children utilizing the Tolerance Principle, adults tend to adopt the probability matching strategy (Newport, 2020; Schuler et al., 2021). Newport (2020) has also found that older children (7-8 years old) are in between, producing answers that are not as categorical as the young children, but also slightly more extreme than the adults. This is in line with the pattern in the present study, in which older children in one of the conditions tend to go with probability

matching more than the others. One account raised in the paper is that children first acquire a clear, definite rule with the Tolerance Principle, and once the rule is formed, probability matching is learned by older children to allow for more flexible language usage. Admittedly, the distinction could be due to adults overthinking in experiments, too. Choosing an answer categorically requires "confidence in conviction," as one has to select one option over the other in every question. Adults who are more used to evenly distributed answers might hesitate to agree to the same answer over and over again. Or, adults could be overly conscious under an experiment setting, and keep track of the token frequency more than they should in real life, causing them to respond with probability matching. These possibilities lead to several insights: first, if future studies aim to explore the "pure" Tolerance Principle effect, they should avoid recruiting children older than 7 to achieve best results. Besides, a production test could probably work better than forced-choice questions, because it allows free response from the children, and therefore is easier to see if they have actually acquired the rule.

In summary, the present study did not succeed in investigating children's use of the Tolerance Principle with more than one productive rule, as the experiment was overwhelmingly difficult for them in an online setting. However, similar to the previous studies, a trend in probability matching was found among the older children. The experiment could give insights into future research in terms of test difficulty, length, and distinction between adults' and children's rule learning.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bureau, U. C. (2020). Educational Attainment in the United States: 2020. *The United States Census Bureau*. Retrieved from https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 23, pp. 378–388).

Carlucci, L., & Case, J. (2013). On the Necessity of U-Shaped Learning. *Topics in Cognitive Science*, *5*(1), 56–88. https://doi.org/10.1111/tops.12002

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable: Export tables to latex or html*. Retrieved from https://CRAN.R-project.org/package=xtable

Elsen, H. (1998). The Acquisition of Past Participles: One or Two Mechanisms? In R. Fabri, A. Ortmann, & T. Parodi (Eds.), *Models of Inflection*. Berlin, Boston: DE GRUYTER. https://doi.org/10.1515/9783110919745.134

Fuson, K. C., Richards, J., & Briars, D. J. (1982). The Acquisition and Elaboration of the Number Word Sequence. In C. J. Brainerd (Ed.), *Children's Logical and Mathematical Cognition: Progress in Cognitive Development Research* (pp. 33–92). New York, NY: Springer. https://doi.org/10.1007/978-1-4613-9466-2_2

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

https://doi.org/10.1016/j.cognition.2005.03.003

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362–366. https://doi.org/10.1016/j.cognition.2010.01.006

Gorman, K., & Yang, C. (2019). When Nobody Wins. In F. Rainer, F. Gardani, W. U. Dressler, & H. C. Luschützky (Eds.), *Competition in Inflection and Word-Formation* (pp. 169–193). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02550-2_7

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186. https://doi.org/10.1016/S1364-6613(00)01467-4

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Henry, L., Wickham, H., & Chang, W. (2020). *Ggstance: Horizontal 'ggplot2' components*. Retrieved from https://CRAN.R-project.org/package=ggstance

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

Hugh-Jones, D. (2021). *Huxtable: Easily create and style tables for latex, html and other formats*. Retrieved from https://CRAN.R-project.org/package=huxtable

Kam, C. L. H., & Newport, E. L. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, *1*(2), 151–195.

Kaplan, D., & Pruim, R. (2021). *Ggformula: Formula interface to the grammar of graphics*. Retrieved from https://CRAN.R-project.org/package=ggformula

Kassambara, A. (2021). *Rstatix: Pipe-friendly framework for basic statistical tests*. Retrieved from https://CRAN.R-project.org/package=rstatix

Kiparsky, R. P. V. (1973). “Elsewhere" in phonology.

Koulaguina, E., & Shi, R. (2019). Rule Generalization from Inconsistent Input in Early Infancy. *Language Acquisition: A Journal of Developmental Linguistics*, *26*(4), 416–435. https://doi.org/10.1080/10489223.2019.1572148

Li, D., Grohe, L., Schulz, P., & Yang, C. (2021). The Distributional Learning of Recursive Structures - lingbuzz/005812. Retrieved from https://ling.auf.net/lingbuzz/005812

Long, J. A. (2020). *Jtools: Analysis and presentation of social scientific data*. Retrieved from https://cran.r-project.org/package=jtools

Lüdecke, D. (2018). Sjmisc: Data and variable transformation functions. *Journal of Open Source Software*, *3*(26), 754. https://doi.org/10.21105/joss.00754

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German Inflection: The Exception That Proves the Rule. *Cognitive Psychology*, *29*(3), 189–256. https://doi.org/10.1006/cogp.1995.1015

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*(11), 465–472.

Miller, K. F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States: Cross-cultural insights into the nature and course of preschool mathematical development. In *Handbook of mathematical cognition* (pp. 163–177). New York, NY, US: Psychology Press.

Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Newport, E. L. (2020). Children and Adults as Language Learners: Rules, Variation, and Maturational Change. *Topics in Cognitive Science*, *12*(1), 153–169. https://doi.org/https://doi.org/10.1111/tops.12416

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Pruim, R., Kaplan, D., & Horton, N. (2021). *MosaicData: Project mosaic data sets*. Retrieved from https://CRAN.R-project.org/package=mosaicData

Pruim, R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic package: Helping students to 'think with data' using r. *The R Journal*, *9*(1), 77–102. Retrieved from https://journal.r-project.org/archive/2017/RJ-2017-024/index.html

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2020). *PsychTools:Tools to accompany the 'psych; package for psychological research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psychTools

Rumelhart, D. E., & McClelland, J. L. (1985). *On learning the past tenses of english verbs.* CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Schuler, K., Yang, C., & Newport, E. (2021). *Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient*. PsyArXiv. https://doi.org/10.31234/osf.io/utgds

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Taatgen, N., & Dijkstra, M. (2003). Constraints on generalization: Why are past-tense irregularization errors so rare? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25).

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A Neural Dissociation within Language: Evidence that the Mental Dictionary Is Part of Declarative Memory, and that Grammatical Rules Are Processed by the Procedural System. *Journal of Cognitive Neuroscience*, *9*(2), 266–276. https://doi.org/10.1162/jocn.1997.9.2.266

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2020a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2020b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wilke, C. O. (2020). *Ggridges: Ridgeline plots in 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggridges

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*(3), 531–556. https://doi.org/10.1017/S0305000900009946

Yang, C. (2016a). The linguistic origin of the next number. Retrieved from https://ling.auf.net/lingbuzz/003824

Yang, C. (2016b). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT Press.

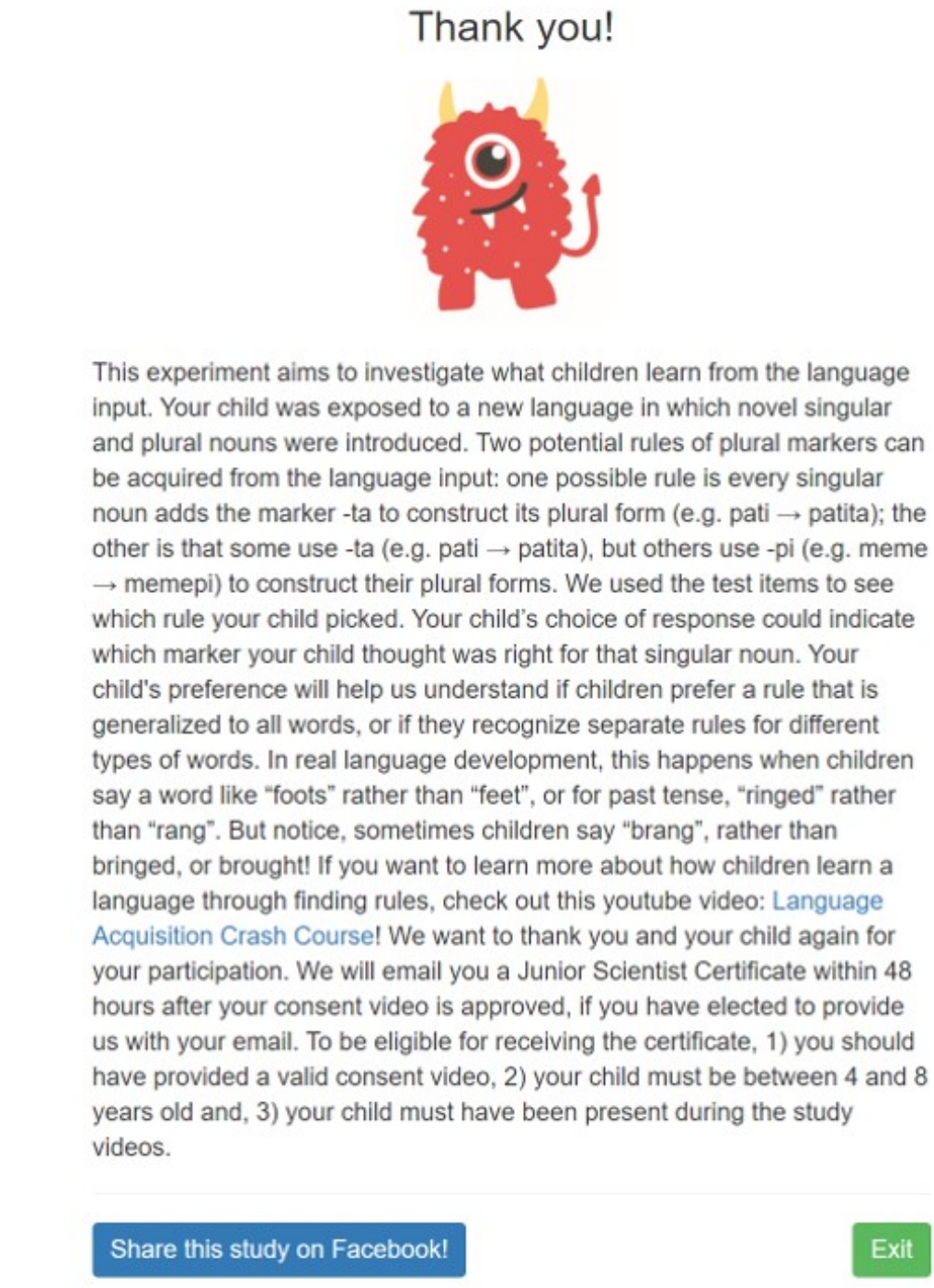
Yang, C., & Montrul, S. (2017). Learning datives: The Tolerance Principle in monolingual and bilingual acquisition. *Second Language Research*, *33*(1), 119–144. Retrieved from https://www.jstor.org/stable/26375874

Zipf, G. K. (2016). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Ravenio Books.

Appendix

Appendix A

Debriefing section that parents or guardians will see when the experiment is over.



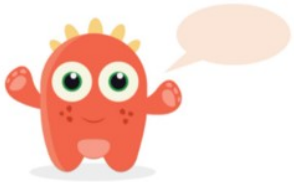
Appendix B

A certificate that the participating child will receive.

Appendix C

The description of the experiment read by the participating family.

Learn an alien language!



Last edited: Apr 03, 2021

Lab: Smith College Lab

In this experiment, your child will watch a group of alien creatures describing objects in an alien language. Then, your child will be asked to answer ten questions to help humans learn the alien language.

Purpose:

Our study investigates how children learn a new mini language from a given set of input. In particular, we want to know when two competing rules are present, which rule the children will pick. Your child's preference will help us understand if children prefer a rule that is generalized to all words, or if they think there are separate rules for different types of words.

Duration: 15 minutes **Exit URL:** https://lookit.mit.edu/studies/history/

Participant eligibility: 4-to-8-year-old children who speak English (Bilingual children are also welcomed to participate as long as one of their native languages is English!) **Compensation:** We will email a Junior Scientist Certificate to families who participate and have elected to provide us with their email. To be eligible for receiving the certificate, please 1) provide a valid consent video (the instructions will be included), 2) your child must be in the age range specified above and, 3) your child must be present during the study videos. There are no additional benefits anticipated.

Minimum age cutoff: 4 years 0 months 1 day **Maximum age cutoff:** 9 years 0 months 2 days

UUID: 4089833d-9102-46bb-a665-51fa0cf41651

Appendix D

Below are several links relevant to the research:

Experiment Stimuli

Experiment Preview on Lookit

Data analysis and results write-up

Appendix E

Below is the list of software and packages used in this research:

R (Version 4.0.3; R Core Team, 2020) and the R-packages *dplyr* (Version 1.0.2; Wickham et al., 2020), *forcats* (Version 0.5.0; Wickham, 2020a), *ggformula* (Version 0.10.1; Kaplan & Pruim, 2021), *ggplot2* (Version 3.3.2; Wickham, 2016), *ggridges* (Version 0.5.2; Wilke, 2020), *ggstance* (Version 0.3.5; Henry et al., 2020), *huxtable* (Version 5.2.0; Hugh-Jones, 2021), *jtools* (Version 2.1.3; Long, 2020), *knitr* (Version 1.30; Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *mosaic* (Version 1.8.3; Pruim, Kaplan, & Horton, 2017, 2021), *mosaicData* (Version 0.20.2; Pruim et al., 2021), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *patchwork* (Version 1.1.1; Pedersen, 2020), *psychTools* (Version 2.1.3; Revelle, 2020), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 1.4.0; Wickham & Hester, 2020), *readxl* (Version 1.3.1; Wickham & Bryan, 2019), *rstatix* (Version 0.7.0; Kassambara, 2021), *sjmisc* (Version 2.8.6; Lüdecke, 2018), *stargazer* (Version 5.2.2; Hlavac, 2018), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.0.4; Müller & Wickham, 2020), *tidyr* (Version 1.1.2; Wickham, 2020b), *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 2019), and *xtable* (Version 1.8.4; Hugh-Jones, 2021; Dahl, Scott, Roosen, Magnusson, & Swinton, 2019)