Adult regularization of inconsistent input depends on pragmatic factors

Amy Perfors
School of Psychology
Level 4, Hughes Building
University of Adelaide
Adelaide, SA 5005, Australia
amy.perfors@adelaide.edu.au

Abstract

In a variety of domains, adults who are given input that is only partially consistent don't discard the inconsistent portion (regularize) but rather maintain the probability of consistent and inconsistent portions in their behavior (probability match). This research investigates the possibility that adults probability match, at least in part, because of two pragmatic assumptions that they bring to the learning problem: (a) that the variation they see is predictable rather than random; and (b) that their goal is to correctly learn that variation. Evidence from two experiments demonstrates that when either assumption is eliminated, people probability match less and therefore regularize more. These results are discussed with respect to age and domain differences in regularization.

Introduction

In a many experimental situations, people given probabilistic input will tend to *probability match*: that is, they respond differentially in a way that is proportional to those probabilities (Herrnstein, 1961, 1970; Baum, 1979; Wearden, 1983; Shanks, Tunney, & McCarthy, 2002; Vulkan, 2000; West & Stanovich, 2003; Ferdinand, Thompson, Kirby, & Smith, 2013). In learning theory, probability matching occurs when they choose a stimulus proportional to the number of times it has been reinforced (e.g., Herrnstein, 1961, 1970; Baum, 1979; Pierce & Epling, 1983; Wearden, 1983). In decision making, probability matching occurs when people are asked to predict the next item in a sequence (e.g., a card drawn from a deck, a flashing light, or marbles drawn from a bag) and they respond by choosing proportionally to the frequency of that item in the past (e.g., Castellan, 1974; Shanks et al., 2002; Vulkan, 2000; West & Stanovich, 2003; Ferdinand et al., 2013). And in language learning, the focus of this paper, probability matching occurs when people are given linguistic input that varies inconsistently (such as an affix that occurs only 60% of the time, for no apparent reason) and they produce that affix proportional to its frequency (e.g., Hudson Kam & Newport, 2005, 2009; Wonnacott, 2011; Perfors, 2012a, 2012b).

Why do people tend to probability match? Answering this question is difficult because under most ways of viewing it, probability matching is surprising and apparently irrational behavior. In learning theory, a person receives more reinforcement if they always choose the more frequently-reinforced stimulus. In decision making, they achieve more successful predictions if they always choose the most frequent item. And in language learning, a person minimizes the burden on the listener as well as the chance of miscommunication if they remove linguistic variability that serves no purpose. So why maintain such variability?

This paper, which focuses on a linguistic context, investigates the possibility that adults probability match at least in part for pragmatic reasons. Those reasons may include multiple factors, but this paper focuses on two possibilities: that (a) people assume that the variation they see is predictable rather than random; and (b) people assume that their goal should be to correctly learn the underlying variation. Two experiments suggest that when either assumption is eliminated, people choose the more frequent form more (i.e., regularize) and therefore probability match less.

Why would either assumption cause probability matching? Let us consider each in turn.

The assumption that variation is predictable

Assuming that variation is predictable corresponds to assuming that there is some real pattern that explains and is associated with the variation. In a linguistic context, this would mean that people think that variations in morphology or lexical choice map onto some phonological, semantic or situational correlate – that a certain morpheme or marker (e.g., an affix) is correlated with some other factor. Put another way, predictable variation is just variation that is conditioned (perhaps probabilistically) on something else, be it a referent, context, phoneme, or word stem. This notion is consistent with the principle of contrast, which holds that variation in (word) forms corresponds to variation in their meaning (Clark, 1988). The idea that variation in language is predictable is a reasonable assumption for other reasons too: truly inconsistent or unpredictable variation in natural language is quite rare (Chambers, Trudgill, & Schilling-Estes, 2003), and languages tend to evolve to eliminate it (Reali & Griffiths, 2009; Smith & Wonnacott, 2010).

But why would the assumption that variation is predictable lead to probability matching in experiments? One might instead expect that it leads to systematicity, since people who believe that it follows some underlying pattern should try to reflect that pattern. However, in typical experiments, the completely inconsistent morphology in the input cross-cuts any patterns that a person might try to find within the individual words, phonemes, stems, or meanings. A learner who fails to find any patterns but still think the variation must be predictable *somehow* would hypothesize that the pattern is something that cross-cuts those things. For instance, one participant in a previous experiment told us that they thought one affix was associated with things that were shiny. They did therefore produce some systematic pattern, but it looked like probability matching on both the local and global level: a metallic cup had a different affix than a ceramic one. Since many participants assign mappings in an idiosyncratic way, and all participants see a randomized assignment of stems and labels to objects in the first place, there is no way to experimentally identify any systematicity people might be imposing. The assumption that variation is predictable would thus lead to behavior that looks to us just like probability matching.

The assumption that the goal is to correctly learn the underlying variation

The second assumption is that the goal of the task is to correctly learn the language. On first glance one might assume instead that the goal of learning a language is to communicate with it. Of course, when a language is not inconsistent, these goals are not in contradiction. There is also substantial evidence that, for adult learners at least, concerns about appearing competent or using the language "correctly" are large drivers of behavior, independent of any actual communicative goals. Many second-language users over-monitor their language use due to "an overconcern with correctness" even though such over-monitoring is linked to poorer overall learning (e.g., Elkind, 1970; Krashen, 1982). A variety of studies provide converging evidence that attention to (and anxiety about) how others will perceive their language competence has significant effects on foreign language acquisition (e.g., Horwitz, Horwitz, & Cope, 1986; Tsui, 1996; Lefkowitz & Hedgcock, 2002).

If people are at least in part motivated by trying to correctly use the language, this implies that they should endeavor to reproduce the variation they observe in that language. They should do this *even if they aren't sure what causes the variation.* As in the "shiny things" anecdote above, this might be because they formed a hypothesis that explains the variation and then produce language according to that hypothesis. But even if they have no idea what causes the variation, as long as they have noticed that the variation is there, reproducing it serves as a signal to their interlocutor (or the experimenter) that they *have* noticed and are tracking it. In that sense reproducing the variation is more "correct" than ignoring or regularizing it would be.

Bringing it all together

The research just reviewed suggests that both of the two assumptions may apply to real-life language learning in adults. It is possible that the setup of typical experiments can reinforce both assumptions even further. In these experiments the adult participants realize they are doing an experiment and that it involves learning a language that was designed by the experimenter. Stimuli are clearly artificial, often presented in such a way – on a computer or video, or coming from a trained experimenter – that makes it unlikely that the words and labels were accidental or full of errors. People in experiments often believe that they are evaluated based on their ability to "correctly" learn or do the task. Overall, the entire situation may carry the strong implication that there is some pattern or regularity for people to learn, and that the experimenters are interested in how well they are learning it. All of these factors could act to create a sense to participants that the variation is predictable, that it is their job to learn that variation, and that their goal is to demonstrate that they have managed to learn it.

This paper investigates the possibility that adults probability match because they hold these two assumptions: first, that variation is predictable, and second, that their goal is to correctly learn the language. This is investigated through experiments with demand characteristics which relax each of these assumptions in turn. The results indicate that when either assumption is eliminated, adults regularize significantly more.

Experiment A: Eliminating the assumption that variation is predictable

This experiment aims to remove the demand characteristic that implies that variation is predictable. Participants in one condition were told that their input came from the previous person, who was operating under time pressure and may have made some errors. To make this manipulation believable, the specific affixes were changed so that they appear more similar to typos. As a control, performance is compared to a situation where the affixes are the same but people are not given any reason to think they are errors (i.e., the instructions are neutral). A final control compares performance to a more standard situation with neutral instructions and more distinguishable affixes. Results show that regardless of the nature of the affix, people regularize more if they think the variation resulted from a previous participant's error and thus have no reason to think it is predictable.

Participants

152 adults were recruited via Amazon Mechanical Turk, an online resource increasingly used and validated for experiments in psychology and linguistics (Sprouse, 2011; Crump, McDonnell, & Gureckis, 2013). Participants were randomly allocated to one of five conditions, described below: STANDARD (29 people), PARTICIPANTSIMILAR (28), PARTICIPANTTYPOS (33), EXPERIMENTERSIMILAR (30) and EXPERIMENTERTYPOS (32). Ages ranged from 18 to 65 (mean: 32.8). Eight of the participants were from India, two were from Canada, and the remaining 142 were from the United States. All participants were paid $1.50US for the 15-20 minute experiment.

Procedure

The standard task, which was the same in all conditions, involved a word learning task originally modelled after Hudson Kam and Newport (2009) and Perfors (2012a, 2012b). The original Hudson Kam and Newport (2009) involved an artificial language taught over multiple days, but the key element for our purposes was that in this language, units (which they called determiners) covaried with the nouns in an inconsistent fashion: participants heard the main determiner only 60% of the time. Participants were asked to provide the noun and determiner associated with a scene and sentence and the frequency with which each determiner was produced after each noun was noted.

As in Perfors (2012a) and Perfors (2012b), extraneous elements of the task were removed so as to focus on the aspect involved in producing the inconsistent units. This language consists of words composed from 10 stems, all one-syllable consonant-vowel-consonant nonsense terms mapped to images representing common categories of item.[1] Each stem was followed by a one-syllable affix: the **main** affix occurred 60% of the time, and each of the four **noise** affixes occurred 10% of the time. Importantly, in all conditions there was no consistency to the variation: each participant saw a different random assignment of affixes to stems and stems to categories, and the distribution of affixes to stems exactly reflected the global distribution of affixes in the language: each stem occurred with the **main** affix 60% of the time and the **noise** affixes 10%.

[1]Stems were: DUT, SIL, ZEG, MAB, YOK, PIM, REN, JAF, WUX, and COV. Items used were: babies, balls, beds, birds, books, cars, cats, cups, dogs, and shoes.

The task consisted of a total of 200 trials of image-label pairs, with each of the ten categories thus appearing 20 times throughout training. On each trial, an exemplar image drawn from a category appeared on the computer screen and at the same time the person saw a label written in all capitals below it. Labels were presented with no space between the stem and the affix: thus, participants would see words like PIMUT or JAFIG underneath the picture. People went to the next trial by clicking a next button. Learning was tested with ten questions (one for each of the objects) every 50 trials, for a total of 40 test questions. At each test, the participant was presented with a completely novel exemplar image and asked to enter the label for it. No feedback was given.

Conditions

The goal of this experiment was to explore the possibility that adult regularization can be affected by removing the cues suggesting that the variation was predictable. Those cues were manipulated by changing the cover story of how the data originated. In order to make the cover story believable, the nature of the affixes was also varied.

The cover story and affixes in the STANDARD condition were modelled after other studies, with the goal of serving as a control and replication (Perfors, 2012a). The affixes were ON, EP, AD, IG, and UT, and which affix was shown 60% of the time (the **main** affix) was randomized for each participant. People were told that we were studying how people learn new words, and that they would be presented with a series of pictures of common objects (balls, dogs, etc) with labels from an artificial (non-English) language. They were asked not to write any labels down, since we were interested in how people learn by relying on their memory. Appendix A contains the exact instructions in all conditions.

The remaining four conditions amount to a 2x2 design that systematically manipulates two different factors: first, the cover story of how the data originated, and second, the nature of the affixes seen. The first factor is the most important, since it is the cover story that suggests whether the variation was predictable. In the EXPERIMENTER condition the cover story was the same as in the STANDARD condition, while in the PARTICIPANT condition people were given the two following additional sentences: "The labels actually come from a previous participant, who had to learn the fake language themselves. Some participants were given a very limited time to provide the labels so there might be errors; we're interested in how you learn the labels even in that case." The goal was that in the PARTICIPANT condition people would have been given a reason to conclude that the variation they saw might have been random or unpredictable; our question is whether this causes them to regularize more.

The second factor, varying the nature of the affixes, is important only by virtue of the fact that it was necessary in order to make the cover story believable. This is because the affixes in the STANDARD condition do not look very much like the natural sort of errors that a human would make. Thus, people in the TYPOS condition saw affixes designed to look like typos (MacNeilage, 1964; Damerau & Mays, 1989). As a control, people in the SIMILAR condition saw affixes that were more similar than in the STANDARD condition but not necessarily so they looked like typos. Table 1 shows all of the affixes in all conditions. The hypotheses predict that the surface appearance of the affix should not drive differences in regularization since it does not affect beliefs about the predictability of the variation.

Main	Noise (TYPOS)				Noise (SIMILAR)			
ON	PN	OB	N	OON	AN	OV	O	ONN
EP	EO	RP	P	EEP	EW	UP	E	EPP
AD	AF	SD	D	ADD	AL	ID	A	AAD
IG	KG	IF	G	IIG	IZ	AG	I	IGG
UT	UR	YT	T	UTT	UM	ET	U	UUT

Table 1: Affixes in each of the conditions in Experiment A. For each participant in each condition, one of the five possible **main** affixes (which occurred 60% of the time) was chosen at random. In the STANDARD condition, the four **noise** affixes (each occurring 10% of the time) were chosen from the other affixes in the **main** column. For participants who saw the TYPOS affixes, each **main** affix was associated with a different four **noise** affixes, shown on the same line, which occurred 10% of the time each: thus, for instance, one participant might have had ON as the **main** affix and PN, OB, N, and OON as the four **noise** affixes. When designed as typos, the **noise** affixes were mutations of the **main** one consistent with having been produced as a typo; when designed simply to look similar the mutations altered the affixes to the same extent, but were less consistent with having occurred as a typo (e.g., because they involved letters far from each other on the keyboard).

Results of Experiment A

The main question of interest is how much participants in each condition regularized by producing any single affix more than 60% of the time. However, it is first necessary to determine whether any differences in regularization are also associated with differences in overall performance or attention between conditions.

Let us therefore begin by examining the accuracy of stem learning within each condition. There are two ways of calculating accuracy. The most straightforward is simply to count a stem as correct if it is identical to the correct stem. As Figure 1 shows, people in all conditions were fairly accurate according to this definition (since answers were free response and there were 10 stems to be learned, chance performance is far below 50%). There was no significant difference in accuracy between conditions (Kruskal-Wallis: $\chi^2(4) = 5.19, p = 0.269$).[2] The second way of calculating accuracy is more graded, and involves computing the Levenshtein distance between the target stem and the produced stem; this measure captures the number of edits required to turn one into the other (Levenshtein, 1966). There are no significant differences in Levenshtein distance[3] by condition either (Kruskal-Wallis: $\chi^2(4) = 5.82, p = 0.213$). Together, these results suggest that any differences in regularization did not result from participants in different conditions paying different amounts of attention or finding the task more or less difficult.

The primary question is whether people show different levels of regularization in each condition. To evaluate this, following Hudson Kam and Newport (2009) and Perfors (2012b), all participants who did not get at least 9 out of the final 20 stems correct on

[2]Normality assumptions for the accuracy scores were violated in all conditions in which the parametric Kruskal-Wallis test and post-hoc Wilcoxon tests were applied.

[3]Complete descriptive statistics for all measures can be found in Appendix C.

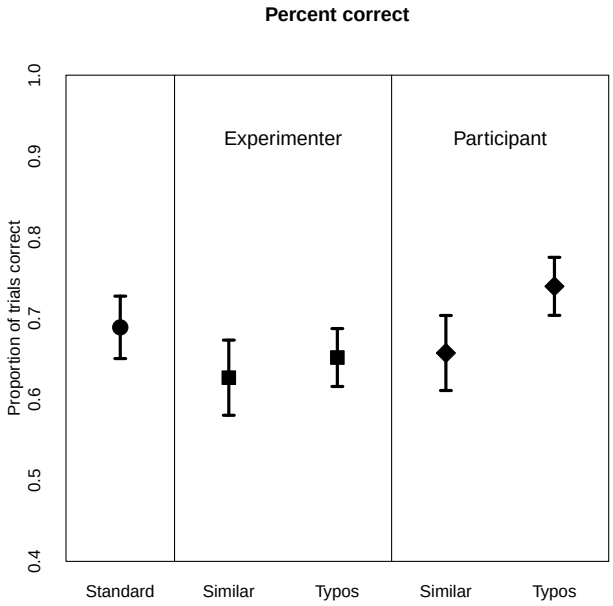

Figure 1. Differences in accuracy by condition. There was no significant difference between conditions in the total proportion of correct stems labeled. Error bars show standard error.

the test trials were excluded.[4] In keeping with the finding that accuracy did not differ by condition, after the exclusion there was no significant difference in the number of people in each condition ($\chi^2(4) = 1.37, p = 0.849$).

Regularization was evaluated by assigning each participant a **regularization index**, defined as the proportion of trials on which that person produced their most frequent affix (out of every trial for which a correct stem was produced). This index is higher for those participants who regularize more. Figure 2 shows the amount of regularization in each condition, revealing that participants in different conditions regularized to different extents (Kruskal-Wallis: $\chi^2(4) = 12.13, p = 0.016$). Post-hoc pairwise tests[5] showed that the difference between the EXPERIMENTER and PARTICIPANT conditions was significant. This suggests that, as predicted, regularization was affected by people's inferences about whether the variation might be predictable (Wilcoxon: $W = 1085, p = 0.016$). The difference between the SIMILAR and TYPOS conditions was not significant, which implies that regularization was less affected by the nature of the affixes (Wilcoxon: $W = 1559, p = 0.592$)

[4]The analyses were conducted without this exclusion. In order to identify any learning effects, all analyses were also repeated for just the second half of trials. In all cases the results were qualitatively identical.

[5]Two pairwise tests (rather than a full 2x2 ANOVA on the four central conditions) were performed for two reasons. First, those two tests correspond to the specific hypotheses this paper focuses on: whether people's regularization was influenced by their perception of whether variation was predictable (i.e., EXPERIMENTER vs PARTICIPANT) or whether it depended on the nature of the affixes (SIMILAR vs TYPOS). Second, the non-normality of the data meant that an ANOVA was inappropriate, and there is no equivalent non-parametric test for 2x2 data. That said, in keeping with the pairwise tests, that ANOVA revealed a significant effect of how the data originated (i.e., EXPERIMENTER or PARTICIPANT) but no effect of the nature of the affix (i.e., SIMILAR vs TYPOS) and no interaction (data origin: $F(2) = 5.79, p = 0.004$; nature of the affix: $F(1) = 0.473, p = 0.493$; interaction: F(1)=0.58, p=0.448).

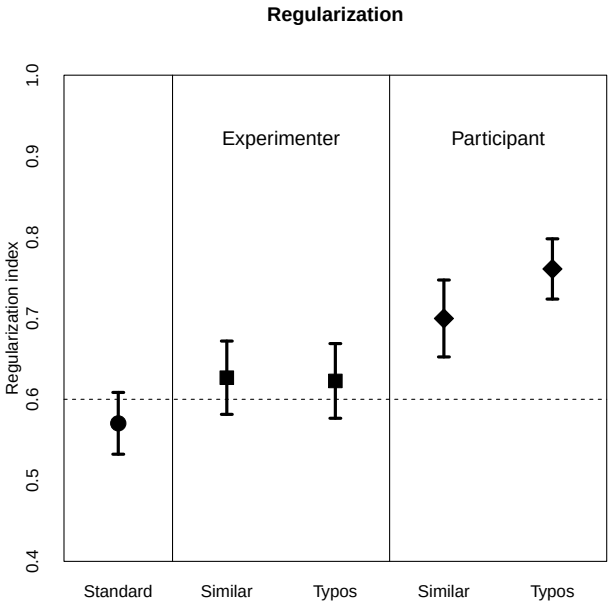

Figure 2. Differences in regularization by condition. The regularization index reflects the proportion of valid trials on which the most frequent affix was produced: an index above the dotted line, which shows the proportion of the **main** affix in the input, indicates regularization. Overall, there was a significant effect of condition on regularization. Error bars show standard error.

Thus far this analysis has just examined regularization on a global level – regularization across all different stems at once. It is possible that lexically-specific regularization might show different patterns. This possibility is inspired by Smith and Wonnacott (2010) and Wonnacott (2011), which found that people may sometimes impose regularization on a lexical level even if the global statistics remain unchanged. Smith and Wonnacott (2010) tested this by calculating the average conditional entropy given each stem: lower absolute conditional entropy reflects more regularization. The conditional entropy measure is very noisy on this dataset, since each stem occurs a maximum of four times (fewer if the participant didn't label each object with the correct stem each time). However, it may be at least suggestive. Results, shown in Figure 3, reveal a significant difference in entropy between conditions (Kruskal-Wallis: $\chi^2(4) = 9.77, p = 0.045$). Post-hoc pairwise tests reveal no significant difference among either pair, although the trends are in the right direction; this may be because of the noisiness of the measure (Wilcoxon: EXPERIMENTER VS PARTICIPANT: $W = 1771, p = 0.083$; SIMILAR VS TYPOS: $W = 1576, p = 0.520$). Overall, this evidence weakly suggests that when people believe that the variation is truly inconsistent, they regularize more on the local as well as global level.

One final possibility is that because these analyses are of mean performance, they may be hiding individual regularization patterns in different directions. To evaluate this possibility, following Hudson Kam and Newport (2009) and Perfors (2012b) participants were counted as **regularizers** if they produced their most frequent affix on at least 90% of the valid trials. There was a significant difference between the EXPERIMENTER, PARTICIPANT, and STANDARD conditions ($\chi^2(2) = 6.04, p = 0.048$), but not when comparing

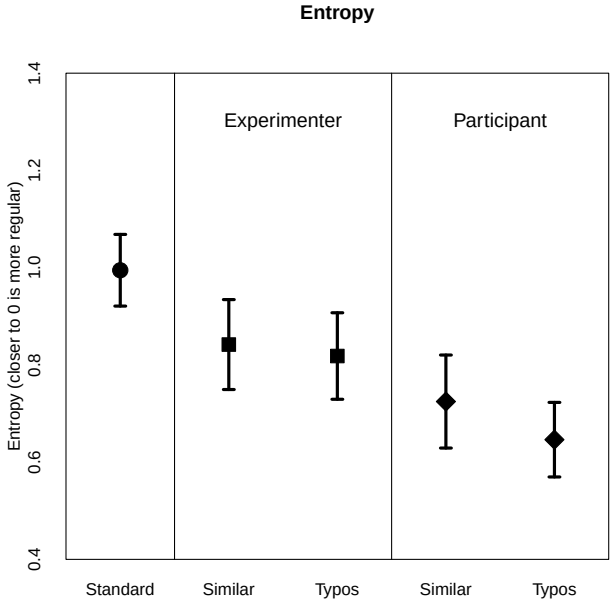

Figure 3. Differences in entropy by condition. Entropy reflects regularization on a lexical rather than global level, with values closer to zero indicating greater regularization. Overall, there was a significant effect of condition on conditional entropy. Large black dots show condition means and error bars show standard error.

the SIMILAR, TYPOS, and STANDARD conditions ($\chi^2(2) = 4.71, p = 0.095$). This effect was even stronger with an 80% threshold (EXPERIMENTER vs PARTICIPANT vs STANDARD: $\chi^2(2) = 9.75, p = 0.008$; SIMILAR vs TYPOS vs STANDARD: $\chi^2(2) = 4.61, p = 0.100$). This analysis supports the idea that differences in regularization were driven more by the possibility that the data originated in such a way that the variation was unpredictable.

Experiment B: Removing the assumption that the goal is to correctly learn the underlying variation

The previous experiment addressed the first assumption – that the variation participants see is predictable. Experiment B investigates the second by removing the demand characteristic that implies that people's goal should be to correctly reproduce the language. Participants were brought into the lab in pairs and told that the pay of each is dependant not on how "correct" they are according to some objective standard, but rather on how closely they matched the other's labels. This is a sensible alternative goal for a language task, since coordination is one of the principal goals of a communicative encounter. As a control, other pairs of people were told that their pay depended on the extent to which one partner could infer the correct meaning when given the other partner's labels. Consistent with the hypothesis of this paper, the results show that if people are pressured to match each other their regularization increases, but if they are pressured to be accurate it decreases.

Participants

52 adults (13 pairs in each condition) were recruited from the University of Adelaide and surrounding community and were paid $10 for their time. Once in the lab, participants were divided randomly into one of two conditions, CORRECTGOAL and MATCHGOAL, described in more detail below. One person in the CORRECTGOAL condition had a computer error causing a failure to save data, leaving 25 people in that condition and 26 in the other.

Procedure

The main task was very similar to Experiment A. Participants were shown images from ten different categories randomly paired with one of ten stems (the same stems as in Experiment A). The same affixes as in the STANDARD condition were used, with one exception: due to a computer error, some participants saw the affix RY instead of one of the other affixes. This occurred equally often in both conditions so is unlikely to drive any differences between condition. As in Experiment A, one of the affixes was chosen at random to be the **main** affix (occurring 60% of the time) and the other four occurred 10% of the time each. There was no consistency to the variation and the mapping of stems to objects was randomized between participants. Training lasted for 200 trials and there were 40 test questions, this time occurring in two blocks of 20 after every 100 trials. At each test, the person was shown a completely novel exemplar image and asked to enter the label for it. No feedback was given.

Conditions

The goal of this experiment was to investigate whether regularization can be affected by explicitly changing the goals of the task. Each pair of participants was therefore up into one of two conditions defined by different task goals.

CORRECTGOAL: This condition increased the pressure to be correct by pairing each person with another participant person who was in the lab at the same time. Both people were informed that the goal of this experiment was to learn a new language and then successfully use it to communicate with the other person. They were asked to imagine they were scientists who had just discovered a community speaking this language, and they had gotten an informant to label a series of pictures for them. They were to learn these labels, and then they would be tested on how well they had learned them based on their responses to new pictures. Full instructions for both conditions are in Appendix B.

Participants sat at different computers and did the standard task individually. However, at the end of the standard task each person was given the labels the other person generated during their test questions, and asked to match each of those labels with the correct image. People were told in advance that this would happen, and that they would get paid proportionally to how many of this final set of questions both of them got right. This created a strong social pressure to learn the language correctly, since not only was each individual's payment dependent on it, so was their partner's.

It is important to note that this manipulation in itself does not favor either regularization or probability matching. Since the affixes did not correlate to the images at all, people could get 100% correct on the test regardless of what they did with the affixes, as long as the stems were matched to the correct image. Thus, any effect on regularization is due to

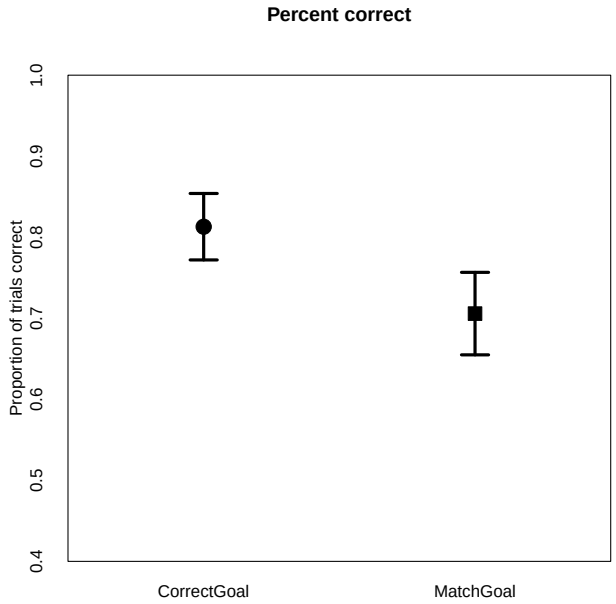

Figure 4. Differences in accuracy by condition. There was no significant difference between conditions in the total proportion of correct stems labeled. Error bars show standard error.

increasing the sense that there is a "right answer" along with the social pressure to find that right answer – not because participants could make more money using one strategy.

MatchGoal: The aim of this condition was to decrease the pressure to be correct. This was also accomplished by placing people in pairs with another participant. This time, however, they were told that both they and their partner would be given the same images during test. Their payment would be proportional to the number of images they produced the same label for. The experiment was otherwise identical to the correctGoal condition, with the same instructions and the same procedure. As in the other condition, participants were prevented from talking to each other or otherwise communicating during the task itself by sitting in different computers in adjoining rooms.

This manipulation removes the pressure to use the "correct" labels – indeed, in theory people could achieve a 100% match if both simply called all of the objects the same random word. Since people were prevented from communicating during the experiment, agreeing on a code beforehand was impossible. They instead had to think about how their partner would tend to respond and try to choose that label. Regularizing the **main** determiner is an easy way to maximize payment in this condition, but if both participants used variable endings in the same way they could also maximize the money earned. Regardless, this condition changes the typical task demands away from trying to be as correct as possible.

Results of Experiment B

As before, the first question is whether there are differences in accuracy between conditions. As Figure 4 shows, when accuracy is calculated according to whether the stem is identical to the correct stem, people were reasonably accurate, but the difference between

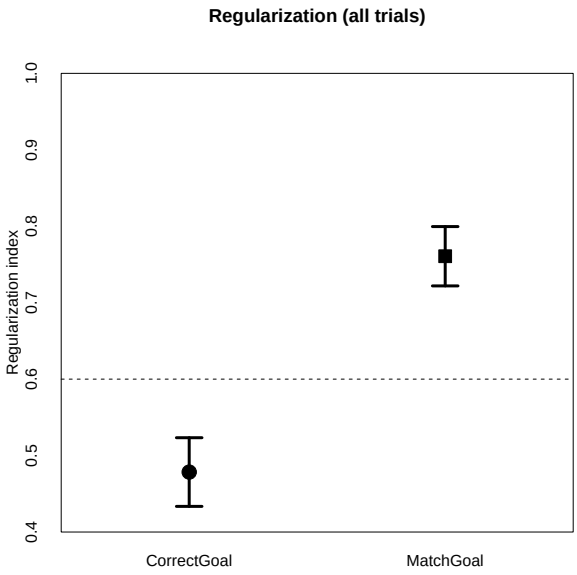

Figure 5. Differences in regularization by condition. The regularization index reflects the proportion of valid trials on which the most frequent affix was produced: an index above the dotted line, which shows the proportion of the **main** affix in the input, indicates regularization. People regularized significantly more in the MATCHGOAL condition. Error bars show standard error.

conditions was not significant ($\chi^2(1) = 3.20, p = 0.073$). These results are the same when accuracy is calculated based on Levenshtein distance ($\chi^2(1) = 3.73, p = 0.054$). These differences are close to statistically significant, but this is not surprising given that the effort participants were meant to put toward being correct was manipulated between conditions.

Do people show different levels of regularization in each condition? In order to investigate this the lowest-performing participants were excluded, as in Experiment A and previous research. This results in 24 people in the CORRECTGOAL condition and 22 in the MATCHGOAL condition.[6] As Figure 5 shows, the degree of regularization was strongly affected by the goal of the participant (Kruskal-Wallis: $\chi^2(1) = 13.9, p < 0.001$).

What about lexically-specific regularization? This is explored by calculating the average conditional entropy in each condition. Results, shown in Figure 6, reveal a significant difference in entropy according to goal condition (Kruskal-Wallis: $\chi^2(1) = 10.66, p = 0.001$). This evidence indicates that on both the global and local levels, people regularized more when their goal was less about being correct and more about trying to match what the other participant was likely to produce.

To explore the possibility that these analyses of mean performance are hiding individual regularization patterns, participants were classified as **regularizers**, as in Experiment A, if they produced their most frequent affix in at least 90% of the trials or more. Over all trials, the difference between conditions was not significant ($\chi^2(1) = 2.04, p = 0.153$), but if the threshold is set to 80% rather than 90%, large differences between conditions emerge ($\chi^2(1) = 7.42, p = 0.006$). This suggests that, although people may not be regularizing *all*

[6] As in Experiment A, all analyses were qualitatively identical when the exclusion criteria were dropped or were restricted to the second half of test trials.

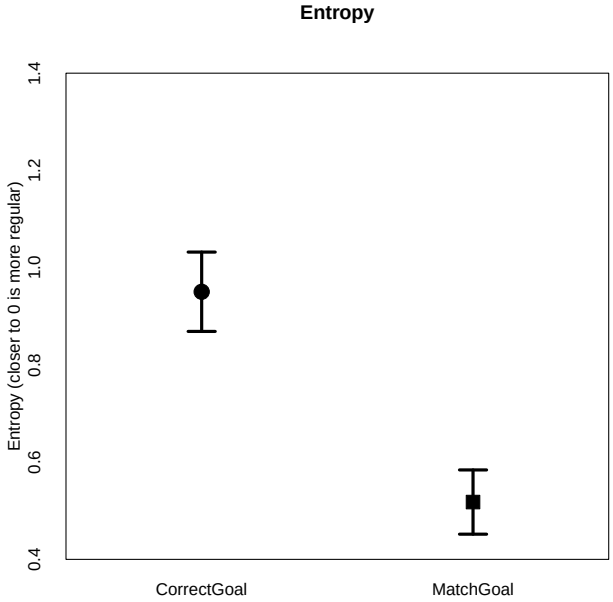

Figure 6. Differences in entropy by condition. Entropy reflects regularization on a lexical rather than global level, with values closer to zero indicating greater regularization. Overall, there was a significant effect of condition on conditional entropy. Large black dots show condition means and error bars show standard error.

of their productions, they are still doing so more often on an individual level when the goal of the task emphasizes coordinating with their partner rather than being correct.

Discussion

This experiment tested the hypothesis that probability matching behavior in adults is driven, at least in part, by pragmatic factors – in particular, two specific assumptions adults may be bringing to the language learning task (both in real life and in these experiments).

One assumption is that the variation in the language is predictable. This is a reasonable assumption in real life: languages tend to evolve to eliminate unpredictable or random variation, rendering it quite rare (Chambers et al., 2003; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). The experimental context may also reinforce that assumption: many aspects of the situation, from the artificially designed stimuli to the presence of an experimenter evaluating performance, carry with them the implication that there is something predictable to be learned. These results show that when this assumption is eliminated by telling people instead that their input came from a (possibly erroneous) previous participant, they regularize significantly more.

The other assumption people may have is that their goal is to correctly learn the underlying language, and to demonstrate that correctness by reproducing it accurately. This also is a reasonable assumption in real life: since variation in language tends to be predictable, accurately learning and using that variation improves communicative success. Moreover, there is consistent evidence that adult language learners are motivated by the desire to be correct and not look foolish to others, often experiencing significant anxiety

around the issue (Krashen, 1982; Horwitz et al., 1986; Tsui, 1996). This research shows that when the pressure to be correct is eliminated by telling participants to try to produce the same labels as each other, regularization increases. This is also consistent with other work finding that when people are allowed to coordinate with each other rather than achieve some objectively correct standard, regularization also increases (Kamps, Ferdinand, & Kirby, 2014). Of course, it is also possible that the increase in regularization in the matchGoal condition was due to the desire to coordinate rather than the elimination of the desire to be correct; this confound was impossible to fully eliminate without replacing it with something else since people found the concept of a purposeless experiment to be nonsensical. That said, the coordination explanation doesn't capture why regularization went *down* in the correctGoal condition, nor would it change the fundamental point that pragmatic pressures affect regularization behavior.

The main motivation of this research is to understand why adults would probability match, since it otherwise seems an inexplicable behavior. Removing linguistic variability that serves no purpose minimizes the burden on both speaker and listener; indeed, this is probably why languages tend to evolve to eliminate it (Smith & Wonnacott, 2010). These results here suggest the people maintain this variability not because (or at least not only because) of cognitive factors, but because of pragmatic assumptions they bring to these experimental contexts as well as, probably, real-life language learning.

Do these results also explain why people might be probability matching even in a non-linguistic context? It is always dangerous to speculate without data, and the final word requires replicating these manipulations in a reinforcement-learning or decision-making paradigm. That said, it seems likely to us that one or both assumptions and many of the same demand characteristics probably apply in those experiments as well. In them, people are also given many cues (both implicit and explicit) suggesting that there is some pattern to learn and that the experimenters are interested in how well they learn it. Just the task of predicting flashing lights or cards in a deck implies that there is something to predict, some underlying pattern to pick up on. Indeed, in these domains it has already been suggested that participants probability match because they think the variation can be predicted based on some underlying factor (Restle, 1961; Koehler & James, 2003). As Restle (1961) stated, "The subject seems to think that he is responding to patterns... even if he is told the sequence is random he does not understand this information clearly, nor is there any strong reason to believe it." Indeed, people can be persuaded to stop probability matching if they are paid for successful performance or given explicit feedback comparing their performance to an optimal responder (see Vulkan, 2000; Shanks et al., 2002, for an overview).

Some of the interest in regularization, particularly in a linguistic context, comes from evidence suggesting that young children may be more likely to regularize by producing or responding to the more frequent item closer to 100% of the time (Hudson Kam & Newport, 2005, 2009). What do these results imply about child-adult differences in regularization?

First, the evidence that children regularize at all is based on relatively few published studies within language (Singleton & Newport, 2004; Hudson Kam & Newport, 2005, 2009), and exceptions have been found (Wonnacott, 2011). Even the evidence of child regularization in non-linguistic contexts is less compelling than it may appear at first glance. Most importantly there are multiple studies showing children of all ages *failing* to regularize, from infants (Davis, Newport, & Aslin, 2011) to preschoolers and kindergarteners (Craig

& Myers, 1963; Messick & Solley, 1957; Offenbach, 1964; Stevenson & Odom, 1964) to older children (Atkinson, Sommer, & Sterman, 1960; Craig & Myers, 1963; Messick & Solley, 1957; Offenbach, 1964; Stevenson & Odom, 1964). There are also studies that show adults regularizing more than older children and often as much or more than young children (Moran & McCullers, 1979; Stevenson & Hoving, 1964; Weir, 1964). Only a few studies find evidence of regularization only in young children, and in those, the regularizers were generally five or under (Derks & Paclisanu, 1967; Jones & Liverant, 1960; Stevenson & Weir, 1959; Bever, 1982). This is younger than the five-to-seven year olds in the Hudson Kam and Newport (2005, 2009) studies. Within the studies that find regularization in young children, it is not always a large effect: for instance, Jones and Liverant (1960) found that the group that regularized most (at nursery age) had only 13 of 40 regularizers, compared to 5 of 40 older children. Given this, we should consider the possibility that there are no robust differences in regularization between children and adults, and hence nothing to explain.

Nevertheless, let us suppose that the child-adult regularization difference is real. The most common hypothesis about what that factor might be relates to possible cognitive changes that occur between childhood and adulthood like differences in metacognitive control (Jones & Liverant, 1960; Ramscar & Gitcho, 2007) or memory (Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009). However, these accounts are hard to reconcile with other research indicating that limitations in these abilities do not necessarily lead to more regularization (West & Stanovich, 2003; Perfors, 2011, 2012b). It is possible that memory and/or other cognitive factors may still matter as a necessary (but not sufficient) condition, but it seems unlikely given that research and the results here that they are the *only* or even the *main* driving forces behind probability matching.

Pragmatic factors, by contrast, may actually differ between adults and children – in real life but especially in an experimental context. Children may be less likely to notice or care about the cues within these experiments that suggest that the variation is predictable. They probably do not carry the same assumptions about laboratory experiments, and they may also feel less pressure than adults to perform "correctly" and not appear foolish. The imitative nature of children might also make them more likely to have coordination as a goal rather than correctness. Indeed, the ability to monitor oneself for correctness has been theorized elsewhere to account for at least some of the differences between child and adult language acquisition (Elkind, 1970; Krashen, 1982). This is speculative, but these experiments give reason to think that pragmatic factors underlie regularization behavior to some extent. It also predicts that children should regularize less if they are given cues suggesting that the variation is predictable and they are expected to learn it well.

So far we have mainly been considering how this research bears on the experimental literature. Of course, especially within the area of language, much of the interest in regularization arises because children and adults appear to differ in their propensity to regularize *outside* of the lab as well as in it. Deaf children exposed to inconsistent sign language will regularize what they hear (Singleton & Newport, 2004), and the process of creolization has been argued to result from children regularizing inconsistent pidgin languages (Bickerton, 1981). By contrast, adult language learners are known to produce variable, inconsistent utterances, even after years of experience with the language and after their grammars have stabilized (Johnson, Shenkman, Newport, & Medin, 1996). If children's regularization in these experiments is driven by differences in the assumptions they bring to the task, how

do we explain these differences in real life?

One answer is that the attested instances of children regularizing in real life are direct analogues of the conditions in which adults regularized in these experiments here: both groups regularize in situations in which the variation is obviously *not* predictable. It is precisely when children must learn from non-fluent speakers, as in the case of inconsistent parents and pidgins, that the variation is truly unpredictable; it therefore may not be surprising that children are sensitive to this and regularize in exactly those situations (Senghas & Coppola, 2001; Singleton & Newport, 2004). How might they be able to tell they are in such a situation? Unreliable or not fully fluent speakers are perceptually distinct in many ways, including things like speech rate, the nature of their disfluencies, and the number and length of pauses (Towell, Hawkins, & Bazergui, 1996; Ejzenberg, 2000; Freed, 2000). Indeed, people react differently to disfluencies depending on the reason for them (Arnold, Hudson Kam, & Tanenhaus, 2007). Adults learning a pidgin language are in this situation, and indeed some argue that creolization occurs due to the adult language learning of such languages over many generations (Arends, 1993). We also know that children are sensitive to the reliability of their teacher when determining what linguistic generalizations to make (Jaswal & Neely, 2006; Jaswal, McKercher, & VanderBorght, 2008). It would not be surprising if children, like the participants in these experiments, were also capable of subtle judgments about degrees of linguistic proficiency, leading them to regularize inconsistent but not consistent input from an imperfect speaker.

Another possibility is that because these experiments more directly reflect word learning while the real life situations (and potentially the Hudson Kam & Newport studies) are more about grammar learning, they may be reflecting or measuring fundamentally different things. This seems unlikely, though. In terms of the task in Hudson Kam and Newport (2005, 2009), although it is embedded within learning a grammar, the task itself still involves consistent stems (nouns) associated with inconsistent affixes (determiners). People didn't have to have learned the grammar to do the sentence completion task in which regularization was measured – the language was VSO and the experimenter provided the initial verb, so participants only had to produce the same stem-and-affix combination as in these experiments. Perhaps the fact of being embedded within a grammar-learning task makes a difference? We cannot rule the possibility out, but are unaware of any literature that suggests that word learning under variation should make people probability match more or even differently than they do while grammar learning. As discussed in the introduction, just as there are multiple words that refer to the same thing, many grammatical constructions have linguistically contextualized variability and probabilities, and there is substantial evidence that people pick up on (Ellis, 2002; Ellis, O'Donnell, & Römer, 2014).

One final issue is how to reconcile these results with the literature finding that adult regularization is affected by apparently non-pragmatic factors like complex input (Hudson Kam & Newport, 2009; Ferdinand et al., 2013) or retrieval pressure (Hudson Kam & Chang, 2009). The most likely possibility is that adequate memory at retrieval and certain pragmatic assumptions are each individually *necessary* but neither alone is *sufficient* to cause probability matching. Without the belief that the variation is predictable, people will not try to learn or reproduce it; without the capacity to remember what they have seen or reported earlier, people will not be able to. But even beyond that, at least in the retrieval pressure experiment pragmatic factors may come into play as well (Hudson Kam & Chang,

2009). In that research, retrieval pressure was eased (resulting in less regularization) by either giving people flashcards with all of the nouns (stems) and determiners (affixes), or by providing the nouns alone. Either of these manipulations can serve to underline to participants that there is a correct answer and their job is to find it. The flashcard manipulation may have another effect as well, since showing all of the determiners creates a demand characteristic that implies they should all be used.

That said, the contention in this paper is that pragmatic factors in general, and the two assumptions manipulated here in particular, are important factors underlying people's tendency to probability match. They are not the *only* factors that matter: for instance, participants will not be able to reproduce variation that they cannot remember, so memory retrieval must be important in some way. The point is that these two assumptions are necessary conditions for probability matching, and that they may explain adult behavior, at least in part.

This research highlights the importance of considering the pragmatic assumptions people may be operating under when they approach language-learning tasks. It is critical to continue to explore the extent to which these and similar assumptions might explain people's behavior in both experimental and real-world contexts.

Appendix A

Instructions

The instructions in the STANDARD and EXPERIMENTER condition were as follows:

> We are studying how people learn new words. Thus you will be presented with a series of pictures of common objects (balls, dogs, etc) with labels from an artificial (non-English) language.
>
> The plan for the experiment is as follows: first you will see 50 picture-label pairs, which you should try to learn. Then you will be shown 10 pictures without labels, and will be asked to write the label for the picture. This will be repeated four times (so you'll see 200 total pictures and have to provide 40 total labels). It should take you about 20 minutes. Please do NOT write any labels down. We are interested in how people learn labels by relying on their memory, not by writing things down.

The PARTICIPANT condition had very similar instructions with a few modifications made to the first paragraph:

> We are studying how people learn new words when they are given the labels by other people rather than a computer. Thus you will be presented with a series of pictures of common objects (balls, dogs, etc) with labels from an artificial (non-English) language. The labels actually come from a previous participant, who had to learn the fake language themselves. Some participants were given a very limited time to provide the labels so there might be errors; we're interested in how you learn the labels even in that case.
>
> The plan for the experiment is as follows: first you will see 50 picture-label pairs, which you should try to learn. Then you will be shown 10 pictures without labels,

> and will be asked to write the label for the picture. This will be repeated four times (so you'll see 200 total pictures and have to provide 40 total labels). It should take you about 20 minutes. Please do NOT write any labels down. We are interested in how people learn labels by relying on their memory, not by writing things down.

Appendix B

Instructions

The instructions in the CORRECTGOAL condition were as follows:

> This experiment is about how people learn and use new words. Imagine that you are a scientist who has discovered a community of people who made up their own words for common objects. You would like to figure out what their words mean. For that, you have presented a person in the community with a set of objects, and asked them to label them. Your task is to try to learn the labels for each of the kinds of objects that you see.
>
> There will be 200 trials in total. After 100 trials, and again at the end of the task, you will be shown some pictures without labels and you will be asked to enter the correct label for that picture.
>
> To make this more interesting (and more like real language) you have been paired with another participant. After the final test, you will exchange computers with the other person, and they will be given the labels you entered for the pictures. Their job will be to match the label with the correct picture. At the same time, you will go to their computer and shown their labels, which you will have to match with the correct picture.
>
> Your payment will depend on the JOINT performance of you and your partner on this last test! That means that you will want to (a) make sure you do your very best to learn the labels; and (b) make sure you do your best to label your pictures clearly and correctly, so your partner can answer their questions correctly. [Note: if you just use normal English words you will get paid the same as if your partner got none right; you need to use the words you have learned.]

In the MATCHGOAL condition the first two paragraphs were identical. The third and fourth paragraphs were replaced by the following:

> To make this more interesting (and more like real language) you have been paired with another participant. At the end of the experiment we will look at the labels that you and your partner have produced. Since in real language it is important that everyone use the same words for things, your payment will depend on the extent to which you both produce the same label for the same picture.
>
> Note that it is important that you produce EXACTLY the same label; otherwise this will not be counted as the same. Also, if you just use normal English words you will get paid the same as if you matched on no labels at all; you need to use the words you have learned. It is important to learn the new labels!

		PARTICIPANT		EXPERIMENTER	
Measure	STANDARD	SIMILAR	TYPOS	SIMILAR	TYPOS
Accuracy	0.69 (0.21)	0.66 (0.24)	0.74 (0.21)	0.63 (0.25)	0.65 (0.20)
Levenshtein dist	0.63 (0.49)	0.71 (0.68)	0.50 (0.50)	0.74 (0.61)	0.69 (0.44)
Regularization index	0.57 (0.19)	0.70 (0.24)	0.76 (0.21)	0.63 (0.22)	0.62 (0.25)
Conditional entropy	0.99 (0.37)	0.72 (0.48)	0.65 (0.43)	0.84 (0.45)	0.82 (0.48)

Table 2: Descriptive statistics for all measures in Experiment A. Numbers denote means, while items in parentheses are standard deviations.

Measure	CORRECTGOAL	MATCHGOAL
Accuracy	0.81 (0.21)	0.71 (0.26)
Levenshtein dist	0.31 (0.41)	0.58 (0.59)
Regularization index	0.48 (0.22)	0.76 (0.18)
Conditional entropy	0.95 (0.40)	0.52 (0.31)

Table 3: Descriptive statistics for all measures in Experiment B. Numbers denote means, while items in parentheses are standard deviations.

Appendix C

This appendix contains all of the descriptive statistics for all conditions, shown in Tables 2 and 3.

In Experiment A, the number of people classified as regularizers according to the 90% threshold were: 17 out of 56 people (30%) in the PARTICIPANT condition, 9 out of 53 (17%) in the EXPERIMENTER condition, and 2 out of 25 (8%) in the STANDARD condition. Within the SIMILAR condition, 9 out of 49 (18%) were classified as regularizers, while in the TYPOS condition 17 out of 60 were (28%). The 80% threshold yielded 12 out of 53 regularizers (23%) in the EXPERIMENTER condition, 24 out of 56 (43%) in the PARTICIPANT condition, 3 out of 25 (12%) in the STANDARD condition, 15 out of 49 (31%) in the SIMILAR condition, and 21 out of 60 (35%) in the TYPOS condition.

In Experiment B, 5 of 22 people (23%) in the MATCHGOAL condition were regularizers compared to 1 of 24 (4%) in the CORRECTGOAL according to the 90% threshold. If the threshold is set to 80% rather than 90%, 12 of 22 (55%) of people in the MATCHGOAL condition are regularizers, compared to 3 of 24 (13%) in the CORRECTGOAL condition.

Acknowledgements

This work was supported by ARC DECRA Fellowship DE120102378.

References

Arends, J. (1993). Towards a gradualist model of creolization. In F. Byrne & J. Holm (Eds.), *Atlantic meets pacific: A global view of pidginization and creolization* (pp. 371–380). Amsterdam: John Benjamins.

Arnold, J., Hudson Kam, C., & Tanenhaus, M. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914—930.

Atkinson, R., Sommer, G., & Sterman, M. (1960). Decision making by children as a function of amount of reinforcement. *Psychological Reports*, *6*, 299-306.

Baum, W. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, *32*, 269-281.

Bever, T. (1982). *Regressions in mental development: Basic phenomena and theories.* Lawrence Erlbaum Assoc.

Bickerton, D. (1981). *Roots of language.* Ann Arbor, MI: Karoma.

Castellan, N. J. (1974). The effect of different types of feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *11*, 44-64.

Chambers, J., Trudgill, P., & Schilling-Estes, N. (2003). *The Handbook of Language Variation and Change.* Blackwell.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317-336.

Craig, G., & Myers, J. (1963). A developmental study of sequential two-choice decision making. *Child Development*, *34*(2), 483-493.

Crump, M., McDonnell, J., & Gureckis, T. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *e57410*.

Damerau, F., & Mays, E. (1989). An examination of undetected typing errors. *Information Processing & Management*, *25*(6), 659–664.

Davis, S., Newport, E., & Aslin, R. (2011). Probability-matching in 10-month-old infants. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3011–3015). Austin, TX: Cognitive Science Society.

Derks, P., & Paclisanu, M. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, *73*(2), 278–285.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–314). The University of Michigan Press.

Elkind, D. (1970). *Children and adolescents: Interpretive essays on Jean Piaget.* New York: Oxford University Press.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N., O'Donnell, M., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Cognitive Linguistics*, *25*(1), 55–98.

Ferdinand, V., Thompson, B., Kirby, S., & Smith, K. (2013). Regularization behavior in a non-linguistic domain. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 436–441). Austin, TX: Cognitive Science Society.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). The University of Michigan Press.

Herrnstein, R. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267-272.

Herrnstein, R. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*, 243-266.

Horwitz, E., Horwitz, M., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125-132.

Hudson Kam, C., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *35*(3), 815–821.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning & Development*, *1*(2), 151–195.

Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*, 30–66.

Jaswal, V., McKercher, D., & VanderBorght, M. (2008). Limitations on reliability: Regularity rules in the English plural and past tense. *Child Development*, *79*, 750–760.

Jaswal, V., & Neely, L. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science*, *17*, 757–758.

Johnson, J., Shenkman, K., Newport, E., & Medin, D. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, *35*, 335–352.

Jones, M., & Liverant, S. (1960). Effects of age differences on choice behavior. *Child Development*, *31*, 673-680.

Kamps, C., Ferdinand, V., & Kirby, S. (2014). The origins of regularity in language: why coordination matters. In *Proceedings of the 10th International Evolution of Language Conference.*

Koehler, D., & James, G. (2003). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*, 123-127.

Krashen, S. (1982). *Principles and practice in second language acquisition.* Pergamon Press.

Lefkowitz, N., & Hedgcock, J. (2002). Sound barriers: influences of social prestige, peer pressure and teacher (dis)approval on FL oral performance. *Language Teaching Research*, *6*, 223-244.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *6*, 707-710.

MacNeilage, P. (1964). Typing errors as clues to serial ordering mechanisms in language behaviour. *Language and Speech*, *7*, 144–159.

Messick, S., & Solley, C. (1957). Probability learning in children: Some exploratory studies. *The Journal of Genetic Psychology*, *90*, 23-32.

Moran, J., & McCullers, J. (1979). Reward and number of choices in children's probability learning: An attempt to reconcile conflicting findings. *Journal of Experimental Child Psychology*, *27*, 527-532.

Offenbach, S. (1964). Studies of children's probability learning behavior: I. effect of reward and punishment at two age levels. *Child Development*, *35*, 709-715.

Perfors, A. (2011). Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3274–3279). Austin, TX: Cognitive Science Society.

Perfors, A. (2012a). Probability matching vs over-regularization in language: Participant behavior depends on their interpretation of the task. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 845–850). Austin, TX: Cognitive Science Society.

Perfors, A. (2012b). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(486-506).

Pierce, W. D., & Epling, W. F. (1983). Choice, matching, and human behavior: A review of the literature. *The Behavior Analyst*, *6*, 57-76.

Ramscar, M., & Gitcho, N. (2007). Develomental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*(7), 274-279.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Restle, F. (1961). *Psychology of judgment and choice: A theoretical essay.* New York: Wiley.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan sign language acquired a spatial grammar. *Psychological Science*, *12*, 323–328.

Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Singleton, J., & Newport, E. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *216*, 444–449.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods*, *43*(1), 155-167.

Stevenson, H., & Hoving, K. (1964). Probability learning as a function of age and incentive. *Journal of Experimental Child Psychology*, *1*, 64-70.

Stevenson, H., & Odom, R. (1964). Children's behavior in a probabilistic situation. *Journal of Experimental Psychology*, *68*(3), 260-268.

Stevenson, H., & Weir, M. (1959). Variables affecting children's performance in a probability learning task. *Journal of Experimental Psychology*, *57*(6), 403-412.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119.

Tsui, A. (1996). Reticence and anxiety in second language learning. In K. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language acquisition.* Cambridge University Press.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101-118.

Wearden, J. (1983). Undermatching and overmatching as deviations from the matching law. *Journal of the Experimental Analysis of Behavior*, *40*, 332-340.

Weir, M. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, *71*, 473–490.

West, R., & Stanovich, K. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *1*, 1–14.