Language Learning and Development

ISSN: 1547-5441 (Print) 1547-3341 (Online) Journal homepage: https://www.tandfonline.com/loi/hlld20

Reconsidering Retrieval Effects on Adult Regularization of Inconsistent Variation in Language

Carla L. Hudson Kam

To cite this article: Carla L. Hudson Kam (2019) Reconsidering Retrieval Effects on Adult Regularization of Inconsistent Variation in Language, Language Learning and Development, 15:4, 317-337, DOI: 10.1080/15475441.2019.1634575

To link to this article: https://doi.org/10.1080/15475441.2019.1634575

LANGUAGE LEARNING AND DEVELOPMENT
2019, VOL. 15, NO. 4, 317–337
https://doi.org/10.1080/15475441.2019.1634575

Routledge
Taylor & Francis Group

Reconsidering Retrieval Effects on Adult Regularization of Inconsistent Variation in Language

Carla L. Hudson Kam

Department of Linguistics, The University of British Columbia, Vancouver, Canada

ABSTRACT

The phenomenon of regularization – learners imposing systematicity on inconsistent variation in language input – is complex. Studies show that children are more likely to regularize than adults, but adults will also regularize under certain circumstances. Exactly why we see the pattern of behaviour that we do is not well understood, however. This paper reports on an experiment investigating whether it is possible to induce regularization in adults by varying the conditions of learning and/or testing in ways that made retrieval more difficult, something predicted by previous studies of regularization. The data show that interfering with learning does not lead to regularization, but that interfering with retrieval at test does, although only to a small degree.

Introduction

Regularization of inconsistent variation in language was first examined by Newport and colleagues (Ross & Newport, 1996; Singleton & Newport, 2004); they described the language development of a child called Simon who received input from his parents (and importantly only his parents) that contained unpredictable variation that he did not replicate. Specifically, while his parents' productions contained numerous morphological errors, his own signing was much more consistent; he seemed to filter out the errors and use only their predominant morphological patterns in his own signing. His is not an isolated case, however; subsequent work by Ross and Newport shows the same pattern in a different set of children (Ross, 2001). Although these are the best described cases at the level of individual learners (in terms of input and learning outcomes) there is reason to believe that the same process is involved in cases of language emergence and change, e.g., some of the developments in Nicaraguan Sign Language can be described as learners imposing a systematicity on previously meaningless variation present in their input (Senghas & Coppola, 2001), and some aspects of the evolution of French-based creoles have also been described this way (Becker & Veenstra, 2003).

These real-life cases of regularization have inspired laboratory experiments designed to better understand regularization in language learning. Most laboratory studies have explored when regularization happens, thus providing a picture of what needs to be explained. In the earliest of these studies, Hudson Kam and Newport (2005) exposed child and adult learners to a miniature artificial language in which meaningless determiners (single syllable forms that followed the nouns in a sentence) occurred probabilistically. That is, they occurred (or not) with the nouns at a rate specified by the researchers, unrelated to any phonological or meaning distinctions in the language. They tested learners after several days of exposure to the language. The primary test was a production task in which learners were asked to produce novel sentences. Hudson Kam and Newport then analysed participants' determiner usage. They found that children were very likely to

CONTACT Carla L. Hudson Kam Carla.HudsonKam@ubc.ca Department of Linguistics, The University of British Columbia, 2613 West Mall, Vancouver V6T 1Z4, Canada.

regularize this unpredictable variation; they imposed patterns on the language that were not present in the input, e.g., using the determiners all of the time rather than probabilistically. In contrast, the adult learners were quite good at replicating the same patterns of variation in their own production; if they heard determiners with 75% of the nouns in their input they produced determiners about 75% of the time in the novel sentence production task.

That study seemed to suggest that there is a categorical difference between children and adults, however, subsequent studies have demonstrated that it is not the case that children regularize while adults don't. Hudson Kam and Newport (2009), for instance, found that adults will also regularize given slightly different, but still variable, input. Specifically, when the input was variation between multiple determiners occurring in the same contexts (rather than a determiner alternating with no determiner) adults overproduced the most common form, with the amount of overproduction increasing as the number of alternatives increased, even when the probability of the most frequent form was held constant. Children regularized this sort of input too, just as they had with the simpler variation. Similarly, Ferdinand, Kirby, and Smith (2017) found increased regularization in a label learning task when the number of labels applied to a single object increased. Other work has shown that details of the production task also lead to more or less regularization by adults. Wonnacott and Newport (2005), for instance, found that adults asked to produce sentences using novel nouns produced determiners in a regular pattern in contrast to those asked to produce familiar sentences at test (who probability matched). They also found that adults asked to produce both kinds of sentences regularized even when producing the familiar sentences, suggesting that the production of sentences with novel nouns had an impact on the production of the familiar sentences. And studies using an iterated learning paradigm (in which one learner's output is given to a new learner, then their output is given to yet another learner, etc., in a diffusion chain) have found that variation of even the simple sort will slowly get smoothed out and removed from the language by adults (Reali & Griffiths, 2009; Smith & Wonnacott, 2010).[1]

What causes regularization?

Hudson Kam and Newport (2009) suggested that regularization might emerge from retrieval issues related to memory inefficiencies. Children have more difficulty retrieving items from memory and so over-retrieve some items. This can also explain why adults might regularize when faced with multiple items in competition – each individual lower-frequency form is harder to retrieve, making the retrieval of the most frequent form increasingly likely. Hudson Kam and Chang (2009) tested one of the predictions of this idea. They exposed adult learners to a miniature artificial language (MAL) containing numerous variable forms in competition, a type of variation Hudson Kam and Newport showed adults regularize. However, they used testing techniques that minimized the retrieval demands and showed that adults no longer regularized the language under these conditions. Specifically, for the production test, in which participants had to produce an utterance describing a novel-to-them scene or event, one group was provided the verb and noun(s) orally, and had to recall any other words to be included and produce the entire sentence. Another group of participants were given flashcards on which all the words of the language were written. They too heard the verb (to ensure they had the same target sentence), but their task was to select and arrange cards on which the words were written to "produce" the sentence. Participants in these two conditions matched the probability of the determiner in their productions, in contrast to participants given the standard version of the production test (being given the verb and having to produce the rest from memory). These data thus show that reducing retrieval demands in adults reduces regularization, consistent with Hudson Kam and Newport's (2009) proposal.

[1]The design and specific outcomes (i.e., the nature of the "smoothing out") differ between these two studies in ways that are potentially theoretically quite interesting. However, they both demonstrate the removal of unpredictable variation by adults in diffusion chains.

Perfors (2012) tested another possible prediction from Hudson Kam and Newport's (2009) proposal; that making learning harder, simulating inherent memory limitations during encoding, would increase regularization. She exposed adults to a miniature artificial noun lexicon, consisting of 10 nouns each of which was accompanied by 5 different determiners, one of which occurred 60% of the time, the others each occurring 10% of the time. This mimics one of the input patterns found by Hudson Kam and Newport (2009) to lead to over-production of the most frequent form. In Perfors' study, the input phase included concurrent tasks interleaved with the input, in an attempt to interfere with learning. Over seven experiments with different concurrent tasks, learning was clearly impacted as shown by performance on vocabulary tests. However, there was no greater regularization of the determiners by participants given the concurrent tasks compared to those with no concurrent task. Based on the experimental results and a series of computational models, Perfors concludes that memory constraints during learning are unlikely to explain regularization by themselves, and suggests that there must also be a prior bias to regularize involved.

There is another possibility raised by Hudson Kam and Newport (2009) and Hudson Kam and Chang (2009), however, namely that regularization results from issues during retrieval that are caused by or related to retrieval processes themselves. The cognitive abilities underlying self-directed retrieval are well established to be poorer during childhood than adulthood (Davidson, Amso, Anderson, & Diamond, 2006; Munakata, Snyder, & Chatham, 2012), and have been proposed to be related to other differences between adult and child language learners (Finn, Lee, Kraus, & Hudson Kam, 2014; Thompson-Schill, Ramscar, & Chrysikou, 2009). The basic idea here is that even when the representations themselves are fine (i.e., there were no memory constraints impacting learning), retrieval difficulties inherent to retrieval itself can affect production processes, resulting in regularization. A person who has fewer executive function resources, either inherently (as is the case with children) or due to in-the-moment task related effects (e.g., split attention) will be less able to access and retrieve the relevant information in memory. This will result in the over-retrieval of easily retrieved forms as they require less cognitive effort, and this then will result in regularization.

Assessing retrieval effects

I present results from a study contrasting these two possibilities. Adult language learners were exposed to a miniature artificial language containing unpredictable variation previously shown to lead to low levels of regularization in adults. In one experimental condition the participants were given a concurrent task at the time of learning (as in Perfors, 2012), and in the other, a concurrent task at the time of the production test. The first condition was designed to cause interference with learning, possibly resulting in weaker representations that are therefore more difficult to access and retrieve. The second condition was designed to interfere only with retrieval itself, leaving the representations themselves unperturbed.

As described above, Perfors (2012) showed that adults given a concurrent task during learning did not show increased regularization, suggesting that we should not see regularization from participants given the concurrent task during learning. However, her artificial language only included nouns and determiners/particles. Real languages, where we know that regularization can occur (see, e.g., Singleton & Newport, 2004), involve producing more than small phrases (or single words with their particles). It is not clear we would expect to see the effects of retrieval difficulties on production when participants are given such a simple task, as fewer steps are involved. Producing a sentence involves selecting the content words and their accompanying function words, arranging them into phrases and assembling the phrases into a sentence in the correct order (see, e.g., Bock, 1995). Given that the order is different in the MAL and English, the native language of participants, even simple things such as getting the order right are not necessarily error-proof and so involve some thought and effort (Hudson Kam, 2009). Perfors' (2012) task, in contrast, can be solved much more easily and without any syntactic or hierarchical representation. Another difference is the length of the studies: Perfors' lasted a single day. Participants' memories were fresher than in a longer study with testing conducted on a separate day. And there was no

opportunity for sleep in between exposure and testing. Sleep has been shown to increase abstraction and generalization (Fenn, Nusbaum, & Margoliash, 2003; Gómez, Bootzin, & Nadel, 2006), something apparent in the regularization seen in previous laboratory studies of regularization. Whether or not sleep is crucial, length of exposure does seem to be a relevant variable. In a previous study that found a concurrent task during learning increases generalization, this effect was only significant in a longer version of the study (Pitts Cochran, McDonald, & Parault, 1999). Note that the use of a richer miniature artificial language in the present work also allows for a more direct comparison with the results of Hudson Kam and Newport (2009) and Hudson Kam and Chang (2009). In addition, it leaves greater scope for participants to impose idiosyncratic rules on the language when they are producing sentences (e.g., conditioning on syntactic type), and so evince greater regularization. Thus, it is possible that a different methodology might produce different results than Perfors. If I find regularization where Perfors did not, it would suggest that the nature of the task used is important. If, like her, I fail to find evidence for regularization in the interfering with learning condition, then it provides a conceptual replication of her results.

Methods

Participants

Fifty-two native English speakers (average age = 22.02 years, SD = 8.5, range = 18–66 years), participated in the experiment. They were recruited from a participant pool via an e-mail describing the study, or after having responded to recruitment posters. They were paid for their participation. Participants were randomly assigned to one of three conditions: a control condition (N = 18) and two experimental conditions (N = 18 & N = 16). Age did not differ significantly between the three conditions ($F(2,49) = 0.840$, $p = .438$).

The language and exposure stimuli

The basic vocabulary of the miniature artificial language comprised 28 nouns referring to objects and substances (e.g., block, cotton batten) which were divided into two noun classes, each of which occurred with its own distinct main determiner (like a noun class marker or gendered article). There were also 6 intransitive and 4 transitive verbs referring to actions, relationships, and properties (e.g., move, be under, be big), and a negative word. Nouns were assigned to a class on a completely arbitrary basis (that is, there was no difference in meaning or sound correlated with class assignment), with 15 nouns in Class 1 and the remaining 13 in Class 2. (This was a subset of the language used by Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005, 2009. A few nouns were in different classes in this study and those.) This asymmetric size is typical of real noun class systems. The only consequence of class is main determiner choice. There were also two other determiner forms which I refer to as noise determiners that occurred with nouns in both classes.

As in previous work using this methodology, the word order in sentences is (NEG)V-S-(O), where NEG represents negative, V represents verb, S represents subject, and O represents object. (Brackets indicate that the constituent is not obligatory in all sentences.) As is typical for VSO languages, the determiner (det) follows the noun within the noun phrase (Greenberg, 1963). This ordering ensured that participants would have to learn aspects of the language's grammar as well as the vocabulary. An example sentence is presented here in (1):

(1) mirt misna poe
move snake (NC2) maindet2

"The snake moves."

The exposure set contained 60 intransitive and 60 transitive sentences, each repeated six times over the course of exposure. The main determiner forms each occurred 60% of the time nouns in their respective classes occurred (NC1 main det = kaw, NC2 main det = poe). This was true of each syntactic position, and each input session. The two noise determiners (tay and meg) each occurred with 20% of the nouns in each class, and again, this was true by syntactic position and input session. Because the main determiners only ever occurred with nouns in their respective class, if we compute the percentage of noun tokens overall that occurred with each main determiner separately, the percentage is approximately 30% (30.64% for NC1 and 29.26% for the NC2 main form), not 60%. The noise forms, by contrast, occurred with nouns in both classes, so they occurred with 20% of the nouns overall as well as with 20% of the noun tokens within a class. Thus, the noise forms were less frequent and less predictable than the main forms. Each instance of any sentence was independent of the other instances of the same base sentences, and so had the potential to be different from the other instances (within reason – given that there are only three possible determiner forms for each noun, the possible variants are limited, although they are more numerous for transitive than intransitive sentences).

Experimental manipulation

There were three conditions, one control and two experimental. The control condition was a replication of the 2 noise determiner condition in Hudson Kam and Newport (2009). Participants in this condition were exposed to the language and tested in the usual way, that is, with no interfering tasks at any point in time. They were expected to produce the main determiner forms slightly more often than they had heard them. In the first experimental condition the interference occurred during learning (IntLearn). In this condition, low and high tones were interspersed throughout the exposure videos. Tones never occurred at the same time as a sentence. Participants were prompted to count the number of high (or low) tones within a block. They were explicitly told to keep the count in their heads, not to use fingers or other things to keep the tally. At the beginning of the block participants were informed as to which kind of tone to count in that block. At the end of the block a prompt came on the screen telling them the block was over, at which time they were to write the number of tones on an answer sheet. Next they were told which kind of tone to count in the next block. Blocks randomly varied as to whether participants were supposed to count the high or low tones. They also varied in terms of the number of sentences contained within a block (3–13 sentences), the number of tones within a block (2–8), and the number of target tones (1–6). The blocks were the same for all participants. The task was based on one used by Pitts Cochran et al. (1999). They found greater generalization in a language-learning task when participants were engaged in a similar (but slightly simpler) tone counting task while learning, as compared to participants who were not engaged in the secondary task. In the other experimental condition the concurrent task occurred during testing (IntTest). In this condition, participants were given a set of three numbers prior to each item in the production test. They were to keep the numbers in memory, in order, until after they had produced their response (i.e., the sentence in the sentence production test), at which time they would be prompted to recall the set of numbers. Then they were given a new set of numbers and this repeated. This set size was selected to be within all participants' normal short term memory capabilities (see Appendix A), and has been shown to lead to regularization in a non-linguistic probability-learning task (Wolford, Newman, Miller, & Wig, 2004). The IntLearn condition tests the idea that regularization might result from retrieval issues caused by representational strength effects resulting from learning difficulties. The IntTest condition tests the possibility that retrieval difficulties may (instead) be due to differences in retrieval efficacy related to executive function; in this case, executive function difficulties imposed by the experimenter on all participants in the condition.

The task in the IntLearn condition is most similar to Perfors' (2012) concurrent operational load condition. In that condition participants were given a simple math problem to solve at the same time

they were presented with an image-label pairing, and had to quickly solve it before moving on to the next item. The two tasks are similar in that both involve some mental arithmetic and reporting numerals. The IntLearn task differs in that participants never know how long they are going to have to hold onto the count for. In addition, it involves suppression (not counting the irrelevant tones) and switching (from high to low or low to high tones), which can themselves exert some cognitive cost. So while counting and retaining a count over a short span of time might seem easier than solving a math problem, the IntLearn task is not as easy for participants as it might at first seem. There is not an exactly analogous condition to the IntTest condition in Perfors (2012) because she only interfered during learning. However, the task in her low concurrent load condition is almost the same as that used in the IntTest; she asked people to hold onto three letters for later recall, whereas I asked people to remember three numbers. A major difference between how the task was used in the two studies concerns the nature of what people were doing while holding on to the three items. In Perfors (2012) they saw an image on a computer screen and heard a recorded label for that image. In the present study they had to produce a novel sentence describing a scene on a video monitor. So it is possible that same amount of cognitive interference will have a different impact because of what the participant was doing while performing the interfering task.

Procedure

Participants were exposed to the language by videotape for eight sessions, each lasting 20–25 minutes and each occurring on a different day. Participants were seated in front of a video monitor, on which played a scene or event accompanied by a spoken sentence. Participants were asked to repeat each sentence after hearing it. They were told that this practice would be helpful, as they would have to produce their own sentences at the end of the experiment. There was no explicit instruction in either grammar or vocabulary and participants never saw anything written. Testing occurred in a separate, ninth, session. Participants completed the MAL study in 11–13 days, depending on their schedules.

Tests

Vocabulary test

Participants were tested on their knowledge of 12 nouns relevant to the later production test (the test of primary interest) twice. The first test was administered at the end of the fourth session (half way through the exposure sessions). Participants were asked to provide a name for each object as it appeared on the screen in front of them. They had as much time as they wanted to respond. Participants were told that the results of this test would not be analyzed, that it was just to give them some idea of how they were doing up to this point.

The second vocabulary test followed the same format. Participants were tested on the same 12 items as in the first vocabulary test, but the order in which the items appeared was different. This test was used to evaluate whether participants had learned enough vocabulary to complete the sentence production test. It was administered at the start of the final session. Responses were video-recorded and later transcribed for analysis. However, the research assistant kept a running tally of the number of correct responses; participants needed to be able to correctly produce at least 5 of the 12 nouns to go on to take the sentence production test.

Sentence production test

Participants' use of the variable part of the language, the determiners, was assessed in a sentence production task. In this test they saw a novel scene, were provided (aurally) with the verb to ensure that they were producing the intended sentence, and asked to produce a sentence describing the scene or event using the word they heard. Thus, they had to recall the noun(s), the associated determiner(s), and arrange the words into an utterance. Because this test occurred in a session separate from the input sessions, participants had to rely on their memory for the novel language,

nothing had been reactivated (by the experimenter) in the test session, and so responses were unlikely to reflect things such as the last sentence heard in exposure. Twelve nouns each occurred three times in this test, once in each of the three syntactic roles (IT Subj, Tr Subj, Tr Obj). The 12 transitive test items were followed by the 12 intransitive test items. Responses were video-recorded and later transcribed for analysis.

Results

Concurrent tasks

It is important for the question under investigation that the secondary tasks were difficult enough to interfere with the other things participants were attempting to do at the same time. Thus, I begin by analyzing participants' performance on the concurrent tasks.[2]

Tone counting (IntLearn)

Performance on the tone counting task was assessed by computing the percentage of tones the participant correctly noted. This was simply the number of tones reported divided by the number of target tones actually present, in the exposure set overall. Performance on this varied, but no one counted every target tone, suggesting that this was of sufficient difficulty to create potential interference with learning. Participants reported 71.83% of the tones on average (sd = 14.22; range = 39.83%-93.22%, individual data points shown in Appendix B).

Random numbers (IntTest)

Performance on this was scored using the scale reported in Appendix C. The basics of the system are this: A perfectly correct answer including all three numbers recalled in the correct order received a zero, anything else received a score from 1–9, where lower numbers correspond to answers that were closer to the correct response (9 represents no correct numbers reported). Scores for the 24 test sentences were added together, with lower scores indicating better performance. The best possible score is 0, and the worst is 216. The mean score here was 25.78 (sd = 23.35; range = 0–72, individual data points shown in Appendix B). Only 2 of the 18 people in this condition achieved a perfect score. Again, the scores suggest that this task was of sufficient difficulty to interfere with production.

Vocabulary

Performance on the second vocabulary test is shown by condition in Figure 1.[3] Perfors (2012) found that people given concurrent tasks during learning performed worse on the vocabulary test than those without a concurrent task during learning. There is a numerical trend towards this in the present data as well, but the difference is not significant ($F(2,49) = 0.646$, $p = .229$, partial Eta-squared = .026). Note also that the participants who had the concurrent task during learning (IntLearn) look the same as those who had the concurrent task during their sentence production task (IntTest) who had no reason to show worse learning than participants in the control condition. Thus, it appears that the concurrent task the IntLearn participants faced did not significantly affect their ability to learn the vocabulary. This does not say that their

[2]What counts as "difficult enough" is unclear. Showing the predicted effect on regularization would be evidence for interference, but that is circular. Moreover, we would not want a failure to show effects on regularization to be taken as evidence for a lack of interference; if we did so, then we could not disprove the hypothesis that regularization results from cognitive limitations. Thus, we need another demonstration that the tasks were difficult enough to at least have the potential to cause interference. Showing that performance was not perfect demonstrates, at the very least, that they did cause some difficulty for participants. I return to this issue in the discussion.

[3]When people got an item wrong it was usually a non-production (e.g., saying "I can't remember"), although occasionally they would produce a different MAL word. Slight mispronunciations were not counted as incorrect (e.g., flomba for flombut).

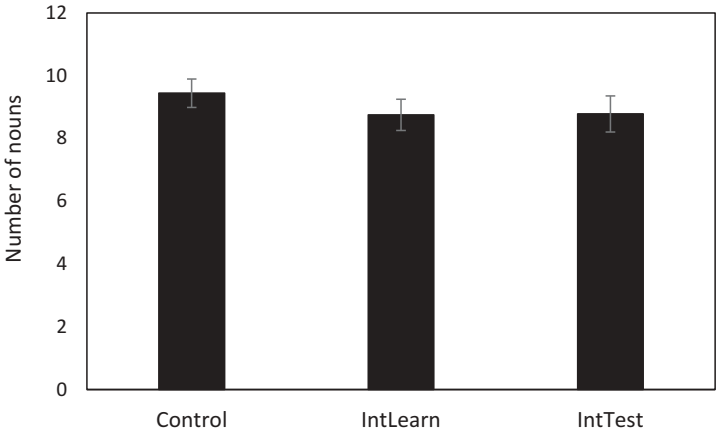


Figure 1. Vocabulary test performance by condition (max = 12; error bars = se).

representations are as strong or robust as the other conditions, however, which is the proposed underlying cause of any potential differences in regularization.

Sentence production

Concurrent task effects

Figure 2a shows the average percentage of noun tokens produced in the sentence completion task that were accompanied by the correct main determiner forms (i.e., the form matching the class of the noun) by condition. A glance at the figure shows that participants given the concurrent tasks did not produce more main determiners than the control participants; instead, they produced the same percentage or less. The main effect of condition is significant ($F(2,49) = 3.392$, $p = .042$, partial Eta-squared = .122), and follow-up contrasts comparing each of the two experimental conditions to the control condition show that while participants in the IntLearn condition produce significantly fewer main determiner forms than those in the control condition ($p = .015$), participants in the IntTest condition do not ($p = .509$). These data, therefore, do not show any evidence for the hypothesis that retrieval issues are a cause of regularization in language.

This first analysis was looking for only one kind of regularization – overproduction of the main determiner form. But regularization can take different forms, e.g., the use of determiners only with nouns with particular meanings or in certain kinds of sentences. If different learners regularize in different ways, an overall analysis could obscure regularization occurring at the individual level, and it is clear from Figure 2b, which shows individual percentages, that there is variation between individuals. So I next examined each participant's productions for evidence of any sort of internal consistency in determiner usage using the categories and coding scheme in Hudson Kam and Newport (2009). Assignment into these categories was done by a research assistant who was trained to use the metric. That is, it was clearly laid out before analysis and so should not have been subject to manipulation by the experimenter; a participant either fell into a category or they did not.

There was evidence for three different kinds of determiner usage. Most common were *scattered* users, people who used both main and noise forms, and used them in unpredictable ways. There were also a number of *systematic main* users, people who always produced the main determiner forms. And finally, there were people who imposed their own idiosyncratic patterns on the language, things like consistently conditioning determiner choice on individual specific nouns. These participants were classified as *systematic other*. As in Hudson Kam and Newport (2009), participants were allowed one exception to any pattern, that is, one noun that is

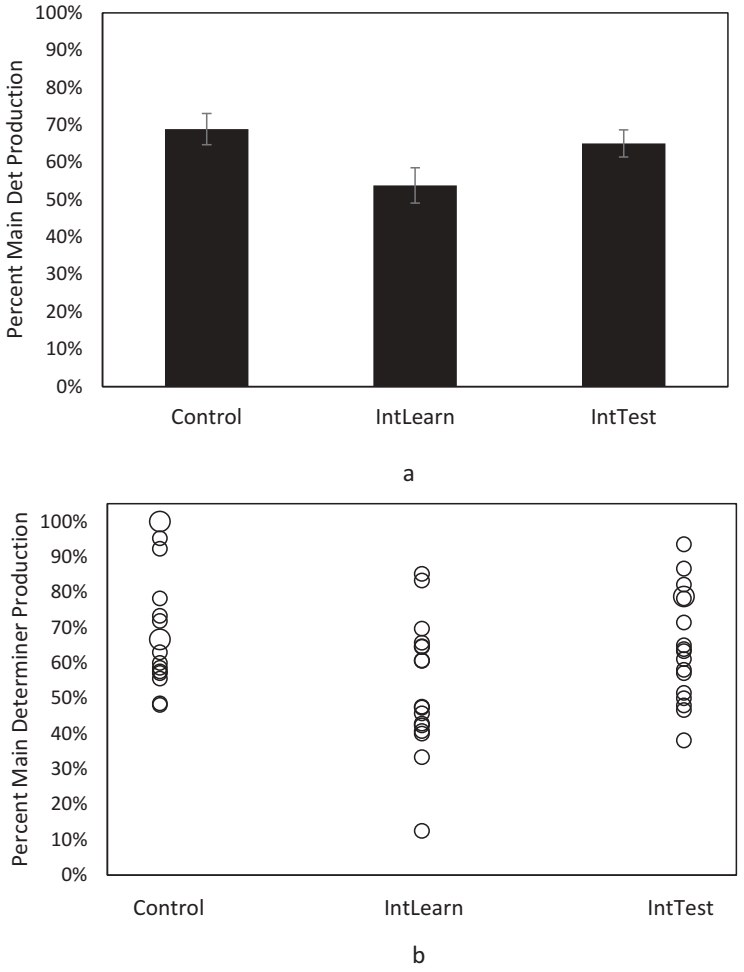


Figure 2. Mean percentage of noun tokens produced with correct main determiner form by condition (error bars=se) (a), Percentage of noun tokens produced with correct main determiner forms by condition - Individual data (b). (The smaller circles represent one participant, the larger circles represent two participants at that value).

different. For example, if a participant used the main determiner associated with each noun for all nouns, except for one production of one noun, that person would be classified as a systematic main determiner user. They were also classified as a systematic main user if all uses of one noun were consistently produced in a way that was inconsistent with the others, for example, if the participant produced all nouns with their respective main determiners except for one, which was always produced with the same noise form, that person would be classified as a systematic main user. If a person produced all but two nouns with the main determiner forms all of the time, the person was not categorised as a systematic main user, as those would constitute more than one exception.

Table 1 shows the number of participants who fell into each of these three categories, by condition. A chi-square test on these data is significant (X^2 = (4, N =52) = 14.703, p =.005). This overall significant result could again be due to the IntLearn condition: a look at the table shows that participants in the IntLearn condition are less likely to be systematic than those in the other two conditions. In fact, there is not a single participant in the IntLearn condition who is systematic in his or her productions. However, there is another, potentially interesting,

Table 1. Determiner production pattern types by condition.

	Production Pattern		
Condition	Scattered	Systematic Main	Systematic Other
Control	14	4	0
IntLearn	16	0	0
IntTest	11	2	5

difference in these data. Specifically, the IntTest condition contains several participants (5) who imposed their own idiosyncratic forms of consistency on the language as they produced it, a category that was not seen in the other two conditions. To see if IntTest differed from the control condition, I ran a second chi-square analysis, this time including only the control and IntTest data. It too is significant, (X^2 = (2, N =36) = 6.027, p =.049). Thus, there does seem to be an effect of increasing the difficulty of retrieval on participants propensity to regularize, or at least, on the way that they regularize.[4]

In response to a reviewer's query about whether the coding was driving the results, I recoded participants who had exceptions and reran the analysis. Participants whose exception comprised a single noun that was produced in multiple ways (e.g., once with a noise determiner and the other times with a main determiner) were recoded as noise users. Participants who were consistent within their single exception (i.e., they consistently used a particular noise determiner with a single noun and the appropriate main determiner with all of the other nouns) were recoded as systematic other. This led to three changes in categorization: one systematic main user in the control condition and one in the IntTest condition became scatter users and one systematic user in the IntTest was recoded as systematic other. A chi-square test on these data too is significant (X^2 = (4, N =52) = 18.104, p =.001). Separate analyses on pairs of conditions showed that the comparison between control participants and those in the IntTest condition remained significant (X^2 = (2, N =36) = 9.333, p =.009). The comparison between the control condition and IntLearn was not significant (X^2 = (1, N =36) = 2.925, p =.087), while the comparison between the two control conditions was (X^2 = (1, N =36) = 6.476, p =.011).[5] This follow-up analysis demonstrates that slight changes in coding did not affect the results, and so the specific coding scheme does not appear to be responsible for the outcome of the analysis.

However, it must be acknowledged that the number of participants in each condition, while larger than in my previous studies examining regularization, is small, and despite the categories being developed a priori based on previous studies, categorization of any kind involves decisions made by experimenters, and so are potentially open to manipulation. At the suggestion of a reviewer, I computed two additional measures that were used by Perfors (2012) for comparison. Convergence in the results of various measures should provide more confidence in the findings, at least as they apply to these particular subjects.

I first report a Regularization Score, similar to the Regularization index used by Perfors (2012). This is simply the proportion of productions occurring with the participant's most common determiner form, whatever that is. Because of the specifics of the language used in the current work, i.e., two distinct main determiners that each occur with a different set of nouns, instead of computing it using specific determiners as "forms", it is computed using types. So both main forms (when used appropriately) counted as a single category, as did both noise forms.[6] The mean proportion and SDs by condition are presented in Table 2. These data present

[4]I am not presenting an analysis of the relationship between the concurrent tasks and regularization here. However, for those who are interested, see Appendix B for more detail.

[5]The different degrees of freedom for the comparisons are due to the fact that in two of the analyses neither condition had any participant in the systematic main category and so the category was dropped from the analysis.

[6]Zero determiner, incorrect main form, and other were also potential types, so that, e.g., a person who mostly produced unmarked forms would have zero marking as their most common type and thus, the proportion of unmarked forms would be their score. However, no participant had one of these three as their most frequent type.

Table 2. Mean regularization score by condition.

	Proportion of Productions with Most Common Determiner Type	
Condition	Mean Proportion	SD
Control	0.6887	0.1772
IntLearn	0.6207	0.1514
IntTest	0.6207	0.1106

the same general picture as the data for correct main determiner production: Participants in the two experimental conditions are not regularizing more than participants in the control group. If anything, it is the opposite; numerically, participants in the control group produce their most common form slightly more than those in the experimental groups, although the effect of condition on regularization scores was not significant ($F(2,49) = 0.874$, $p = .424$, partial Eta-squared =.034).

I also analyzed participants' conditional entropy of determiners given nouns. This is a measure of participant-specific consistency in determiner patterns that does not rely on coding decisions by a researcher.[7] Perfect consistency yields a conditional entropy value of 0, and as values increase above zero consistency decreases. That is, as the relationship between a noun and the determiner(s) that follow it becomes less predictable, the conditional entropy goes up.

This measure shows results that are in between the consistency categorization and the regularization score analyses: looking first at the means (shown in Figure 3a) it seems that participants who experienced interference during testing had much lower conditional entropy than IntLearn participants and slightly lower conditional entropy than control participants. Conversely, the IntLearn participants appear to show much higher conditional entropy than participants in the other two conditions. Indeed, the 95% CI of the mean for the IntLearn (0.3967–0.6127) does not extend to include the mean value of either of the other two conditions. However, the overall ANOVA was not significant ($F(2,49) = 2.570$, $p = .087$, partial Eta-squared =.095).The individual level data shown in Figure 3b show a more nuanced picture than the means. Again, numerically, conditional entropy is lower in the IntTest condition than the IntLearn condition and slightly lower than the Control condition. The variability is higher in the Control condition than in the other two, (Control: SD = 0.317, 95% CI of the mean = 0.1985–0.5185, range of 0.32; IntLearn: SD = 0.2027, 95% CI of the mean = 0.3967–0.6127, range of 0.22; IntTest: SD = 0.2333, 95% CI of the mean = 0.1927–0.4248, range of 0.23), which is only partly a result of the outlier at 0.99. In summary, the manipulation that occurred during the learning phase of the experiment did not increase regularization. If anything, interference during learning led to less regularity in production: there was a slight reduction in the use of the main determiner forms or whatever the individual's most commonly produced form was (regularization score), and fewer individuals were consistent in their productions. Conditional entropy, however, was not higher in the IntLearn than the Control condition. The general pattern is consistent with the findings in Perfors (2012). However, the current data also show that the manipulation that occurred during the test phase increased regularization behaviour, when regularization was measured as consistency at a fine-grained level (consistency types and conditional entropy). Sometimes the differences between the IntTest and Control were statistically significant, but not always, however, the trend was always numerically present.

[7]This is actually only true for a certain kind of consistency. Specifically, this measure captures consistency in the relationship between individual nouns and determiners, and by extension, patterns where a single determiner is used for all nouns. It does not, however, capture patterns where determiner usage is based on things like grammatical categories. However, no participants in this study used an idiosyncratic pattern based on anything other than individual nouns, so it is adequate for these data.

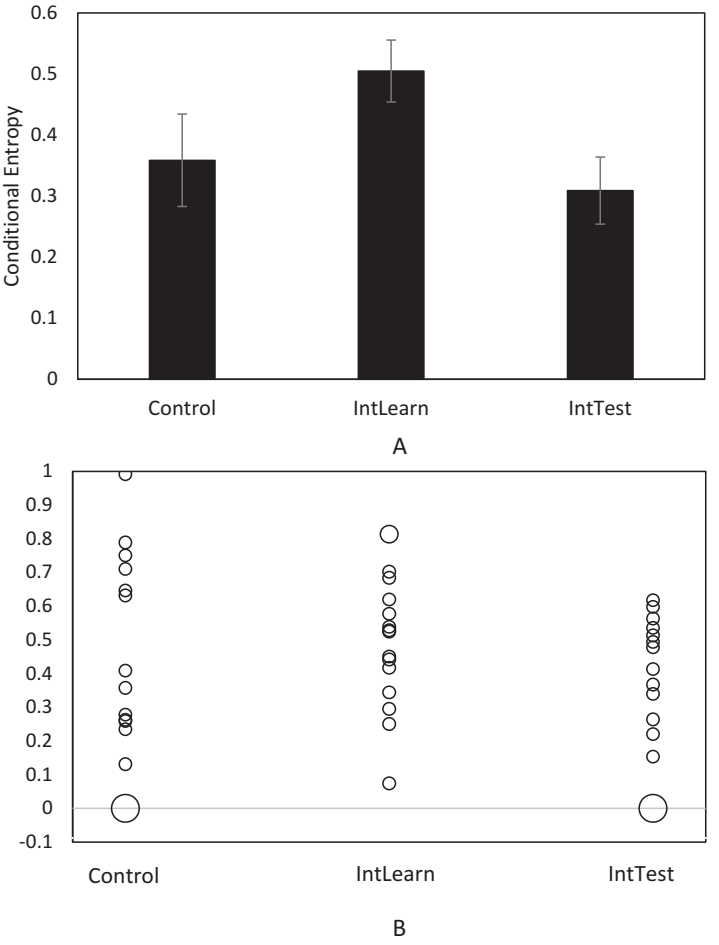


Figure 3. Mean average conditional entropy by condition (error bars=se) (a), Individual average conditional entropy by condition (b). (Circle area indicates the number of participants at that value: the smallest circles represent one participant, the middle-sized represent two participants, and the largest represent five).

Discussion and conclusion

In this study, I examined whether adult learners would show more evidence of regularization of variable aspects of language when language production was made more difficult, either by manipulating the learning conditions or by manipulating the production situation. Participants were exposed to a language with two main determiner forms, each consistently associated with a specific subset of nouns, alternating unpredictably with two noise determiner forms that occurred with nouns in both classes and were less frequent in the input. Hudson Kam and Newport (2009) found that this language lead to average main determiner production that was slightly higher than the input percentages – that is, it already elicited some regularization – suggesting that this language was a good candidate for inducing increased regularization by making production more difficult. In one experimental condition, IntLearn, participants had to perform a concurrent cognitive task during learning. The idea behind this is that it would make learning less effective, and lead to weaker or less stable representations, which in turn would make it more difficult for participants to access these representations and so lead to over-retrieval of some forms. In the other experimental condition, IntTest, the concurrent cognitive task occurred during testing, and so interfered with

retrieval in a much more direct fashion. Participants in the control condition were exposed to the same language but with no concurrent tasks at any time. The control condition served as a comparison group for the two experimental groups. In addition, it served as a replication of one condition in Hudson Kam and Newport (2009).

The data from the control condition replicated the basic findings from Hudson Kam and Newport (2009); participants produced the main determiner forms slightly more often than they had heard them on average. The prediction was that participants given either concurrent task should regularize more than participants in the control condition. The data, however, show that participants given a concurrent cognitive task during learning did not regularize more. Instead, they used the main determiner forms even less than the control participants. This is consistent with Perfors (2012). She had participants do a variety of concurrent tasks during learning and similarly found no evidence of increased regularization. She used a much simpler language, potentially less interfering concurrent tasks, and exposed people for a much shorter period of time, making regularization potentially less likely in her paradigm than the present one. However, the data from the present study replicate her findings, interference during learning does not lead to increased regularization and if anything, can lead to slightly less regularity in productions, whether regularity is measured in terms of the production of the main forms, patterns in the productions, regularization score, or conditional entropy.

Retrieval was manipulated much more directly in the IntTest condition, and this condition produced a different pattern of results than the IntLearn condition. The average main determiner production did not differ between the IntTest participants and the control participants; it was not higher, as was predicted, or lower, as was the case for the IntLearn (and Perfors', 2012) participants. However, there were more participants in the IntTest condition that imposed their own idiosyncratic rules on the language and numerically, they had lower entropy scores than participants in the other two conditions. Interference at the time of production does seem to have an effect on regularization behaviour, albeit a small one.

As pointed out by a reviewer, the two concurrent tasks are quite different from each other, potentially making it difficult to draw strong conclusions from the data. This was deliberate: I started by thinking of the primary tasks (learning and production), and selected secondary tasks that would interfere with the primary one in the right ways. For IntLearn participants, in any one trial their task is actually just to repeat the sentence they heard. Thus, their immediate primary task is quite simple. Ideally, however, the secondary task would interfere not with the ability to repeat the sentence, but instead, to think too deeply or consciously about what they were hearing, forcing more implicit learning. It was, therefore, intended to be quite taxing and involve various aspects of executive function and short-term memory. The demands of the primary task during testing are quite different. The participant has to watch the scene, listen to the word they are provided with, figure out what it means, recall the relevant nouns and any related determiners from memory, and then plan and produce the sentence. The secondary task needed to be one that would require participants to divert some cognitive resources, but not so much that production would be impossible, as without successful novel productions there would be no way to assess regularization. Therefore, I used a short-term memory task that was well within the capabilities of every participant were it the only task being done. The intention in using these two tasks was not to equate difficulty of the secondary tasks, it was to ensure that each task disrupted the primary task, but not so much that the primary task was impossible, nor so much that participants gave up on the secondary task. I have no independent measure that the two tasks did what they were intended to do, other than the data showing that most participants found them at least somewhat challenging and thus, that attention and memory resources were directed at the secondary tasks, to the potential detriment of the primary ones, leading to increased regularization of a particular type only in the IntTest participants.

Nevertheless, it is possible that it is the difference in tasks, rather than the difference in timing of the tasks, that is responsible for the differences we see between the productions of the IntLearn and IntTest participants. It is not possible to examine this with the current data, but data from Perfors

(2012) would suggest that this is not the case. She had participants do a variety of secondary tasks during exposure to the language and assessed whether and to what degree the interference led to increased regularization. Although none of her tasks were exactly the same as the ones used here, there were two that were very similar and we can look at how those tasks influenced learning and regularization. The first similar task was her "concurrent operational load" task. In that condition, participants were presented with an equation at the same time that they were presented with an image and its accompanying label, and had to solve the equation (quickly) before moving on to the next item. The second similar task was the "low concurrent load" task. In that condition participants first saw a list of three letters, then were presented with an image and its label. The task was to remember the letters, in order, and recall them after the image disappeared from the screen. The concurrent operational load task is very similar to the secondary task in the IntLearn condition, and the low concurrent load task is quite similar to the secondary task in the IntTest condition. Given the similarities between her two tasks and the two in the present study, we can use performance in her study as a gauge for the nature and degree of interference, at least as it is relevant for thinking about their impact on regularization.

So what was the influence of the two tasks in Perfors (2012)? Both had a negative influence on vocabulary learning. However, when the data from the latter vocabulary trials only was considered only the concurrent operational load condition was significantly different from the no load condition. Over time, performance of participants in the low concurrent load condition improved to the point where they were no longer significantly worse than participants in the no load condition. (This was not the only condition in which scores improved over time such that they were no longer significantly different from the no load condition by the end of the session, but it is the one that is relevant to the current study.) Thus, the vocabulary data show that the low concurrent load condition had less of an effect than the concurrent operational load task on learning. What was the effect of the two tasks on regularization behaviour? Perfors has several different analyses of production that assess regularization in different ways. First she looked at the proportion of responses that were of the participant's most frequent type, the Regularization index. Numerically, both relevant experimental groups had lower mean Regularization index values than the control condition (people with no secondary task during learning), and the low concurrent load condition was numerically lower than the concurrent operational load condition, although neither was significantly lower than the control condition. Perfors also classified participants in terms of their consistency according to whether they used a particular determiner form consistently. She analysed the data in several different ways, with different cutoff points for consistency (80% vs. 90%) and different inclusion criteria for participants and productions. Overall, she found that people who performed additional memory tasks during learning were not significantly more likely to be classified as consistent than people who did not do an additional task. Because the question at issue here is the relative impact of the two different tasks, however, the overall analysis is not the most relevant, rather, the difference between the concurrent operational condition and the low concurrent condition is what is important. Although it is unlikely that the difference is significant, numerically, there were more individuals classified as consistent in the concurrent operational condition than the low concurrent condition.[8]

Taken together, her data suggest that the concurrent operational task is slightly more disruptive to learning and non-significantly more likely to lead to regularization than the low concurrent task. Both conditions are often less regular than the control condition, although again, not significantly so. Given the similarities between the tasks used by Perfors and those in the present study, it therefore seems unlikely that it is differences in the task difficulty that are driving the difference in productions seen in the present study. In her study, the concurrent operational task caused more interference and therefore led to more (albeit only slightly more) regularization. Thus, if task differences were driving

[8]Perfors (2012) also presents a non-significant overall analysis of conditional entropy, but no as entropy data are displayed it is not possible to assess the numerical trends.

the results in the current study we would expect more regularization in the IntLearn condition than in IntTest condition, which is the opposite of what I found. It does appear then, that concurrent task timing was the driving factor in the differences between the IntLearn and IntTest conditions. Nevertheless, the fact remains that the effects of secondary task difficulty and secondary task timing are confounded in the present study, making it impossible to definitively locate the source of the difference in regularization between the IntLearn and IntTest conditions.

Although the possibility being tested here was that a task concurrent with learning might lead to regularization, the present data and those of Perfors (2012) suggest that it's the opposite – people given concurrent memory tasks during learning are less regular in their productions – and it is interesting to think about why. In the present study the major difference in productions between the IntLearn and Control and IntTest conditions is actually in their production of the noise determiners: the IntLearn participants are much more likely to produce noise determiners than either of the two other conditions (Control: Mean = 24.93%, SD = 15.36; IntLearn: Mean = 39.5, SD = 15.75; IntTest: Mean = 24.68, SD = 14.59). The overall ANOVA is significant ($F(2,49) = 5.166$, $p = .009$, partial Eta-squared = .174), as are the contrasts between IntLearn and each of the other two conditions (IntLearn vs. Control: $p = .008$; IntLearn vs. IntTest: $p = .007$). This suggests that participants' knowledge of the probabilities was less strong in the IntLearn than in the other conditions; that is, that, as opposed to weakening their lexical representations and thereby leading to retrieval issues, the interference either weakened the links between lexical representations such that the strength of the relationships became more opaque, or possibly left them less able to compute the probabilities in the first place. Interference during learning caused learning issues, but not necessarily representational issues. In any case, whatever the reason for the decreased regularity in production, this kind of interference does not lead to increased regularization.

This is in contrast to the effects in the IntTest condition, where participants were more consistent in their productions. This increased consistency was of a particular type however: participants in the IntTest condition were more likely to have noun-specific patterns. This is quite interesting when considered alongside Hudson Kam and Newport's (2009) results. They found that adults given two noise forms alternating with the two main forms, the same as the input in the present study, did not show any evidence of idiosyncratic rules, but that participants with increasing numbers of noise forms alternating with two main forms did sometimes produce these kinds of patterns. Thus, it seems that making retrieval harder at test has a similar effect as increasing the number of noise determiner forms. Exactly why this is the case is unclear. A strict retrieval hypothesis would predict that the main forms would be regularized under these circumstances: due to their frequency, they should be relatively easier to retrieve and so we should simply see an increase in the production of the main forms, not an imposition of rules/productions that seem like rules. (See Perfors, 2012, for a clear and detailed discussion, including modelling results, of other issues with the retrieval hypothesis.) So although the retrieval hypothesis predicts some of the data (e.g., the results of Hudson Kam & Chang, 2009), it may do so for the wrong reasons, or at least, it is not a complete explanation, a point made by Perfors (2012).

It may be the case that the form of initial productions has stronger implications for later productions when the participant is under pressure. That is, that whatever determiner happened to be produced with a particular noun the first time is more likely to be retrieved again later because that connection has been strengthened and the participant has less time or resources available to think consciously about producing a different pairing. Impressionistically, it is often clear that adults in these studies think about what they say before they say it. If they are busy holding some digits in mind, they are less able to be as conscious about their productions, resulting in the production of the easiest pairing. Given a production system with reinforcement of recently produced forms and structures, the pairing already produced will be the easiest. This can also explain children's greater likelihood of appearing to impose rules: they have worse executive function and so are less likely than adults to consciously reflect on the forms of their productions (Ramscar & Gitcho, 2007). On

this story, item-level regularization is not an imposition of rules, per se, or a difference in learning processes, but rather, the outcome of a production system operating under constraints.

While this can explain the adult regularization data, it is not necessarily what explains differences between adults and children. Although adults do have more executive function abilities and so are better at self-directed retrieval, children in artificial language studies do not generally impose noun-specific regularities (Hudson Kam, 2015; Hudson Kam & Newport, 2005, 2009). Thus, it appears that this cannot be the (only) difference between adults and children. Moreover, it leads to a transient effect – regularization that is not based on representations that are consistent, at least initially. But the cases of regularization in the real world that led to the laboratory investigations suggest long-term effects that are about knowledge differences, not simply momentary effects. This can be accounted for by different a general cognitive effect, entrenchment (Langacker, 1991). If a learner says something often enough in a specific way, that way will become entrenched in their memory. (This is a longer-term effect similar to what I have suggested might be responsible for noun-specific patterns.) And if the way they say it is consistent with exemplars in their input (which is the case when variability is present), then it can remain, unchallenged, in their representation of the language. When productions conflict with the actual patterns that exist in the language, as long as the true patterns are heard often enough they will win (see, e.g., Ambridge, Pine, Rowland, Chang, & Bidgood, 2013; Ramscar & Yarlett, 2007). If there are underlying patterns hidden in the variability (e.g., in the case of sociolinguistic variation), those too will be learned, but the more complicated the patterns are the longer learning will take (Schwab, Lew-Williams, & Goldberg, 2018; Shin, 2016). In this way, short term production effects can produce long term changes in a language, but they need not, and exactly how many positive exemplars consistent with a regularized production would be enough for the regularization to stick around in the language is not something the current data can speak to. In sum, the production effects investigated here cannot, by themselves, fully explain regularization.

There are several other suggestions for what causes regularization in the literature. Rische and Komarova (2016), for instance, suggest that adults and children differ in their propensity to regularize due to differences in sensitivity to evidence that is consistent with or conflicts with the learner's existing rule; children are more sensitive to positive exemplars than adults and adults are more sensitive to negative evidence than children. They model several different experimental outcomes and manage to replicate several different patterns in the human experimental data, including children's greater propensity to appear to invent rules that do not just involve overuse of the most frequent form (by including noise). Their model differs from what I am suggesting in that it is about differences in what adults and children extract or retain from the input. Note that, although both the present work and Perfors (2012) demonstrate that manipulating learning fails to increase regularization by adults which would seem to contradict Rische and Komarova's proposal, this is not the case. Differences in sensitivity to consistent or conflicting information are not proposed to be dependent on working memory or executive function, which is what both I and Perfors (2012) manipulated. Therefore, manipulations that affect these more general cognitive functions should not affect regularization under their model. However, it is not clear how their model can account for the fact that both children and adults who regularize show sensitivity to the underlying probabilities, even when they do not seem to guide productions, as has been found in several studies (Ferdinand et al., 2017; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Schwab et al., 2018).

Perfors (2016) made a suggestion about regularization that is unrelated to memory or learning differences and which is quite interesting. She flipped the question around: instead of asking why children (and adults under some conditions) regularize, she asked why adults probability match, taking regularization (resulting from a combination of memory limitations and an inherent bias to regularize, Perfors, 2012) as the basic result. She proposed two pragmatic assumptions that

people have that might lead to probability matching, namely, a bias to think that variability is predictable and an assumption that the goal of learning is to correctly learn the language, and so, the variation. She ran two experiments using the same learning task as in Perfors (2012) but with different instructions. In one she told some participants that the training data they were receiving came from a previous participant and that it might contain some errors. In the other experiment she put people in pairs (who did not communicate with each other) and varied whether participants thought the goal was for both of them to be as correct as possible or as similar to each other as possible. She found that adults were more likely to regularize (and so less likely to probability match) when they thought the variation was not systematic (i.e., due to errors) and when they thought the goal was to be as similar to the other person as possible, consistent with her hypothesis. Although very interesting, the meaning proposal cannot explain why regularization is affected by the number of noise determiners (as in Hudson Kam & Newport, 2009), why it decreases when production constraints are decreased even given a large number of noise determiners (Hudson Kam & Chang, 2009), or why it increases (albeit only slightly and in a specific way) when production is made more difficult, as in the IntTest condition. Nor can it explain why children, who have been shown to strongly expect that language forms are shared (Sabbagh & Henderson, 2007) regularize more than adults (Hudson Kam & Newport, 2005, 2009), or why adults' behaviour is exactly the same when the variation is versus is not predictable, in contrast to children who are less likely to regularize contextually conditioned variation (Hudson Kam, 2015). Thus, at this point, we are left with several explanations, none of which can explain all of the existing data.[9]

In summary, the data in the current paper suggest that differences in learning efficacy that result in differences in representational strength of individual items are unlikely to be the or a cause of regularization. This concurs with the findings in Perfors (2012). The data also suggest that general cognitive constraints acting at the time of production might explain at least some aspects of regularization. It is encouraging that the data replicate the findings of previous studies when there is overlap in the tasks, especially given the sample size, leaving one with greater confidence in the novel findings. Although the intent of this study was to shed some light on the mechanisms responsible for regularization, it is also relevant to thinking about regularization of real languages. For this, the degree of difference between the conditions is relevant. The differences between the IntTest condition and the control condition are not always significant, and even when they are, they are not dramatic: they are small differences, a only few people doing different things from the others. It might appear that such small differences are not really relevant for instances of real languages and how they change. However, studies of regularization using iterated learning techniques show that small differences can very quickly turn into large ones over time (e.g., Reali & Griffiths, 2009; Smith & Wonnacott, 2010). In this way, small differences in regularization behaviour at an individual level can have broader implications for regularization in real language learning situations, and so the small differences found in the present study can still have relevance for understanding the real world phenomena that inspire this work.

Acknowledgment

Thanks to Allison Kraus, Whitney Goodrich Smith, and Amy Finn for their assistance conducting this research.

[9]Something that none of the proposals can explain are individual differences. In a study examining factors affecting probability learning in linguistic and non-linguistic stimuli, Ferdinand et al. (2017) found that individual differences in input could explain some of the variation between individuals. As all of the participants in the present study received exactly the same input, that cannot explain the inter-participant differences in my data. However, it does suggest that we might be able to find differences related to things like attention to specific exemplars in the input or learning rate (which could affect the effectiveness of specific exemplars). But what causes differences in attention or learning rate is another question. It is clear from Perfors' (2012) (as well as the present study, see Appendix A) that the underlying causes are unlikely to be found in working memory differences.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This research was supported by National Institutes of Health Grant HD 048572 to Carla L. Hudson Kam.

ORCID

Carla L. Hudson Kam http://orcid.org/0000-0002-7638-3656

References

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *WIREs: Cognitive Science*, *4*, 47–62. doi:10.1002/wcs.1207

Becker, A., & Veenstra, T. (2003). The survival of inflectional morphology in French-related creoles. *Studies in Second Language Acquisition*, *25*, 283–306. doi:10.1017/S0272263103000123

Bock, K. (1995). Sentence production: From mind to mouth. In L. Joanne & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 181–216). New York, NY: Academic Press.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*, 2037–2078. doi:10.1016/j.neuropsychologia.2006.02.006

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*, 614–616. doi:10.1038/nature01951

Ferdinand, V., Kirby, S., & Smith, K. (2017, March 9). *The cognitive roots of regularization in language*. Retrieved September 12, 2018 from the arXiv database. arXiv:1703.03442v1 [cs.CL].

Finn, A. S., Lee, T., Kraus, A., & Hudson Kam, C. L. (2014). When it hurts (and helps) to try: The role of effort in language learning. *PLoS ONE*, *9*(7), e101806. doi:10.1371/journal.pone.0101806

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science*, *17*, 670–674. doi:10.1111/j.1467-9280.2006.01764.x

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.

Hudson Kam, C. L. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development*, *5*, 115–145. doi:10.1080/15475440902739962

Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language*, *91*, 906–937. doi:10.1353/lan.2015.0051

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of regularization in adults: Memory constraints or learning differences? *JEP: Learning, Memory, and Cognition*, *35*, 815–821. doi:10.1037/a0015097

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195. doi:10.1080/15475441.2005.9684215

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*, 30–66. doi:10.1016/j.cogpsych.2009.01.001

Langacker, R. W. (1991). *Foundations of cognitive grammar. Vol. 2, descriptive application*. Stanford, CA: Stanford University Press.

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control. *Current Directions in Psychological Science*, *21*, 71–77. doi:10.1177/0963721412436807

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*, 486–506. doi:10.1016/j.jml.2012.07.009

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*, 138–155. doi:10.1080/15475441.2015.1052449

Pitts Cochran, B., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, *41*, 30–58. doi:10.1006/jmla.1999.2633

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*, 274–279. doi:10.1016/j.tics.2007.05.007

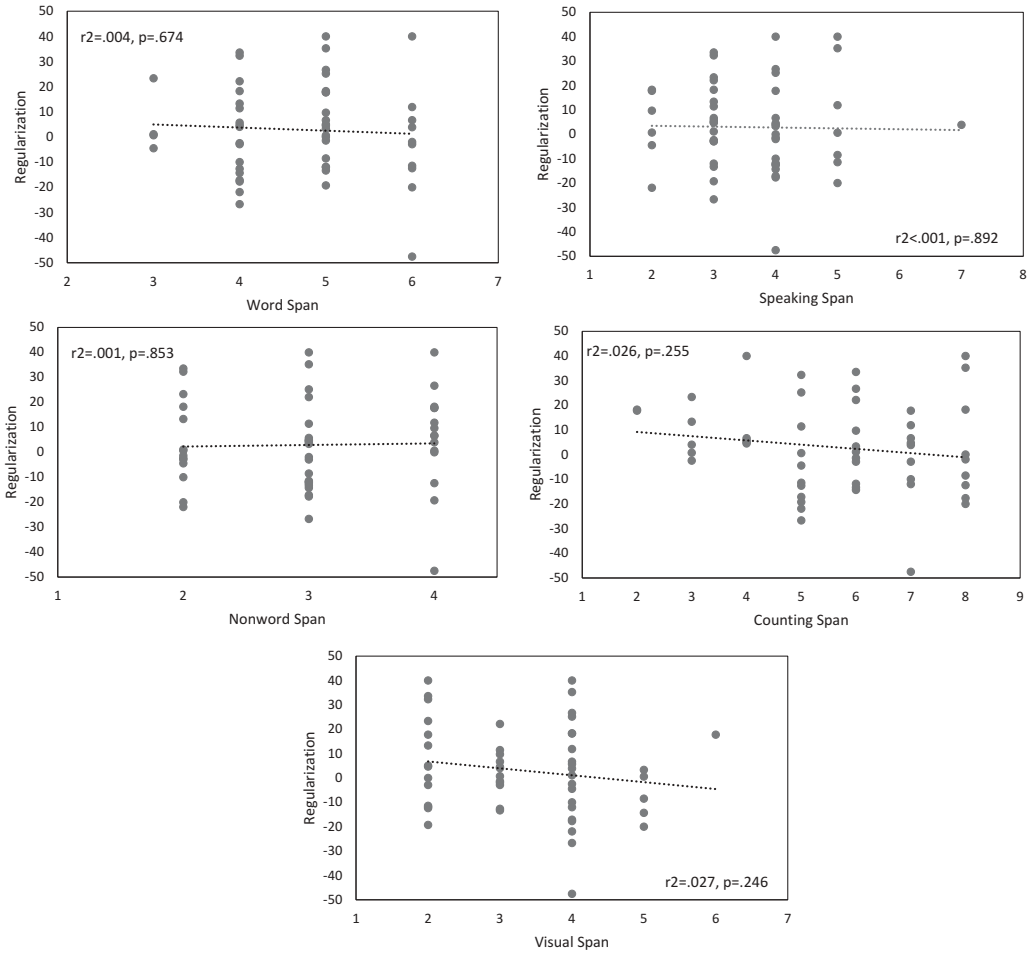
Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*, 927–960. doi:10.1080/03640210701703576
Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328. doi:10.1016/j.cognition.2009.02.012
Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1–30. doi:10.1016/j.cogpsych.2015.10.001
Ross, D. S., & Newport, E. L. (1996). The development of language from non-native linguistic input. In A. Stringfellow, D. Cahana-Amitay, E. Hughs, & A. Zukowski (Eds.), *Proceedings of the 20th annual Boston University conference on language development* (pp. 623–645). Somerville, MA: Cascadilla Press.
Ross, D. S. (2001). *Disentangling the nature–Nurture interaction in the language acquisition process: Evidence from deaf children of hearing parents exposed to non-native input*. (Doctoral dissertation, University of Rochester, 2001). *Dissertation Abstracts International, 62-07B*, 3402.
Sabbagh, M. A., & Henderson, A. M. E. (2007). How an appreciation of conventionality shapes early word learning. *New Directions for Child and Adolescent Development*, *115*, 25–37. doi:10.1002/cd.180
Schwab, J., Lew-Williams, C., & Goldberg, A. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language*, *45*, 1054–1072. doi:10.1017/S0305000918000041
Senghas, A., & Coppola, M. (2001). The creation of Nicaraguan Sign Language by children: Language genesis as language acquisition. *Psychological Science*, *12*, 323–328. doi:10.1111/1467-9280.00359
Shin, N. L. (2016). Acquiring constraints on morphosyntactic variation: Children's Spanish subject pronoun expression. *Journal of Child Language*, *43*, 914–947. doi:10.1017/S0305000915000380
Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from impoverished input. *Cognitive Psychology*, *49*, 370–407. doi:10.1016/j.cogpsych.2004.05.001
Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–449. doi:10.1016/j.cognition.2010.06.004
Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science*, *18*, 259–263. doi:10.1111/j.1467-8721.2009.01648.x
Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*, 195–217. doi:10.1006/jmla.1998.2574
Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, *58*, 221–228.
Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. R. Clark-Cotton, & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development* (pp. 663–673). Somerville, MA: Cascadilla Press.

Appendix A

Like Perfors (2012), I suspected that it might be the case that the concurrent tasks were only interfering enough to induce regularization for some participants, e.g., those with lower working memory capacity. To check for this possibility, I collected a variety of working memory task data from participants in a single session preceding the MAL study. At first, all people who completed the memory testing went on to participate in the MAL study. Later on, however, in an effort to include more variation in memory performance, only people with more extreme scores on the memory tests were invited to continue in this study. (Others were invited to participate in a different MAL study.)

The five memory tests used were the following: a standard word span task (using sets of 2 syllable, non-semantically related words), a non-word task (using 2 syllable non-words that were not part of the MAL), a speaking span task in which participants were given a list of bisyllabic words and asked to produce a sentence using each one, a counting span task (based on Towse, Hitch, & Hutton, 1998; participants saw a visual array consisting of two kinds of objects, had to count one kind, and remember the count as this process was repeated), and a visual span task. In this final task, participants were shown a series of spatial arrays, one by one. The arrays comprised a 3×3 grid of squares, with some squares filled and others left unfilled. Each array was a unique pattern. After presentation of the arrays, participants were shown a panel containing multiple arrays and asked to indicate, by pointing, the arrays that they had just seen, in order. Two similarly designed distractors were always included in the test panels. Although this task is not a verbal working memory task, it shares the aspect of serial presentation with the verbal tasks, something very relevant to language, and thus might be expected to correlate with performance on the MAL task.

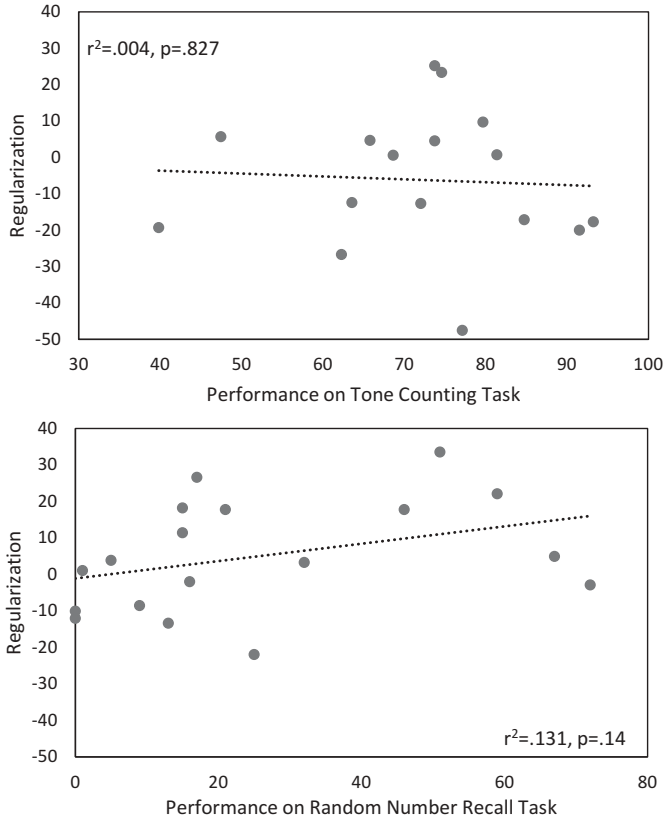
Like Perfors (2012), who used two memory assessments, I found no evidence of a relationship between participants' propensity to regularize (as measured by increased production of the main determiners) and their working memory abilities. The relationship between performance on the various WM measures and regularization (percent main determiners produced – input percentage) is shown in the figures below, including the r^2 values for each data series. I also computed r^2 values using deviations from the input percentage for main determiners (|percent main determiners produced-60|) as the measure of regularization instead, and the picture changed only slightly: they ranged from .027 (for word span) to < .001 (for visual span). The main take-away from these analyses is that these measures of WM are not directly related to regularization in adults.



Appendix B

The concurrent tasks were never intended to be an individual-level predictor of regularization performance; although it seems that people who are most affected by the concurrent task might be expected to regularize more, this is too simple an expectation. What would it mean to be most affected? It could mean that the participant found the concurrent task so overwhelming that they failed at both tasks, or that they stopped trying on one of the tasks so that they could focus exclusively on the other, or it could mean that they performed at only a middling level on the concurrent task. Thus, it is not clear what kind of performance on the interfering task would be expected to correlate with regularization behaviour. Nevertheless, from an exploratory perspective some might be interested in the correlations. The figures below show the relationship between regularization (measured in terms of overproduction of the main determiners) and performance on the concurrent tasks. Keep in mind that this is only one measure of regularization, and not the one that showed the predicted effects. The data are presented and analyzed separately for the two conditions because the two concurrent tasks are quite different. There is no standardized scale we can use for both of them (indeed, in one, higher scores are good and in the other they are bad, although we could easily adjust for that difference).

R^2 values for each set of data are shown on the figure, along with the associated *p* value. The relationship between concurrent task performance and regularization is stronger for the IntTest condition than the IntLearn condition, which is consistent with the overall condition differences. However, neither relationship is very strong (keeping in mind that these are small ns to be performing correlation analyses on).



Appendix C

The coding scheme for the random number concurrent task was based on the following reasoning: recalling more numbers is better than recalling fewer, knowing when you don't know something is better than not (so leaving out a number is better than reporting an incorrect number), and getting things in the right order is better than getting them in the wrong order. This lead to the following coding scheme.

0 – all three numbers reported correctly (the right numbers in the right order)
1 – all three numbers reported, but in the wrong order
2 – 2 numbers correctly reported, 1 missing, recalled numbers in the correct order
3 – 2 numbers correctly reported, 1 missing, order incorrect
4 – 2 numbers correctly reported, 1 number incorrectly recalled, correct order
5 – 2 numbers correctly reported, 1 number incorrectly recalled, order incorrect
6 – 1 correct number, 2 missing (order was not relevant here, as we did not ask people to report the serial location of the number they recalled)
7 – 1 correct number, 2 incorrect numbers reported, the correct number in the correct location
8 – 1 correct number, 2 incorrect numbers reported, order incorrect
9 – nothing correct (this last category includes a variety of response types, i.e., no numbers recalled or some number of incorrect numbers reported)