Journal of Memory and Language 67 (2012) 486–506

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

When do memory limitations lead to regularization? An experimental and computational investigation

Amy Perfors

School of Psychology, University of Adelaide, Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 1 December 2011
revision received 5 July 2012
Available online 29 August 2012

Keywords:
Regularization
Less is More
Computational modeling
Language acquisition

ABSTRACT

The Less is More hypothesis suggests that one reason adults and children differ in their ability to learn language is that they also differ in other cognitive capacities. According to one version of this hypothesis, children's relatively poor memory may make them more likely to regularize inconsistent input (Hudson Kam & Newport, 2005, 2009). This paper reports the result of an experimental and computational investigation of one aspect of this version of the hypothesis. A series of seven experiments in which adults were placed under a high cognitive load during a language-learning task reveal that in adults, increased load during learning (as opposed to retrieval) does not result in increased regularization. A computational model offers a possible explanation for these results. It demonstrates that, unless memory limitations distort the data in a particular way, regularization should occur only in the presence of both memory limitations *and* a prior bias for regularization. Taken together, these findings suggest that the difference in regularization between adults and children may not be solely attributable to differences in memory limitations during learning.

Introduction

In many ways, ranging from phonetic perception to aspects of syntax, children are superior language learners than adults. Adults have difficulty with many aspects of language acquisition, from phonetic perception (Kuhl, 2004; Werker & Lalonde, 1988; Werker & Tees, 1984), to language processing (Clahsen & Felser, 2006), to certain aspects of syntax (e.g., Birdsong, 2006; Johnson & Newport, 1989; Johnson, Shenkman, Newport, & Medin, 1996). Scientists have proposed many theories to account for the difference between children and adults; these theories differ in both the degree and type of contribution made by preexisting language-specific biases.

Some argue that language acquisition in children is guided by language-specific acquisition procedures, whereas adult acquisition is directed by more domain-general learning mechanisms (e.g., Bley-Vroman, 1990). However, there are many other possibilities, since children and adults also differ profoundly in their cognitive capabilities, knowledge, assumptions, and typical linguistic input. Learning a second language is made more difficult by interference from the first language (e.g., Hernandez, Li, & MacWhinney, 2005; Iverson et al., 2003; Mayberry, 1993; Tan, 2003; Weber & Cutler, 2003). Adults and children also differ in the plasticity of their brains (Elman et al., 1996; MacWhinney, 2005), their style of learning (Ullman, 2004), and the nature of the social support (Snow, 1999) and linguistic input (Fernald & Simon, 1984) they receive.

One hypothesis, often called Less is More, suggests that the relative cognitive deficits in children may actually *help* with language acquisition. There are several versions of the Less is More hypothesis. The most general suggests that "starting small" – whether via restricted input, or as a byproduct of limited cognitive capacity – can help a learner to isolate and analyze the separate components of a linguistic stimulus (Newport, 1988, 1990). There is considerable support for this version of the hypothesis.

E-mail address: amy.perfors@adelaide.edu.au

0749-596X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2012.07.009

Cochran, McDonald, and Parault (1999) taught adults a novel sign language normally as well as under conditions of memory load. They found that adults who were not under memory load learned faster but made more errors caused by producing signs holistically, rather than analyzing the individual components. In a similar study, Kersten and Earles (2001) found that adults taught a miniature artificial language learned better if they were first presented with individual words and only later presented with complex sentences. Starting small has also been found to help adults learning recursion in artificial grammars (Lai & Poletiek, 2010) and foreign natural languages (Chin & Kersten, 2010). In addition, Elman (1993) discovered that neural networks could be trained to process complex sentences, but only if, during the initial stages of training, the network had limited memory and was given limited input (though other modelers have found different results; see, e.g., Rohde & Plaut, 1999).

Another version of Less is More suggests that limited capacity may lead children to regularize inconsistent input (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005, 2009). Regularization may be a beneficial strategy when the variability in the observed forms is not conditioned on a previous linguistic context. Unpredictable variation of this sort is not commonly found in most languages, at least in the speech of native speakers (e.g., Chambers, Trudgill, & Schilling-Estes, 2003); however, it is much more common when learning from non-native speakers (Wolfram, 1985; Johnson et al., 1996). In such circumstances, when the input is *truly* inconsistent, regularization can be beneficial.

It is precisely in those circumstances that regularization is often observed. Deaf children produce regular grammatical forms despite being exposed to the inconsistent sign language of their hearing parents (Singleton & Newport, 2004), as will children exposed to inconsistent input in an artificial language (Hudson Kam & Newport, 2005, 2009). Childrens' tendency to regularize may even lead to the creolization of initially inconsistent languages (Senghas & Coppola, 2001). By contrast, adult language learners are known to produce highly variable, inconsistent utterances, even after years of experience with the language and after their grammars have stabilized (Johnson et al., 1996).

This difference between children and adults has also been found in non-linguistic domains. If adults must predict some phenomenon, like a light flashing or a certain card being drawn from a deck, in most circumstances they will tend to probability match: if the phenomenon occurs 70% of the time, they will predict it 70% of the time, even though predicting it 100% of the time would result in more correct predictions (see for an overview Myers, 1976). Children are more likely to predict that the phenomenon will occur closer to 100% of the time (e.g., Derks & Paclisanu, 1967; Weir, 1964), although many still do not. A similar pattern has been found in causal reasoning: children regularize by assuming that causes are deterministic, while adults do not (Schulz & Sommerville, 2006).

Although children's tendency toward regularization is fairly well-established, the reason for the difference between adults and children is less clear. The Less is More hypothesis suggests that regularization may be due to some limitations on children's cognitive capacity, but exactly *how* and *why* such limitations should lead to regularization is somewhat underspecified. Memory is often identified as a possible culprit. For instance, in one of the clearest statements of how Less is More relates to regularization, Hudson Kam and Newport (2009, p. 61) suggest that "one possibility is that children are worse at directed memory search than adults. Another possibility is that children are less efficient at laying down memory traces, with the consequence that they have more difficulty retrieving specific forms (therefore especially those that are lower in frequency or less broadly or consistently used)".

This clearly identifies memory as a potential issue, but the precise details are still unclear: what predicts which forms should be hard to lay down or access, and why? Is the issue with encoding, storage, retrieval, or all of the above? Are the relevant limitations in working memory, short-term memory, long-term memory, or an interaction between them? What specific model of memory is being assumed, where does the limitation lie, and why would that limitation result in regularization?

Hudson Kam and Chang (2009) are slightly more specific, suggesting that memory limitations may interfere with accurate retrieval (although not ruling out the possibility that differences during encoding might also have an effect). Consistent with this, they found that adults who were given cues that made retrieval easier ended up probability matching more precisely than adults who were not given such cues. However, the details of how and why retrieval limitations should lead to regularization are still somewhat underspecified. The most precise explanation suggests that "when retrieval is difficult, the most easily accessible form is likely to be retrieved repeatedly, resulting in regularization." (page 816) However, why this is likely or what assumptions about memory should lead to this are not made clear.

This relative lack of detail is natural given the newness of the Less is More hypothesis as regards regularization. However, it does mean that many open questions remain. If regularization is due to limitations on memory, under what circumstances and for what assumptions about memory might we expect it to occur? Should we expect it to be limited to retrieval, or might effects arising during learning – while information is being processed, encoded, and stored – matter as well? If so, why, and under what circumstances?

These questions lead to the two main goals of this paper. The first goal is to empirically explore whether and to what extent memory limitations during encoding might affect regularization. This has not been investigated before and is an open question. To that end, the first section reports the result of seven experiments in which adults were placed under memory load while simultaneously learning a simple artificial "language" composed of nouns paired inconsistently with determiners. This load, which occurred during the encoding phase of language learning, did not increase regularization in any of the conditions.

The second goal of the paper is make precise – and then systematically investigate – a range of possibilities about *how* and *why* memory limitations during learning could affect the generalizations (and hence the extent of regularization) of the learner. This is accomplished with a computational model that examines a variety of different theories about how memory limitations during learning might affect the pattern of data available to the learner. The modeling also explores how these memory limitations might interact with prior biases for or against regularization. Results indicate that regularization only occurs when *both* memory limitations *and* a prior bias for regularization are present; neither alone is sufficient. Regularization can only occur without a prior bias if the memory process itself distorts the pattern of data available to the learner in a particular way, which does not at present appear to correspond to any well-established models of memory encoding.

Taken together with the experimental findings, these results suggest that adult-child differences in regularization probably do not emerge from differences in memory limitations during encoding. I conclude the paper by discussing other possibilities, among them that adults and children have different prior biases about how to respond to inconsistent input; that regularization is caused by limitations in other cognitive capacities; and that any effects due to memory are more likely to occur during retrieval rather than encoding.

Experiments

Method

179 English-speaking adults were recruited from the University of Adelaide and surrounding community. They were paid $10 or received course credit for their participation; of these, four were excluded due to equipment failure (2) or a refusal to say anything during the word-learning task (2). Thus there were 25 participants in each of the conditions. All conditions within the experiment consisted of two parts, both presented on a computer in the Computational Language and Cognition Lab at the University of Adelaide. The experiment was programmed in MATLAB and auditory stimuli were presented over headphones.

In the first part of the experiment, individual differences in working memory capacity were estimated using a standard complex span task (Conway, Jarrold, Kane, Miyake, & Towse, 2007; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). In the second part of the experiment, subjects completed a word-learning task modeled on the paradigm described by Hudson Kam and Newport (2009) in which they were taught 10 two-word labels from a new language. Interspersed with the word-learning task, participants in six of the seven conditions completed an interference task designed to tax their working memory; these conditions will be described in detail later. In a control condition, the NO LOAD condition, participants performed the word-learning task only.

Complex span task

Complex span tasks are widely used to measure the capacity of the working memory system (Conway et al., 2005; Unsworth et al., 2009). In a complex span task, items to be remembered (e.g., random letters, digits, shapes, or spatial locations) are interspersed with an unrelated cognitive activity (e.g., solving equations, reading sentences, or evaluating the symmetry of patterns). After several trials, participants are asked to recall the items to be remembered in the correct serial order. This sort of task is differentiated from a simple span task (e.g., Digit Span from the Wechsler scales), which only includes the memorization component. It has been argued that complex span tasks provide a measure of working memory, as opposed to span memory, because they entail the requirement to process as well as to store information, although both types of task provide measures of memory capacity and maintenance. Complex span tasks have good internal consistency (Conway et al., 2007; Kane et al., 2004) and test–retest reliability (Klein & Fiss, 1999). They have been shown to correlate with cognitive processes that are believed to depend on working memory (Conway et al., 2007; Unsworth & Engle, 2007), and are linked to disorders including Alzheimer's disease (Rosen, Bergeson, Putnam, Harwel, & Sunderland, 2002) and schizophrenia (Stone, Gabrieli, Stebbins, & Sullivan, 1998). They have also been widely used to explore age differences in working memory capacity (Case, Kurland, & Goldberg, 1982; Salthouse & Babcock, 1991).

Two common span tasks incorporate demands on either operational span (Turner & Engle, 1989) or on verbal span (Daneman & Carpenter, 1980). In an operational span task, participants are presented with equations such as $4/2 + 2 = 3$ and told to say, as quickly as possible, whether the equation is correct. In a typical verbal span task, subjects are presented with an 11–15 word sentence and told to say, as quickly as possible, whether the sentence makes sense. In order to enable comparison across participants, in the first part of the experiment all participants were presented with an operational span task regardless of condition. On each trial people first saw an equation and were asked whether it was correct or not. After each response, a random letter was shown. At the end of a set of n letters, participants were asked to repeat the list of letters in order, given unlimited time to do so. To make sure that they understood the task, they were first trained on two sets of two trials each. The full task comprised two sets each of sizes ranging from an n of three to an n of seven, for a total of 50 trials. For each participant a working memory capacity score was calculated, reflecting the number of correct letters recalled in the correct position.

Word-learning task

After the complex span task, all participants took part in an artificial language learning task modeled after a similar task described by Hudson Kam and Newport (2009). Their language contained 51 words, including 36 nouns and 12 verbs, among other lexical items, taught over the course of eight separate sessions extending for 9–12 days. Of critical interest in their study was the production of the determiners, which were associated with nouns in an inconsistent fashion: participants heard the main determiner only 60% of the time. In one condition, they heard nothing the other 40% of the time; in other conditions, they heard increasingly more **noise** determiners: for instance, two determiners (each 20% of the time), and so forth up to 16 determiners (each 2.5% of the time). Performance was measured in a sentence completion task in which participants had to provide the noun and determiner associated with a scene after being prompted with the beginning part of the sentence (the verb).

The present research sought to remove extraneous elements of the task so as to focus only on the production of the inconsistent input. Participants were therefore presented with a "language" of 10 items that for simplicity I will call "nouns", all two-syllable nonsense words[1] mapped to images representing common objects.[2] Each noun was followed by a one-syllable consonant–vowel–consonant (CVC) "determiner": the **main** determiner occurred 60% of the time with each noun, and each of the four **noise** determiners occurred 10% of the time with each noun.[3] The distribution of determiners across items thus precisely matched the global distribution of determiners, making the determiners completely inconsistent. After seeing the image-label pairs, participants were asked to produce novel labels of their own; the key question was whether they would regularize by producing a determiner (or no determiner) close to 100% of the time, rather than the amount it occurred in the input. The specific details of which word mapped to which meaning and which determiner was the **main** determiner were randomized for each participant. Participants were not told there were two "parts" (noun and determiner) to each label, and since the labels were presented orally this was not made obvious through visual presentation.[4]

All labels were recorded by a female speaker on a Windows computer from mono input using the software program Cubase. In order to ensure that the speech was as natural as possible, each noun was recorded with each determiner, rather than recording them individually and playing them one after another, which would produce odd coarticulation effects. Each label was spoken clearly and with normal intonation at the pace of standard adult-directed speech.

Over the course of the task, participants saw 200 trials of image-label pairs; since there were 10 labels and 10 objects, participants saw each possible object-label pair 20 times during the experiment. There were 10 different image tokens for each object (e.g., 10 different pictures of babies) and each image appeared once in the first 100 trials and once in the second 100 trials. On each trial, an image appeared on the computer screen and, at the same time, the person heard a female voice provide the label: for instance, participants might see a picture of a baby and hear the words churbit mog.

[1] Noun words used were: dragnip, raygler, churbit, tramdel, shelbin, pugbo, wolid, foutray, nipag, and yeetom.

[2] Objects used were: babies, balls, beds, birds, books, cars, cats, cups, dogs, and shoes.

[3] The five determiners were: mot, ped, sib, kag, and zuf.

[4] It is not critical that these be called nouns and determiners, respectively; in fact, since the determiners lack any semantics, it might be more appropriate to call them particles. However, for clarity, I chose to follow the convention established by Hudson Kam and Newport (2009). The critical point is that part of the label (the "noun") maps consistently to the image, while the other part (the "determiner") does not; we are interested in how participants respond to this sort of inconsistent input.

In the NO LOAD condition, participants went to the next trial by clicking a Next button. In the load conditions, explained below, participants had to perform additional tasks interspersed with the word learning. In all conditions, learning was tested with 10 questions every 50 trials, for a total of 40 test questions. At each test, the participant was presented with a novel image and asked to verbally produce the label for it, which the experimenter wrote down as accurately as possible. No feedback was given, and the experimenter was blind to the correct mapping of labels and objects for that participant.

Conditions

There were six different load conditions, described below and illustrated in Fig. 1. A wide range of conditions were investigated in order to thoroughly explore the space of possible ways that memory could be tasked in a word-learning experiment such as this. The goal was to be as certain as possible that at least some of the load tasks substantially taxed working memory while still allowing some learning.

In all conditions the adults were told that we were interested in how well people could learn words when the task was difficult. Thus, while they were learning the words, they would be asked to do something else (the details differed by condition as explained below). Participants were informed that they would be tested on their understanding of the language by the experimenter every 50 trials. They were also told that would be expected to simultaneously do as well on the load task as possible and would be given feedback throughout on their performance on it. It was acknowledged that high performance in *both* tasks might be difficult, but we were interested in how well they could achieve this.

The first two load conditions were modeled after the operational and verbal span tests used to measure working memory. These conditions taxed load by interspersing word-learning trials with items from these working memory tasks.

Verbal load. This task was modeled after the verbal span test of Daneman and Carpenter (1980). After each image-label pair, participants were presented with an 11–15 word sentence, told to read it aloud, and then asked to respond as quickly as possible whether it was sensible or not. Half of the sentences were sensible, and half were made non-sensible by replacing a content word with a semantically inappropriate one. For example, a typical sentence is "Cats really love to sit in the sun, since they are desert animals" while the corresponding non-sensible sentence would replace animals with chimneys. No participant saw both the sensible and non-sensible version of a sentence. Accuracy and elapsed time was displayed in order to encourage peak performance.

Operational load. This condition was modeled after the operational span test of Turner and Engle (1989). After each image-label pair, participants were presented with an equation and told to respond as quickly as possible whether it was correct or not. Half of the equations were correct, and half gave an answer that was one digit away

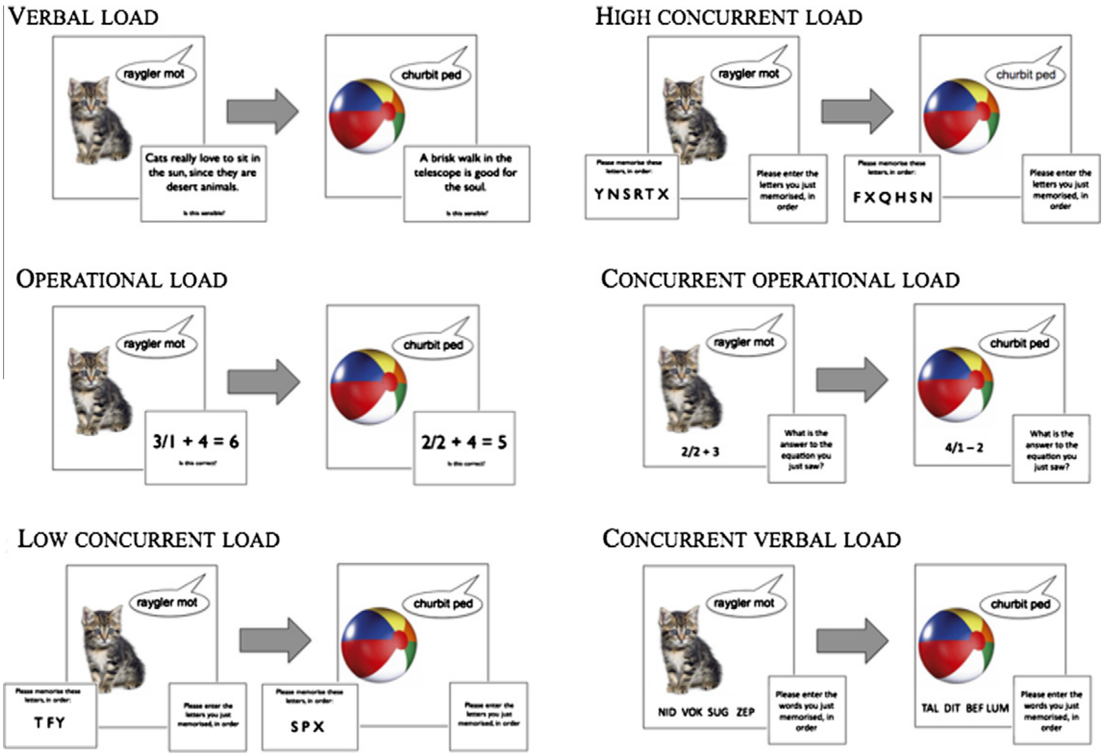

Fig. 1. Schematic illustration of the six load conditions. Each box represents a different screen seen by the participant. In all conditions, participants were presented with the word learning trials. The load conditions differed according to the task required of the participant either in between or concurrently with the word learning task.

from correct. In order to encourage participants to be as fast and correct as possible, a running total of their cumulative number correct and elapsed time was displayed on the screen.

These two load conditions have the advantage that they are modeled after tasks designed to load on working memory, but they have the disadvantage that they are interspersed with the word learning task rather than concurrent with it. It is therefore possible that they might not interfere enough with concurrent working memory to have an effect on the pattern of word learning. The next two conditions address this possibility.

Low concurrent load. In this condition, each image was preceded by a list of three letters to memorize, randomly selected from the following set of letters: F, H, J, K, L, N, P, Q, R, S, T, and Y. No single list contained the same letter twice. After viewing the list for 2.5 s, the image was displayed for 1.5 s while the label was heard. This was followed by a response phase in which participants reported the last set of letters in order. At that point memorization accuracy and the time taken to respond were displayed, in order to encourage participants to continue responding quickly and accurately. When the participant pressed Next the next set of letters to memorize appeared.

High concurrent load. This condition is identical to the LOW CONCURRENT LOAD condition except that participants were presented with lists of six rather than three letters. The list was visible for the same duration and subject to the same constraints, and the procedure by which list and image-label presentation were combined was also identical.

Although the two concurrent load conditions are specifically designed to tax concurrent working memory, there is still one shortcoming: the word learning task is linguistic, and this type of load may not provide the best conflict with a language-based task. Therefore, two additional conditions were added that were designed based on the literature investigating what kinds of tasks disrupt linguistic processing.

Concurrent operational load. In a study investigating the extent of the domain-specificity of the verbal working memory resources used during linguistic processing, Fedorenko, Gibson, and Rohde (2007) discovered that linguistic processing is disrupted by tasks that involve arithmetic integration. In their experiments, participants read sentences of varying complexity while simultaneously solving equations. Sentences were presented phrase by phrase, and each phrase was paired with part of an equation; participants were expected to parse the sentence while maintaining a running total for the equation. This task is quite different from the word-learning task in this study, since the Fedorenko et al. (2007) study was focused on information integration during the course of reading a complex sentence. However, because it reported a load task that demonstrably did disrupt some aspect of linguistic processing, I thought it valuable to include a condition that mimicked that load task as closely as possible.

Thus, in the CONCURRENT OPERATIONAL LOAD condition, participants were presented with an equation to solve *at the same time* as they were shown each image-label pair. Each equation was similar to those in the complex span task, and consisted of one term involving a simple multiplication followed by another term involving addition or subtraction (e.g., 6/6 + 4 or 4/2 − 1). As in the complex span task, equations were constrained so that all answers and terms were integers between zero and 10, exclusive. As in previous conditions, the image was visible for 1.5 s. Immediately following it, participants were asked for the solution to the equation. As soon as they entered the solution, they were presented with a new equation and image-label pair. Feedback about whether they answered the previous equation correctly was displayed in the upper left corner of the screen.

Concurrent verbal load. Work by Gordon, Hendrick, and Johnson (2002) suggests that interference with a linguistic task is higher when the interfering task consists of similar items. These authors found that syntactic processing was disrupted when participants had to parse sentences while simultaneously having to memorize lists of words that were similar to the words in the sentences. To illustrate this, consider a sentence like "It was Tony that Joey liked before the argument began". Participants performed more accurately on a comprehension test about the sentence if they were asked to simultaneously remember a list of common nouns (e.g., poet–cartoonist–voter) than if they were asked to simultaneously remember a list of proper male names like those in the sentence (e.g., Joel–Greg–Andy).

I mimicked this situation as closely as possible by requiring participants to memorize lists of four nonsense CVC words. These nonsense words had a similar form as the determiners, and thus potentially provided a high amount of interference. The set of possible nonsense words was: nid, zep, lum, dit, vok, pob, faz, kiv, sug, bef, rin, and tal. No single list presented to participants contained the same word twice. The Gordon et al. (2002) study used lists of three words, but I decided to use lists of four because participants seemed to be able to manage lists of six letters in the HIGH CONCURRENT LOAD condition and I wanted to make the task as difficult as possible; however, since nonsense words are more difficult than letters, I thought six might be too much. As in the HIGH and LOW concurrent load conditions, participants viewed each list for 2.5 s and the image was displayed for 1.5 s. This was followed by a response phase in which participants reported the last set of letters in order, and memorization accuracy and elapsed time were displayed. When the participant pressed Next the next set of words to memorize appeared.

Results

There are three natural questions one must answer in order to properly understand this experiment. First, were the load tasks difficult enough? Second, did participants in any of the load conditions regularize the determiners more? Third, did individual differences in performance on the initial complex span task predict performance on the word learning task? The answer to the first question is an essential pre-requisite to interpreting the answers to the other two because if the load tasks were not challenging enough, comparisons between conditions are meaningless. The answers to the other two bear directly on the questions motivating this work: does putting adults under cognitive load cause them to make the same regularization errors that children do? Were adults with poorer performance on the complex span task, who have lower working memory capacity, more likely to make those errors? I address each of these questions in turn.

Were the load tasks difficult enough?

It is non-trivial to definitively determine whether the load tasks difficult enough to significantly impair working memory capacity while still being easy enough that something could be learned in the first place. What is "enough"? Although this question is difficult to answer, we can at least evaluate converging evidence from several different directions by investigating a variety of possible indicators.

One possible indicator relates to accuracy learning the noun-image mappings. Each of the 10 images was randomly but consistently paired with one of the 10 possible nouns, and the accuracy with which the participant learned those mappings is an indication of how difficult they found the task. One would expect that performance would be substantially worse in the load conditions if the secondary task provided a sufficient challenge to the cognitive capacities of our participants.

To explore this, each person's answers were coded as *correct* if the noun they produced was identical to or phonologically similar to the correct noun for that image (e.g., wolin instead of wolid). Nouns were counted as "phonologically similar" if they had at least one syllable in common with the correct noun, *and* it was obvious which of the 10 nouns was the intended target. For instance, dragler would not be counted as correct if the correct noun was dragnip, because although there is considerable phonological overlap, there is an equivalent amount of overlap with another possibility (raygler). However, dragzoo would be counted as an instance of the word dragnip, because there are no words that end in zoo. All nouns were coded by the author, but in order to ensure accuracy and objectivity, reliability analysis was undertaken. A random subset of approximately 10% of the participants (17 people, two to three from each condition) were recoded by a second coder who was blind to the decisions made by the first coder. The reliability between the two coders was high, reflected in a Cronbach's α of 0.9703 ($N = 680$).

Fig. 2 demonstrates that participants in all of the load conditions got fewer nouns correct than in the NO LOAD condition. This suggests that the interference tasks did indeed impose a significant strain on their cognitive resources. A one-way ANOVA on nouns correct by condition was significant ($F(6, 168) = 6.298$, $p < .0001$). Planned comparisons using the Holm–Bonferroni method to adjust for multiple comparisons indicated that the mean nouns correct in the NO LOAD condition was significantly different from all

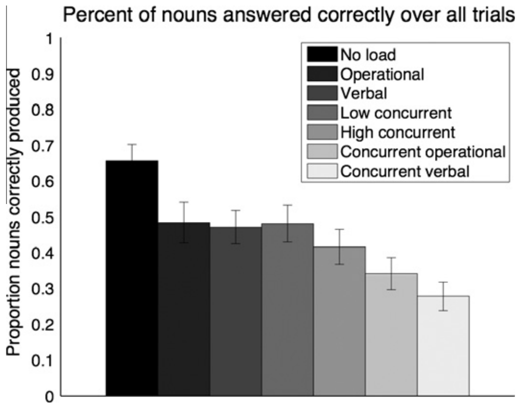

Fig. 2. Performance by condition in the noun-learning task. Participants in the load conditions learned significantly fewer nouns than in the NO LOAD condition, suggesting that the load task provided sufficient cognitive challenge to impair performance.

other conditions.[5] In order to establish that the load tasks were more difficult throughout word learning, rather than just at the beginning, the same analysis was performed for the first half (20) and last half (20) test trials, as well as for the last 10 trials. There was a significant effect of condition in all cases (first half: $F(6,168) = 6.554$, $p < .0001$; last half: $F(6,168) = 5.134$, $p < .0001$; final 10: $F(6,168) = 4.515$, $p = .0002$). Planned comparisons using the Holm–Bonferroni method revealed that the NO LOAD condition was significantly different from all other conditions in the first half of the trials, and from the CONCURRENT VERBAL, CONCURRENT OPERATIONAL, and HIGH CONCURRENT conditions in the second half and final 10 trials.[6]

Another indication that participants were attending to the load task and took it seriously is their performance on it. For three of the conditions, chance performance on the load task was 50%: participants were asked questions with two possible answers. One-sample *t*-tests reveal that in all of these conditions, performance was significantly higher than chance (OPERATIONAL LOAD: $M = 0.95$, $SD = 0.06$, $t = 36.33$, $df = 23$, $p < .0001$; VERBAL LOAD: $M = 0.80$, $SD = 0.18$, $t = 8.42$, $df = 24$, $p < .0001$; CONCURRENT OPERATIONAL LOAD: $M = 0.76$, $SD = 0.11$, $t = 11.63$, $df = 24$, $p < .0001$).[7] The other three conditions required participants to memorize lists. Although it is difficult to elucidate a standard that one would expect participants to attain in this task, people in all conditions succeeded in memorizing many words. They memorized an average of 3.36 letters per trial in the HIGH CONCURRENT LOAD condition (56% of the letters they were presented with), 2.54 letters per trial in the LOW CONCURRENT LOAD condition (84% of the letters they were presented with), and 1.78 words per trial in the CONCURRENT VERBAL LOAD condition (44% of the words they were presented with).

To what extent is performance on the load task correlated with accuracy in learning nouns? A negative correlation between performance on the load task and accuracy might suggest that different participants adopted different strategies during the experiment: perhaps some focused most of their effort on the load tasks and others focused theirs on the word learning task. However, the correlation between performance on the load task and overall accuracy was positive, suggesting instead that the participants who learned more nouns performed *more* highly on the interference task rather than less ($r = .344$, $p < .0001$). This suggests that the participants were performing on the load task to the limits of their abilities, and those participants with greater abilities were able to perform better on both the load task *and* the word learning task.

Did adults regularize more when under cognitive load?

The central question motivating this research was whether adults placed under cognitive load could be made to look more like children. To evaluate this, following Hudson Kam and Newport (2009), I excluded all participants who did not get at least nine out of the final 20 nouns correct on the test trials.[8] Then, on every valid trial (i.e., every trial for which a correct noun was produced) I defined a participants' *regularization index* as the proportion of relevant trials on which that person produced their most frequent determiner (including **none** as one of the possible determiner types). The index is therefore higher for those participants who regularize more.

A one-way ANOVA revealed a significant effect of condition ($F(6,109) = 3.84$, $p = .002$), but as Fig. 3 demonstrates, the trend, if anything, was for people in the load conditions to regularize *less* than in the NO LOAD condition. Planned comparisons using the Holm–Bonferroni adjustment method indicated that the only condition that was significantly different than NO LOAD was CONCURRENT VERBAL – and in that condition people regularized less ($p = .002$).

How much were these results driven by the exclusion of participants who did not get nine out of the final 20 nouns correct? To test this, I performed the same ANOVA on regularization by condition, but included all participants. The results were qualitatively identical.[9] The same results were attained when regularization was evaluated on trials in which participants produced *any* response at all, instead of just the trials where the noun was correct,[10] on the second

[5] CONCURRENT VERBAL: $p < .0001$; CONCURRENT OPERATIONAL: $p < .0001$; HIGH CONCURRENT: $p = .0034$; VERBAL: $p = .0207$; LOW CONCURRENT: $p = .0295$; OPERATIONAL: $p = .0234$.

[6] First half: CONCURRENT VERBAL: $p < .0001$; CONCURRENT OPERATIONAL: $p = .0002$; HIGH CONCURRENT: $p = .0105$; VERBAL: $p = .0074$; LOW CONCURRENT: $p = .0344$; OPERATIONAL: $p = .0235$. Second half: CONCURRENT VERBAL: $p < .0001$; CONCURRENT OPERATIONAL: $p = .0001$; HIGH CONCURRENT: $p = .0028$. Final ten: CONCURRENT VERBAL: $p < .0001$; CONCURRENT OPERATIONAL: $p = .0008$; HIGH CONCURRENT: $p = .0023$.

[7] Note that for one participant in the OPERATIONAL LOAD condition, an equipment failure meant that performance on the load items was not recorded; however, the rest of their data was retained so they were included.

[8] This resulted in 23 subjects in the NO LOAD condition, 19 in LOW CONCURRENT LOAD, 18 in OPERATIONAL LOAD, 17 in VERBAL LOAD, 15 in HIGH CONCURRENT LOAD, 14 in CONCURRENT OPERATIONAL LOAD, and 10 in CONCURRENT VERBAL.

[9] There was once again a significant effect of condition ($F(6,165) = 3.67$, $p = .002$), with the NO LOAD condition having the second-highest regularization index (although this time the planned comparisons showed that no conditions were significantly different from NO LOAD).

[10] There was a significant effect of condition ($F(6,109) = 3.26$, $p = .006$), and planned comparisons (Holm–Bonferroni) indicated that the only condition that was significantly different than NO LOAD was CONCURRENT VERBAL, in which people regularized less ($p = .002$).

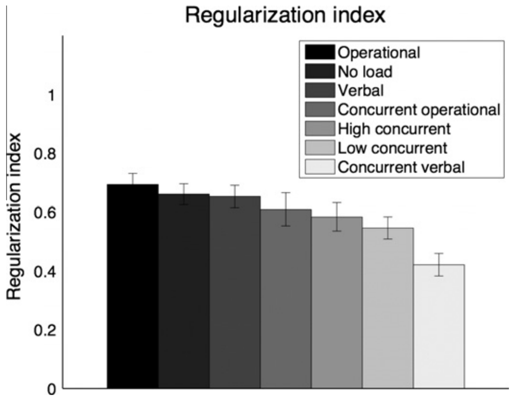

Fig. 3. Regularization by condition. The regularization index (plotted on the *y* axis) is the proportion of trials on which each person produced their most frequent determiner. A higher index therefore reveals more regularization. Participants in the load conditions were no more likely to regularize than participants in the NO LOAD condition, and the trend was toward *less* regularization in the load conditions.

half of test trials,[11] and on the final 10 test trials.[12] There was no significant effect of condition at all on the first half of the test trials ($F(6,108) = 1.07$, $p = .386$).

In order to ensure that these results were not driven by some characteristic of the regularization index, I also calculated percentage of time either the **main** determiner, a **noise** determiner, or **no** determiner was produced. There were no significant differences between conditions in terms of determiner production for any of the three determiner types: that is, participants in the load conditions did not produce any of the three determiners more often according to this measure either.[13] As before, there were no qualitative differences to this result when all of the participants were included, when the trials in which participants produced any response at all were evaluated, and when the analysis was restricted to the first half, second half, or final 10 test trials. In all cases, there was no significant effect of condition on determiner production; as before, if anything the trend was for there to be less regularization in the load conditions.

A final possibility is that, even if global regularization was not affected by memory load, lexically-specific regularization might have been. This possibility is inspired by Hudson Kam and Newport (2009), 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. That is, they might use some determiners consistently with some nouns, even if overall the distribution of determiners matches the input data. Examining lexically-specific regularization was not an original goal of this study, and it is difficult to do with reliability since each noun occurred only four times over the course of all of the test trials. With this caveat in mind, I nevertheless followed Smith and Wonnacott (2010) and calculated the average conditional entropy for each participant of their determiners given each noun; lower absolute conditional entropy reflects more regularization. A one-way ANOVA on the average entropy across conditions revealed no significant effect of condition ($F(6,109) = 1.60$, $p = .154$), suggesting that the load did not increase regularization on the lexical level any more than it did on the global level.

These findings are suggestive, but because it is an analysis of mean performances this outcome may be hiding individual regularization in different directions. To evaluate this possibility, I followed Hudson Kam and Newport (2009) and set a "consistency threshold" of 90%: each participant was coded as **consistent main** or **consistent none** if they produced the main or no determiner, respectively, on at least 90% of the valid trials. They were coded as **consistent noise** if they produced one of the noise determiners consistently on at least 90% of the valid trials, and **not consistent** if they did not produce any determiner type more than 90% of the time. As Fig. 4a shows, few participants were consistent in any way, and differences between conditions were not significant ($p = .796$, Fisher's exact test).

In order to ensure that this result was not a by-product of the exclusion criteria, I ran the same analysis for all subjects, not just those that got nine out of 20 correct, and included all trials in which the participant produced any response at all. This result, shown in Fig. 4b, is qualitatively identical ($p = .373$, Fisher's exact test). I also repeated the two analyses with a consistency thresholds of 80%; the results are shown in Fig. 4c and d. Although the analysis with exclusion criteria showed a significant difference between conditions ($p = .034$, Fisher's exact test), Fig. 4c suggests that this is because participants in some conditions were actually *less* consistent than in the NO LOAD condition. The analysis without exclusion criteria did not show a significant difference between conditions ($p = .059$, Fisher's exact test).

Does working memory span have any effect on performance?

The results presented thus far suggest that people with less available working memory capacity (i.e., those in the load conditions) did not regularize more than did those in the control condition. The experiment also provides another way to evaluate how working memory capacity affects regularization: by analyzing whether individual differences in performance on the initial complex span task predicts differential performance on the word-learning task. As one would expect, performance on the complex span task is positively and significantly correlated with accuracy for nouns ($r = .2653$, $p = .005$) and performance on the load task ($r = .225$, $p = .031$) when considering only participants and trials that fit the exclusion criteria. When evaluating the full dataset, the same is true (complex span to noun accuracy: $r = .343$, $p < .0001$; complex span to load task performance: $r = .235$, $p = .004$). Regardless of whether the exclusion criteria are applied, participants with greater

[11] The effect of condition was significant ($F(6,109) = 3.34$, $p = .005$) and planned comparisons (Holm–Bonferroni) indicated that the only condition that was significantly different than NO LOAD was CONCURRENT VERBAL, in which people regularized less ($p = .002$).

[12] The effect of condition was significant ($F(6,109) = 3.06$, $p = .008$) and planned comparisons (Holm–Bonferroni) indicated that the only condition that was significantly different than NO LOAD was CONCURRENT VERBAL, in which people regularized less ($p = .003$).

[13] Main: $F(6,109) = 1.99$, $p = .074$; none: $F(6,109) = 1.48$, $p = .191$; noise: $F(6,109) = 1.65$, $p = .140$.

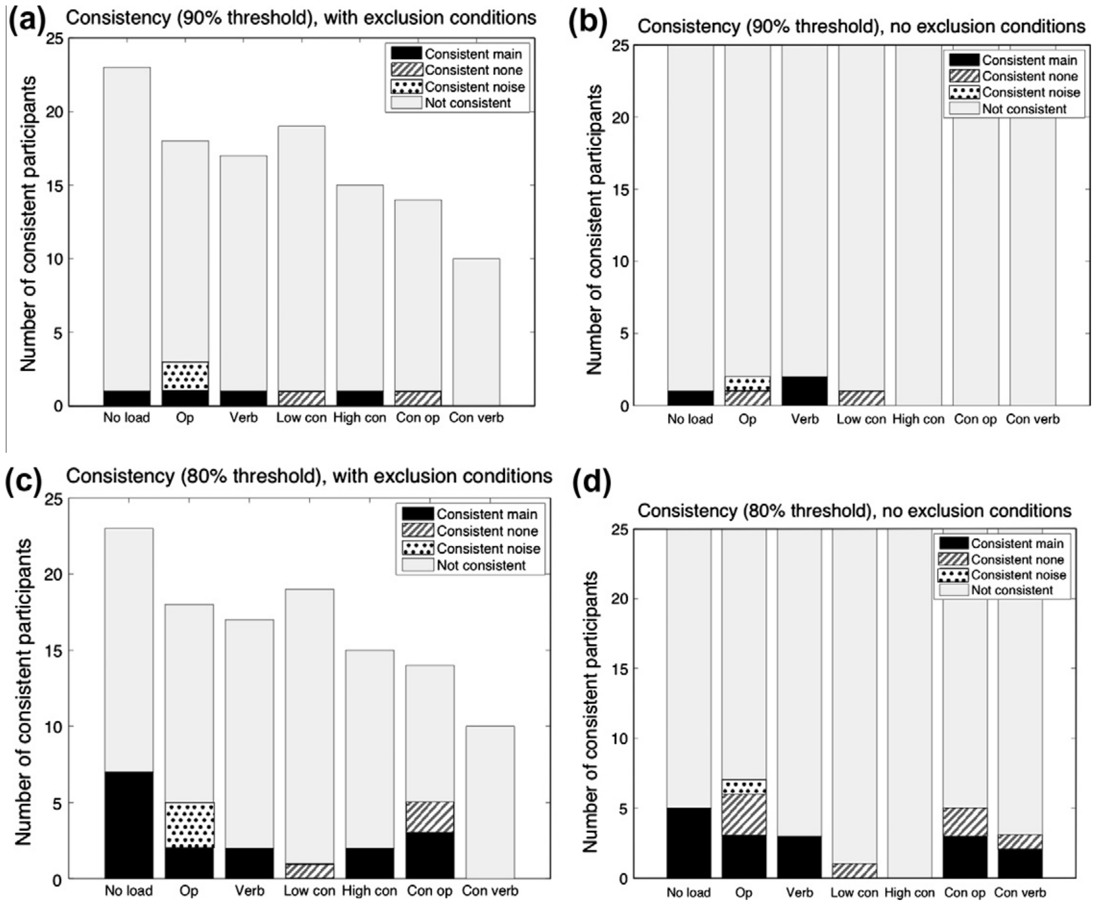

Fig. 4. Individual consistency in determiner production by condition. For the most part, few participants showed any consistency in their pattern of determiner usage, and those in the load conditions did not tend to be more consistent. The top row shows results when performance was coded as "consistent" if a response was the same for 90% of trials; the bottom shows results for a threshold of 80%. (a) Results following the exclusion conditions of Hudson Kam and Newport (2009), in which only participants who got at least nine out of 20 nouns correct were included, and only valid trials (in which the noun was correct) were examined. (b) Results including all participants and all trials in which the participant offered any response. (c) Results following the Hudson Kam and Newport (2009) exclusion conditions when the consistency threshold is 80%. (d) Results without the exclusion conditions for a consistency threshold of 80%.

working memory capacity learned more noun labels, and did better on the interference task.

Did participants with lower working memory capacity regularize more? The correlation between working memory capacity and the regularization index is non-significant regardless of whether the exclusion criteria are applied (with exclusion criteria: $r = .024$, $p = .803$; without exclusion criteria: $r = -.029$, $p = .706$). This suggests that working memory capacity has no relationship to the tendency to regularize in this experiment.

Conclusion

Overall, these results suggest that adult participants placed under cognitive load during language learning do not tend to regularize inconsistent linguistic input. First, the load tasks in this experiment did appear to tax the participants, as evident in their performance on the load tasks as well as their decreased ability to learn nouns when under load. Second, participants under load did not regularize more, regardless of the exclusion criteria used. Finally, there was no relationship between working memory and tendency to regularize in this task.

What is going on here? One possibility is that children simply have a prior bias to favor regularization, whereas adults do not. This bias might be language-specific (e.g., Bickerton, 1984) or more domain-general (which would be consistent with observed age-related differences in probability matching); either way, it would result from something other than age-related differences in memory. It is also possible that memory limitations during learning (as opposed to retrieval, as suggested by Hudson Kam & Chang, 2009), should *not* result in regularization.

I explore this possibility in the next section by using a computational model to investigate the expected effects of both prior biases and memory limitations, and how they trade off against each other. Because the Less is More hypothesis does not specify under what precise assumptions about memory one would expect limitations to result in regularization, the model is designed to evaluate a variety of possible such assumptions, ranging from more to less realistic, and aiming to qualitatively capture effects stemming from both working and long-term memory. The model demonstrates that in the absence of any prior bias for regularization, memory limitations affecting encoding and/or storage should not result in regularization unless they distort the data in a particular way. When there is a prior bias for regularization, a memory-limited learner should show regularization but a non-memory limited learner should not. When there is no such bias, even a memory-limited learner will not regularize. This implies that child–adult differences in regularization are probably not due to memory limitations on encoding or storage, at least given the current well-accepted models of memory considered here.

Computational analysis

Most tasks in which there is the potential for regularization can be described abstractly as tasks in which there are k possible outcomes and the learner must learn the distribution over those outcomes. In the experiment in this paper there are six outcomes associated with each noun (five for each of the determiners, and one for no determiner), while in a typical probability matching task, the outcomes might be the frequency of different colors of flashing lights or cards in a deck.

This situation is captured mathematically by the multinomial distribution, where θ_i denotes the probability of outcome i, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a vector representing the probabilities associated with each outcome category, and $\sum_{i=1}^{k}\theta_i = 1$. Since the experiment has multiple words, each of which could potentially be associated with a different distribution over outcomes, it is necessary to capture a separate $\boldsymbol{\theta}^{(j)}$ for each word j. For notational clarity, the (j) superscript is suppressed throughout the rest of this paper. In a multinomial distribution, the data for the observed outcomes $\mathbf{y}$ (where each y_i is a count of the number of times the ith category occurred) are generated from the underlying vector of outcome probabilities $\boldsymbol{\theta}$ according to the following equation:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \binom{N}{y_1 \cdots y_k}\prod_{i=1}^{k}\theta_i^{y_i}. \tag{1}$$

where N is the total number of observations. The task of the learner is to reason backward from the outcomes $\mathbf{y}$ to infer the nature of the underlying "true" distribution $\boldsymbol{\theta}$. For instance, imagine observing three instances of one determiner and two instances of another: we would write this as $\mathbf{y} = [3\ 2]$. A learner who probability matched would infer that $\boldsymbol{\theta} = [0.6\ \ 0.\text{s}4]$, whereas one that regularized might infer that $\boldsymbol{\theta} = [1.0\ \ 0.0]$.

Which distribution is learned will depend on two things: the nature of the data $\mathbf{y}$ and any prior beliefs about what $\boldsymbol{\theta}$ should look like. The importance of the data is obvious, but the presence of a prior is also key. A complete absence of prior belief would mean that $\boldsymbol{\theta}$ should always match the observed distribution $\mathbf{y}$ precisely; such a learner would never generalize beyond the input at all and would always precisely probability match. It is possible to have very mild prior beliefs – e.g., the weak expectation that any outcome is equally likely – which would still enable some generalization.

Modeling prior biases

The most natural and widely used prior for multinomial data is the Dirichlet distribution (Gelman, Carlin, Stern, & Rubin, 2003). This model uses a symmetric Dirichlet distribution, which imposes no prior bias in favor of any one outcome more than another across the whole dataset. Symmetric Dirichlet distributions have one parameter, α, which captures the degree to which each item (each noun, in this case) is expected to be associated with only one outcome (determiner); it governs the extent of the bias for regularization. If α is very small, the model will assume that each noun is associated with only one determiner (although, because the prior is symmetric, it will have no prior bias about *which* determiner is most likely); this constitutes a strong prior bias for regularization. When $\alpha = 1$, there is no bias for regularization; it is weakly assumed that each outcome will occur with equal probability. I evaluate the role of the prior by considering four values of α: 1 (NO BIAS), 0.1 (WEAK BIAS), 0.01 (MEDIUM BIAS), and 0.001 (STRONG BIAS).

Can a prior ever be learned? In the present case, the prior is simply higher-level knowledge about the expected distribution of determiners (outcomes) across nouns (items). It is indeed possible to learn this sort of information; formally, this corresponds to learning about α rather than pre-specifying it (Kemp, Perfors, & Tenenbaum, 2007; Perfors, Tenenbaum, & Wonnacott, 2010). Doing so would entail making higher-level inferences not just about the nouns and determiners that have been observed, but also about nouns and determiners in general. For instance, if in case A one observed many nouns, each associated with only one determiner, one might learn that α was closer to 0.01 or 0.001; conversely, in case B, observing many nouns, all associated with many determiners, would imply that α is closer to 1. This sort of "learning on multiple levels" can license more appropriate inferences; it enables the learner to correctly respond to a new noun, rather than being guided by a prior bias that might be inappropriate.

Because a prior is necessary, learning about α does not mean removing any bias entirely; it simply means setting the prior bias one level higher. Just as α governs the behavior of $\boldsymbol{\theta}$, so too does a parameter at the higher level (call it λ) govern α. Intuitively, λ places constraints on α in a similar way that α places constraints on $\boldsymbol{\theta}$; an extreme value of λ would constrain which values of α were probable. However, since λ is one level "removed" from the data, the constraints it places on the range of probable values of $\boldsymbol{\theta}$ are correspondingly weaker. Put another way, a learner who could learn α would essentially have a weaker prior about the nature of $\boldsymbol{\theta}$, and more flexibility to account for a *range* of data – being able to respond sensibly to both cases A and B above. There is some evidence that adults and children are capable of learning α in a linguistic context (Perfors et al., 2010; Wonnacott, 2011).

Because of these considerations, in addition to systematically varying values of α, I also model the situation in which α is learned. This model is a special case of a model specified in detail elsewhere (Kemp et al., 2007; Perfors et al., 2010). Formally, α is generated by an exponential distribution parameterized by λ, which is set to 1; this reflects weak prior knowledge that α does not have an extreme value.

Predictions about the expected distribution over outcomes given the data and the priors are given by Bayes Rule, shown in Eq. (2). (During the cases in which α is not learned, $P(\alpha|\lambda)$ is a constant). The integral is approximated by using an MCMC algorithm to draw 10,000 samples of the posterior distribution over $\boldsymbol{\theta}$. The results are calculated based on 100 independent runs for each condition and level of memory limitation.

$$P(\boldsymbol{\theta}|\mathbf{y},\alpha,\lambda) = \frac{P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\alpha)P(\alpha|\lambda)}{\int_{\theta'} P(\mathbf{y}|\boldsymbol{\theta}')P(\boldsymbol{\theta}'|\alpha)P(\alpha|\lambda)d\boldsymbol{\theta}'} \tag{2}$$

Modeling memory limitations

In addition to varying the strength of the prior bias for regularization, it is necessary to also model the effects of memory. The modeling approach in this paper is focused on the computational level of analysis (Marr, 1982): the central question is about how altering or deleting the data available to the learner affects the inferences that can be made. I am not concerned with modeling the time course and/or cause of forgetting (e.g., Anderson & Schooler, 1991; Brown, Neath, & Chater, 2007; Hitch, Burgess, Towse, & Culpin, 1996; Lewandowsky, Oberauer, & Brown, 2009) or capturing how memory is integrated with other aspects of human cognition, like central processing (e.g., Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974; Barrouillet, Bernardin, & Camos, 2004; Oberauer, Süß, Wilhelm, & Wittmann, 2003; Unsworth et al., 2009). This vastly simplifies the nature of the modeling choices that must be made: it is only necessary to capture the ways that different hypotheses about the nature of memory limitations predict changes in the pattern or quantity of data available to the learner. I consider four different ways, inspired by different current theories of memory, in which the pattern and/or quantity of the data might be altered by memory limitations. Note that this framework does not assume that limitations on memory encoding lead to forgetting *per se*; just that the effect of both forgetting and failure to accurately encode is to distort or lessen the data available to the learner.

Drop

One possibility is to assume, as a first approximation, that memory loss means dropping data at random. Memory limitations can therefore be modeled by changing the probability m that a given data point will be dropped (I vary m between 20%, 40%, 60%, 80%, and 90%).[14] Although extremely simplistic, this approach is roughly consistent with theories of memory loss that suggest that it primarily acts on individual tokens, and is a function of temporal processes or interference (e.g., Brown et al., 2007; Murdoch, 1960) or cognitive factors like processing or rehearsal (e.g., Carpenter, Just, & Shell, 1990; Fry & Hale, 1996; Lewandowsky, 2011; Salthouse, 1991). This is because issues like the temporal nature of forgetting or the effects of rehearsal do not impact the overall shape of what data is ultimately remembered by the system, at least assuming all of the data are presented randomly in the first place (as they are in all of the experiments considered here).

Another possibility is to assume that memory limitations result in data being forgotten and then reconstructed by the mind. The following two conditions capture different ways of implementing this possibility.

Random

A trivial way to reconstruct forgotten data would be to randomly reassign it to any of the possible outcomes with equal probability; this is the RANDOM condition. Thus, an error rate of m% would mean that a forgotten determiner is randomly reassigned to another determiner outcome with m% probability. Although this condition is not very realistic, including it serves as a baseline to compare others to.

Prior

Most of the research on memory-based reconstruction suggests that forgotten data is not reconstructed randomly (e.g., Estes, 1997; Schacter, Guerin, & St. Jacques, 2011). Rather, remembering reactivates the brain regions associated with the experience of the information (Wheeler, Petersen, & Buckner, 2000). Memories appear to be reconstructed to line up with the "gist" or associates of the items that *are* remembered (Brainerd & Reyna, 2005; Gallo, 2006; Roediger & McDermott, 1995) or to match the schemas we use to interpret the world (Bartlett, 1932; Lichtenstein & Brewer, 1979; Loftus, 2005). It is possible to roughly capture this basic idea within our framework by assuming that forgotten data is reconstructed according to its prior probability. This can be modeled using the Chinese Restaurant Process (CRP):

$$P(\text{determiner } i|\text{previous data}) = \frac{n_i}{N+\alpha}$$
$$P(\text{new determiner}|\text{previous data}) = \frac{\alpha}{N+\alpha}$$

where n_i refers to the number of observations involving determiner i made so far, N is the number of observations total, and α is the same parameter that captures the prior bias. The Chinese Restaurant Process gives the same distribution that draws from a Dirichlet process do, which is why the CRP is a natural way to capture memory loss within this model. In this condition, an error rate of m% would mean that a data point has m% chance of being "forgotten" and then reconstructed according to the CRP. Note that when α is learned, it is not straightforward to model this sort of reconstruction, since the inferred prior (α) would be constantly changing; I therefore do not attempt to do so.

So far, these conditions have presumed that memory limitations involve distortions to the data on the token

[14] It is occasionally possible for high error rates that all data may be eliminated. Although it would be possible to capture this simply by giving the model no data, it is not straightforward to determine how that model would generate new data, since the model has effectively seen no possible outcomes. We therefore capture the "zero data" case (in all conditions) by giving the model one data point for "no determiner". This is a bit *ad hoc*, but any choice made here would be similarly *ad hoc*. Because it is rare for all data to be completely eliminated, this case is does not drive the main findings here.

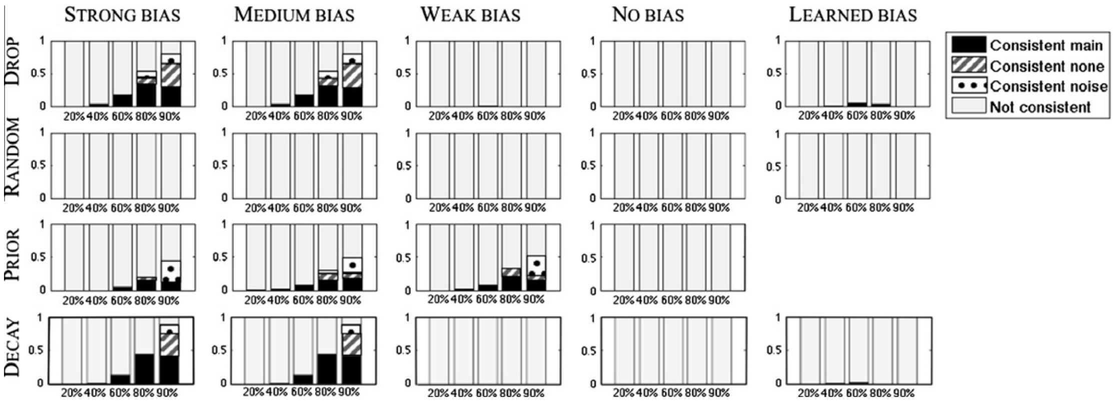

Fig. 5. Model performance varying the strength of the prior bias (columns) and the effect of different kinds of memory limitation (rows). Each graph shows the proportion of **consistent** classifications out of 100 iterations (on the *y* axis) as a function of the percentage of memory affected (on the *x* axis): *m*% means that *m*% of the data are either dropped (Drop), flipped randomly (Random), reconstructed based on the prior (Prior), or had a proportional affect on the decay rate of types (Decay). Across all conditions, regularization only occurs when memory is limited *and* there is a prior for regularization.

level; that is, individual tokens are forgotten or reconstructed. However, some models of memory make a distinction between types and tokens. The final condition attempts to capture the spirit of this distinction.

Decay

More neurologically-inspired models of memory sometimes make an important distinction between types and tokens (e.g., Bowman & Wyble, 2007; Chun, 1997; Kanwisher, 1987). For instance, Bowman and Wyble (2007) propose that memory involves two stages. The first is devoted to processing, which effectively establishes fragile type representations. For an item to be represented more durably, it must make it through a second stage, which is the entrance to working memory. It is in this second stage that the system attempts to associate each type with a discrete episode, or token. Working memory encoding is thus the process of binding a token to a type. Items at the first stage (types) are subject to rapid decay, but are reactivated by new tokens.

There are many complexities within the Bowman and Wyble (2007) model that are not relevant to the grain at which memory is being modeled here. However, the basic picture – of types that decay but can be reactivated by new tokens – is something that can be captured within this framework. To do so, the amount of decay d is modeled as $1 - m$, where m is the error rate. This quantity is multiplied by the distribution of data in the input. Fractional numbers of tokens are converted to integers by rounding proportional to their distance from the nearest integer. For instance, consider a determiner distribution for one noun of [6 1 1 1 1 0], indicating that six tokens of one determiner, one token each of the other five, and no tokens with no determiner have been observed. An error rate of 20% would be modeled by multiplying that determiner distribution by 0.8, producing [4.8 0.8 0.8 0.8 0.80]. Each resulting token would be rounded up with 80% probability (because their fractional portion is 0.8); thus, a possible version of this data with the memory limitation applied would be [5 1 1 0 1 0].

These four conditions are not intended to capture all of the details of current theories of memory. Most of the details are largely relevant on the algorithmic rather than the computational level, and contain many complexities that are beyond the scope of this paper. The conditions *are* intended to capture the range of ways that memory might affect the pattern and quantity of data that is available to the learner, since the central question is how that changes the nature of the inferences such a learner might make.

Results

Fig. 5 shows expected performance by prior bias and memory. To make the model results comparable to the experimental findings, consistency is calculated the same way as in the experiment: e.g., **consistent main** means that on that iteration the model predicted that 90% or more of the determiners should be the **main** one, while **consistent noise** means that on that iteration the model predicted that 90% or more of the determiners should be a particular **noise** one. Each of the stacked bars reflects the proportion of runs (out of 100) in which the model achieved any of each kind of consistency.

The first thing to notice about these results is that they are all quite similar: with one exception, the qualitative pattern is the same regardless of how the memory limitations are modeled. The one exception is the random condition, which shows no regularization at all, ever. Of course, that condition was the least grounded in the literature and was intended mainly as a comparison condition for the more realistic possibilities. All of the other conditions demonstrate two basic effects.

First, simply having a prior bias for regularization is insufficient to cause regularization. In all memory conditions, even when the prior bias is strong, there is no regularization when the error rate is small (i.e., there are few memory limitations during encoding). The reason for this is clear: a prior for regularization only has a noticeable effect when there is a tiny amount of data. Memory limitations have the effect of limiting the quantity of (accurate) data, but other data-limiting factors might also include bottlenecks in the input or attentional restrictions.[15] The reason that a prior bias alone is insufficient is because a sufficient quantity of data will always overcome any prior; a rational learner should think it much more likely that a given determiner actually occurs 60% of the time if it is observed in 600 out of 1000 observations rather than three out of five. Because quantity of the data matters, a prior bias only has an effect when there is little veridical data available. The exact amount of data that counts as "little" will depend a great deal on other characteristics of the learner, but this qualitative pattern – of any effect of the prior being swamped by enough data – is a general hallmark of rational reasoning.

The second implication of these results is that memory limitations alone do not result in regularization either. No models showed any hint of regularization in the NO BIAS or LEARNED ALPHA conditions, and only the prior-based reconstruction model (PRIOR) regularized in the WEAK BIAS condition. The reason a prior bias is necessary is because without it, none of the memory limitations change the overall *pattern* of data towards an regularized one. To illustrate this, consider one noun whose determiners followed the distribution in the data. If the learner randomly forgot 60% of those datapoints (as in the DROP condition), it is unlikely that they would forget all of the **noise** determiners; it is much more likely that they would forget some of the **noise** ones and some of the **main** ones. Reconstructing forgotten data according to the prior partially counterbalances this effect, thus amplifying a weak bias, but if the prior is not biased at all then reconstructing according to it does not impose a bias. Reconstructing data randomly (as in the RANDOM condition) makes the "remembered" data *more* uniform, thus destroying any tendency for regularization, regardless of the nature of the prior. And finally, modeling forgetting as gradual decay (as in the DECAY condition) does not change the pattern of data at all.

What about situations with very high memory loss, like 80% or 90%? It is true that in those situations, so much is forgotten or distorted that the pattern has been somewhat altered. For instance, in most of the memory conditions, at 90% memory loss it is more likely for the **main** determiners to be remembered than any of the **noise** determiners, simply because there are more **main** determiners in the first place. However, when memory loss is that high, there are very few data points remembered at all. When there is little data available, it is outweighed by the prior; and if the prior does not favor regularization, the learner will assume that any outcome is possible. Thus, there is a bit of a catch-22 inherently built in. When memory loss is low, it is not sufficient to change the pattern of the data; when it is high, the pattern is changed, but there is so little data that the prior plays the main role in guiding generalization. In either case, memory limitations should not lead to regularization unless there is already a prior bias favoring it.

Of course, the preceding analysis depends critically on the premise that a high memory loss leaves so little data that the prior plays the main role in guiding generalization. But suppose there were more data to learn from in the first place? In that case, might high memory loss significantly distort the pattern while still retaining enough data to not be outweighed by the prior? To test this, I reran all of the analyses in Fig. 5 with 100 times as much data. The results still show no regularization in the absence of a prior bias. There is less regularization overall – not a surprise, since there is so much data and any prior biases are outweighed by sufficient data. But the same catch-22 applies: if there is enough memory limitation to change the pattern, the amount of data is so low that the prior plays the main role in guiding generalization.

This analysis suggests that the only model of forgetting that would change the underlying pattern would be one in which all of the less-frequent outcomes (like the **noise** determiners) are *preferentially* forgotten – that is, forgotten *more* than would be predicted by any of the memory models considered here. In other words, memory *itself* would have to distort the data towards regularization. This is what happened in the PRIOR condition, which is why we observe more regularization in that condition than in any other. But even this only occurs when there is at least a weak bias for regularization: are there models of memory that predict this sort of distortion regardless of the nature of the prior bias?

One model that might have this effect could be a threshold model, in which types are remembered only if a threshold number of tokens have been seen. If that threshold is higher than the number of **noise** determiners but lower than the number of **main** determiners, such a model could effectively distort the data so that the only input that makes it past the memory filter are **main** determiners.

Such a model, however, does not appear to be very realistic, at least not for explaining this situation. First, the pattern of data it would produce is somewhat odd. It suggests an all-or-nothing type of response in which the **main** determiner is *always* regularized if the data falls into the "sweet spot" where the threshold is higher than the **noise** determiners but lower than the **main** determiners, but is *never* regularized otherwise. In particular, any kind of threshold or all-or-none-type model predicts a sudden and large qualitative shift as the amount of data falls below threshold – from always remembering and using a given type, to never doing so. This appears to contradict decades' worth of experiments showing that memory degradation is gradual. Although it might be possible to address this objection by allowing the threshold to be noisy, there is a more severe problem: any such model would only predict regularization while the quantity of data falls below the threshold. As soon there is enough data to cross the threshold, regularization should cease. Yet creole speakers do not eventually abandon creole and come to resemble pidgin speakers despite spending their entire lives exposed to pidgin, the deaf child Simon did not gradually become more inconsistent with age, and adult learners in our experiment did not show an initial stage of regularization before their data surpassed their threshold. In fact, memory limitations

[15] Another factor might be that there simply is not much data to begin with. However, although children's linguistic input is impoverished in many ways, this alone cannot drive the observed experimental differences between adults and children, since both are generally given the same amount of input (e.g., Hudson Kam & Newport, 2005); nor does it apply to the present experiment.

cannot be an explanation for differences between adult learners and native speakers (who acquired the language during childhood) unless the effects should remain even as the quantity of data increases; this is not the case for a threshold-type model.

Second, there are no models of memory encoding that I am aware of that qualitatively correspond to this sort of threshold model. The threshold models of memory that do exist are dual-process models of the *conscious recognition* of whether a particular item has been seen before (e.g., Batchelder & Riefer, 1990; Swets, 1986; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996); they are not models of the sort of memory relevant to this kind of task, which is generation based and requires a recall model. Moreover, these threshold models presume that the underlying memory trace is continuous and that the threshold governs conscious recognition of the episodic memory event only. Even these models have been criticized (e.g., Dunn, 2008), and continuous analogs have been proposed in an effort to better fit the empirical data (e.g., Wixted & Mickes, 2010).

What if memory operated in such a way that a *separate* forgetting process applied to types and tokens? Although this possibility is not consistent with any memory models I am aware of, it may fit within the learning framework of Goldwater, Griffiths, and Johnson (2011). This approach is a model of inference that envisions language as generated by a two-stage model. The first stage is responsible for generating the allowable word types, while the second generates different numbers of tokens of each type, thus transforming the word frequencies of the first stage so that they more closely match natural language. Although the framework is a model of word generation rather than forgetting, if we make the additional assumptions that types and tokens are also forgotten separately, it can be captured within the modeling framework here. Doing so requires setting an error level for types and a separate error level for tokens. I performed a this analysis and found that for a wide range of error levels, particularly those most consistent with the best performance of the model[16] in Goldwater et al. (2011), regularization again does not occur without a prior bias.

A final type of memory process that is worth thinking about is one that is more neurologically based and therefore difficult to incorporate into the modeling framework here. For instance, memory has been argued to be captured by a network where multiple constraints support gradual learning of repeated, interleaved items (as in, for instance, McClelland, McNaughton, & O'Reilly, 1995). This model hypothesizes two complementary learning systems. One, centered in the hippocampus, is designed for rapidly learning specific events, and assigns distinct representations to individual stimuli. The other, in the neocortex, slowly learns statistical regularities, and thus forms an abstraction based on the shared structure amongst the individual tokens. The closest thing this maps onto (very roughly) in the modeling framework in this paper are for the vector of observed counts $\mathbf{y}$ to be represented in the hippocampus and the inferred distribution $\boldsymbol{\theta}$ to be represented in the neocortex. Notably, these network-type models assume that learning and forgetting in both cases is gradual – which would be most closely captured within this framework in a way analogous to the DROP condition. These network-type models would only distort the data in the peculiar way necessary to cause regularization if, somehow, in the transfer from the hippocampus to the neocortex (i.e., the process of abstracting from $\mathbf{y}$ to $\boldsymbol{\theta}$) the less frequent outcomes were to be forgotten more than their frequency would predict. This does not seem to follow from the structure of the model (McClelland et al., 1995), although we should always be cautious about trying to map such different frameworks onto each other. I return to the larger issue of investigating memory limitations on a more process or neurological level in the discussion.

It is of course always possible that I have missed a model, or that one can be created that distorts data in the pattern necessary to cause regularization. To explain how memory limitations alone lead to regularization it would have to be independently motivated by other memory phenomena, as well as address the difficulties raised earlier.

Conclusion

These modeling results suggest that under a wide variety of assumptions about the nature of memory, memory limitations during encoding and storage alone do not lead to regularization: a prior bias to favor regularization is also necessary. This is because in the absence of such a prior bias, such memory limitations do not change the underlying *pattern* of data. Memory limitations would lead to regularization only if human memory distorted the pattern data in a particular way – remembering frequent items *more* than their frequency would warrant and less frequent items *less* than their frequency would warrant. The next section discusses the implications and limitations of these results, along with the experimental findings.

Discussion

The central question addressed in this paper concerned the relationship between memory limitations and regularization. What assumptions about memory are necessary for memory limitations to lead to regularization, and why? Previous work suggests that facilitating memory retrieval can increase the tendency to probability match (Hudson Kam & Chang, 2009), but it was unclear whether limitations in encoding or storage should affect regularization, or why. This study was designed to explore the effect of this kind of memory limitation using both experimental and computational methods. The experimental results indicate that adults who are placed under memory load while learning an artificial language do not regularize more than adults who are not. The computational results offer one explanation for these findings, suggesting that under realistic models of memory encoding and storage, regularization should only occur in the presence of both memory limitations *and* a prior bias for regularization.

[16] These are errors in which the probability of forgetting types and the probability of forgetting tokens are not widely divergent from each other.

In the next pages I critically discuss these results. I first concentrate on the experiment, followed by the model, and finally conclude by exploring what these findings mean about the role of memory in regularization and the implications for language learning more broadly.

Experiment

Participants were placed under memory load while simultaneously learning a simple artificial "language" composed of nouns paired inconsistently with determiners. In order to explore the effects of different kinds of load – and to ensure that the load tasks actually taxed working memory – there were seven different conditions, which differed according to the type of load. Although all of the load conditions were difficult enough to significantly impair overall learning, participants did not regularize more than participants in a control (non-load) condition. Moreover, complex memory span did not predict regularization. These results suggest that memory limitations during the encoding and storage stage do not lead to regularization.

Although it is in theory possible that the load tasks did not sufficiently challenge our participants, this is unlikely. Divided attention during encoding is known to cause deficits in memory performance (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996). Moreover, it is clear that participants took the load tasks seriously, performing far above chance in them. In addition, people who did well on the load tasks did *better*, rather than worse, at the language learning task, suggesting that participants were not disregarding the load task in order to focus on word learning. This result also implies that participants did not disregard the word learning task in order to focus on the load task; in fact, they appeared highly motivated to do well on the word learning task, since they had to label answers in front of the experimenter and could not hide behind the safe anonymity of a computer screen. Anecdotally, the participants found the task extremely challenging (several complained afterword that it was the hardest experiment they had ever done), and indeed the presence of load had a large and significant effect on noun learning.

What about the converse possibility? Perhaps the task was so difficult that with more training, regularization might emerge. This is also unlikely, since there was no tendency toward increased regularization over the course of the experiment. It is also worth noting that 200 trials was sufficient to observe regularization in much of the probability matching literature (Derks & Paclisanu, 1967; Pecan & Schvaneveldt, 1970; Weir, 1964) and that this experiment had a similar number of observations per noun as in Hudson Kam and Newport (2005, 2009), where children did regularize. More generally, if difficulty of the load tasks is an issue in either direction, it implies that the dependence of regularization on memory limitations must be extremely precisely calibrated: memory limitations cannot be so high as to render learning impossible, nor so low as to not lead to regularization. This is a balancing act that, if nothing else, seems unlikely to precisely describe the state of all child language learners.

Another concern lies in the applicability of this experiment to the previous studies, and to child language acquisition in general. There are a number of differences between the experiments in this paper and the Hudson Kam experiments, and even more differences between our experiments and the process of language learning over developmental time. The Hudson Kam and Newport (2005) studies involved learning and producing noun-determiner pairs in the context of a limited grammar over multiple days rather than one session, and the same is true a thousandfold in the case of child language acquisition in the real world. How can we be certain that the lack of regularization under load that was observed in our studies would occur if the language were richer or the learning process more prolonged?

The answer is that we cannot be certain of this, and this is an area where further work would be very useful. However, there are several reasons this concern does not invalidate the present work. For instance, consider the issue that the language learning task in this experiment was far simpler than learning a real language. Might this cause the participants to therefore treat it more like paired-associate learning rather than like learning a language with rich internal structure? It is hard to say, but even if they did, according to Less is More this should not affect their tendency to regularize: there is nothing in the hypothesis that predicts that people should only regularize in linguistic contexts. Indeed, part of the empirical support for the hypothesis comes from the fact that children but not adults regularize in non-linguistic contexts like prediction tasks (Derks & Paclisanu, 1967; Myers, 1976; Weir, 1964).

It is also worth asking why we should expect the complexity of the system in which the nouns and determiners are embedded to matter. One possibility might be that if learning is embedded in a complex linguistic system, the learner might have fewer resources to apply to learning the pairing. In other words, the complexity of the system might play the exact same role that the load tasks do. Yet here there was no additional regularization with load, and the more difficult load tasks (as measured by impact on noun learning) did not have more regularization than the simpler ones. Another possibility might be that people bring different prior assumptions to word learning tasks than grammar learning tasks. Yet even though the Hudson Kam tasks were embedded within a grammar learning framework, they still consisted of the fundamentally same task: mapping noun-determiner pairs onto referents. Why would people treat one as word learning and one as grammar learning? Plus, even if they did, there is no reason to think that this would make them *less* likely to regularize in the word-learning case. If anything, biases like mutual exclusivity would imply that regularization was more likely in a word learning context; even if such biases did not apply, there is evidence that statistical learning in the presence of variable input looks similar in both situations (e.g., Vouloumanos, 2007).

Another potential factor is that these experiments took place in one session rather than spread over multiple days (as in the case of the previous Hudson Kam work) or multiple years (as with child language acquisition). Is it possible that regularization requires a more extended learning phase, or consolidation due to intervening sleep? There is indeed research suggesting that sleep may be useful for both consolidating specific episodic memories (Gais & Born, 2004) and generalizing to new stimuli (Fenn, Nusbaum, & Margoliash, 2003; Gomez, Bootzin, & Nadel, 2006). Thus, the hypothesis that sleep might be critical for regularization, perhaps because of how it interacts with memory formation, is an intriguing one that cannot be ruled out. If it were correct, it would suggest that whatever is happening during sleep has the effect of distorting the data available to the learner in the way predicted by the modeling results here; left unexplained, however, is why that type of distortion should occur given what we currently know about sleep and the brain (e.g., McClelland et al., 1995; Walker, 2009).

A final question is what the load tasks actually disrupted. In general, they were designed to disrupt many aspects of cognition, all of which can affect what is processed, encoded, and stored. The tasks require people to retrieve information (word meanings in the VERBAL LOAD condition, number and symbol meanings in the OPERATIONAL and CONCURRENT OPERATIONAL conditions, memorized letters and nonsense words in the other conditions), to store information in short-term memory (the numbers in the equation conditions, the words and letters in the others), to manipulate representations (to determine the correct answer to the load questions), and to regulate attention between the load task and the word learning task. These considerations make it quite likely that the load tasks disrupted the process of word learning – in particular, the process of attending, encoding, and storing the information about noun-determiner pairings. One thing that they did *not* directly disrupt was retrieval, precisely because the intent of this work was to explore whether and to what extent limitations on non-retrieval aspects of memory lead to regularization. (Retrieval was still an element of the task, but it did not differ between conditions.) Indeed, the retrieval aspect of this task – the production test – was very similar to that of the studies by Hudson Kam and colleagues. Although their participants learned verbs as well as nouns and determiners, during the production task they were given the verb, and thus only had to produce the noun-determiner pairs, as in the present study (although some of the time they had to produce transitive sentences with two noun-determiner pairs in a row).

Why would limitations during learning, rather than retrieval, not affect regularization? To explore that, we turn to the computational results.

Computational

The computational model systematically explored how different degrees and types of realistic memory limitation affect the pattern of data available to the learner, and how memory limitations interact with prior biases for or against regularization. The model was deliberately designed to be extremely simple in order to minimize the extent to which these results depend on arbitrary modeling choices. The only free parameter in the model, α, governed the extent of the prior bias for regularization and was systematically varied. The underlying distribution being learned, the multinomial, is the most obvious and statistically widely-used way of capturing distributional data when many outcomes are likely, and the Dirichlet distribution is likewise the most widely-used and mathematically elegant prior for multinomial data.

Memory limitations were modeled on Marr's computational level, since the central question was about how a learner's generalizations are affected by the input available to the cognitive system. The underlying premise was that memory limitations have the net effect of distorting the input available to the learner in different ways. Modeling different distortions of that input – mostly different patterns of deletion and alteration – allows us to explore the effect of these distortions on the nature of generalizations made by the learner. We considered several different ways of distorting the input, inspired by leading models of memory. These ranged from simply dropping data at random, to reconstructing it based on one's prior assumptions, to dropping types according to a decay process. Under all of these assumptions about memory, regularization only occurs when *both* memory limitations *and* a prior bias for regularization are present. In general, regularization can only occur without a prior bias if the memory process itself distorts the pattern of data available to the learner so as to remember the most frequent items more than their frequency would warrant and the less frequent items less. However, this type of distortion does not appear to emerge naturally from any of the theories of memory captured here.

One assumption inherent in the model is that it is Bayesian, meaning that it predicts the behavior of a rational learner. This means that the importance of previous biases (the prior) and fitting the data (likelihood) are balanced in a particular way (according to Bayes' Rule). However, every model needs to perform *some* tradeoff between these factors. Because of this, models that weigh these tradeoffs differently might vary quantitatively, but all models except for the most pathological[17] should show that regularization is more likely when the input is limited and the prior bias for it is strong.

It is also worth noting that, although the model is Bayesian, this is not a typical ideal learning analysis; because the model incorporates different kinds of memory limitations, it should be more properly understood as a "capacity limited" rational model. It thus allows us to investigate what a rational learner *with certain capacity constraints* might be expected to do. In particular, it provides a means to systematically evaluate the effect of different kinds of capacity constraints, as I have done here. This sort of approach is an important step toward bridging computational-level and process-level accounts of cognition.

Relatedly, because this is a computational-level model, there are many aspects of memory that I have not attempted to capture here. The model does not include many of the issues that memory researchers most care about, from the time course of memory and forgetting, to the low-level details of how memory is implemented in the human brain, to the step-by-step integration of memory

[17] "Pathological models" include those that do not learn at all from data or never generalize at all beyond the data. Humans, of course, do neither of these things.

and other aspects of cognition. The goal here is to abstract away from these issues and to explore how different patterns of data distortion during learning affect the types of inferences a rational learner makes. These findings can thus inform future work investigating the effects of memory on a more process level.

What do our results suggest about memory? In one way, the findings are more general than may first appear, since they apply to more cognitive capacities than memory; in another, they are more limited, since they do not apply to all kinds of memory. On the first point, these results are about how input is distorted during learning; thus, *any* cognitive capacity that might distort the input during learning is theoretically relevant. The main finding of this research is that the only pattern of data distortion that leads to regularization in the absence of a prior bias is one that retains high-frequency items and drops or alters low-frequency items *beyond* what their frequency would warrant. Because the particular distortions considered in this paper were motivated by the memory literature, I am hesitant to conclude that attentional or executive processes do not have qualitative effects that distort the data in this way (although they might). In terms of memory, none of the theories of memory captured by my model resulted in that pattern of distortion. Threshold or "all or none" models of memory are one type of model that would be able to capture it, although such models have other problems. They do not correspond to any current independently-motivated memory models that I am aware of, and since they predict that regularization should eventually cease, they cannot explain the difference between native speakers and second-language learners (since native speakers eventually have more data than second-language learners, and thus should eventually regularize less rather than more). This is not meant as a conclusive argument that no realistic memory model could distort data in the precise way that would lead to regularization; however, it does seem unlikely at this point.

It is important to note that the model presented here is a model of distortions of the input that occurred *before* generalization. In other words, it is a model of how cognitive or memory limitations affect what is learned – not a model of how such limitations affect what is produced. It is therefore more about processing, encoding, and storage rather than retrieval. This was deliberate, because the main intent of this paper was to explore whether and to what extent *non-retrieval* memory limitations affect the inferences and generalizations a learner should make.

An additional important consideration is that the models captured here mainly focus on memory in general rather than memory *development* in particular. This is mainly because it was unclear how to map what we know about developmental changes in memory onto differences that would make a difference at the level of the model. For one thing, aside from children having smaller working memory spans for the most part, there appear to be few behaviorally observed differences between children and adults on simple working memory tasks (e.g., Davidson, Amso, Anderson, & Diamond, 2006; Gathercole, 1999). The differences that have been observed or postulated are at the neurological level: the pre-frontal cortex of children is more immature, and children recruit more of their hippocampus during working memory tasks than do older adolescents and adults (Finn, Sheridan, Kam, Hinshaw, & D'Esposito, 2010). This is intriguing research, but at this stage it does not make clear predictions about what particular patterns or distortions, if any, should be imposed on the data available to the learner as a result of differential hippocampal involvement and/or PFC immaturity. In fact, there is reason to believe any of the possible alternatives: that they should cause children to regularize more, or regularize less, or have no effect. On one hand, as discussed later, an immature PFC is consistent with the suggestion that adults should probability match more (Thompson-Schill, Ramscar, & Chrysikou, 2009), although this would affect cognitive control more than memory. On the other hand, hippocampal involvement is consistent with the stimuli being more novel or complex for children, in which case one would expect more binding of individual patterns and therefore less abstract learning and generalization on their part.

Conversely, if the relative novelty of the stimuli leads to differences in the actual phonological representation, then this is not a difference which would show up in this modeling framework (in which the word representations are discrete). To the extent that the lack of a detailed phonological representation has an effect on the nature or pattern of forgetting – if it does – this might matter. If words are encoded according to their phonological features in some way, being presented with multiple different novel words could cause interference effects on the level of the phonological encoding, such that none are remembered well. This is probably not what is going on in the experiments in this paper – or if it is, it is not leading to regularization – since if it were, one might expect the CONCURRENT VERBAL condition to show interference effects and thus regularization. But effects due to distributed representations more generally cannot at this point be ruled out, and indeed may be occurring in other experiments, such as Hudson Kam and Newport (2009). Further research investigating the effects of memory limitation using distributed representations is necessary.

Bringing it all together

How do these results compare to previous studies? In Hudson Kam and Newport (2009), adults were found to regularize when presented with extremely small probabilities (for instance, 16 determiners each occurring 2.5% of the time, rather than four determiners each occurring 10% of the time, as in our experiment). It is tempting to conclude that perhaps a threshold model of memory *is* appropriate in this case, and 2.5% of the time is "below threshold" whereas 10% was not. This is probably unlikely, not just because a simplistic threshold model of memory does not appear to otherwise have independent support, but because it would predict that adults in our experiment should have regularized at the beginning of the experiment (while the data were still below threshold). However, there were no differences in regularization over the course of the experiment. Although I can only speculate, it is probably more likely that other cognitive factors, like attention or metacognitive reasoning, explain the Hudson Kam and Newport (2009) results. Perhaps once there are enough alternatives, or each occurs rarely enough, the alternatives are simply disregarded. One possible reason for this might be that adults recognize that real language contains errors; perhaps determiners that occur rarely enough are ignored because they are thought to be errors (Perfors, 2012). It is also possible that such determiners could not be encoded in enough phonological detail (either because of their low frequency or due to interference with the many other determiners) for them to be generated, as hypothesized above. Future work is necessary to explore these possibilities.

In Hudson Kam and Chang (2009), adults were made to probability match more by being given assistance with retrieval: instead of being expected to generate words on their own, they were presented with a list of all possible words and simply asked to pick which words from the list they would say. This converts the task to more of a recognition than a pure production task. In this situation, people probability matched more precisely, producing **main** determiners around 60% of the time rather than around 70% of the time, as they do without help during retrieval. What explains these results?

The most likely explanation is that Hudson Kam and Chang (2009) found the pattern they did because they were manipulating memory retrieval rather than encoding or storage. The present modeling results are not relevant to retrieval – indeed, one can imagine several different retrieval processes that might make a learner regularize in production but still have an underlying representation that more closely matches the probabilities in the input. However, it is also worth noting that Hudson Kam and Chang (2009) aimed to make adults *less* like children by making the cognitive load easier, rather than to make adults act *more* like children by making it harder. There may be an inherent asymmetry to adults' performance: perhaps it is relatively easy to make adults regularize less, but making them regularize more is difficult. This is the case in the decision-making literature, in which great efforts have been made to stop adults from probability matching, often to no avail.

That said, if retrieval entirely drives regularization, it raises some problematic questions. It is unclear how retrieval effects – or, indeed, any sort of performance-driven effect – can constitute the full explanation for children's tendency to learn languages that are more consistent than their input, as in the case of deaf children like Simon (Singleton & Newport, 2004) or creolization (e.g., Mühlhäusler, 1986). If the underlying inconsistency is reproduced in the representation (as the judgment task in Hudson Kam & Newport (2009) might suggest), one would expect that once children's retrieval difficulties lessen – whether due to age or additional practice – then regularization would cease. Although this happens with some regularization (e.g., verbs like "goed" or "maked"), probably because the variation is not truly inconsistent, creole speakers do not turn into pidgin speakers as they get older, and Simon did not gradually become more inconsistent with age. The fact that these learners do not revert to a more inconsistent language as their retrieval limitations decrease implies that retrieval is not the only constraint, and that their actual representations were regularized versions of the input. The analysis in this paper suggests that most known models of memory limitations do not change one's representations in the way necessary for regularization; taken together, this suggests that something else – perhaps a prior bias for regularization – is necessary to explain the difference between children and adults.

But what is meant by "a prior bias for regularization"? The model does not make any claims or statements about where one might originate. As such, there are many possibilities, most of which have yet to be explored. One possibility is that, consistent with a different version of Less is More; such a bias might reflect *other* cognitive differences between children and adults. For instance, in addition to memory and processing speed, children and adults differ in their levels of cognitive control, which has been identified as a potentially important factor during language acquisition (Thompson-Schill et al., 2009). If children struggle to engage in top-down processes to figure out general rules, they may instead rely more strongly on more bottom-up or data-driven thinking. This has been argued to cause them to take up the most frequent and reliable patterns, whereas adults are capable of using their cognitive control to override this tendency (Ramscar & Yarlett, 2007; Thompson-Schill et al., 2009). The effect of limitations on cognitive control is not tested directly in the current paper, although the interference-based load tasks may have taxed cognitive control to at least some extent.

Executive control is related to other cognitive differences between children and adults, such as the ability to use metacognitive strategies (e.g., Flavell, Green, Flavell, Harris, & Astington, 1995). It may be that adults' ability to introspect and reason about their own cognition makes them more likely to rely on explicit rather than implicit learning (Ullman, 2004) – a difference that has been hypothesized to be the root of child–adult differences in language acquisition. Such metacognitive ability might also make adults more likely to try to capture patterns in the input that do not exist; this tendency has been suggested as an explanation for why adults probability match in non-language tasks (Estes, 1976). It might result from a generalized preference for simplicity or tendency to ignore exceptions on the part of children. Might such attentional or strategic differences be themselves related to memory? That is an open question, although at least in the categorization literature, individual differences in working memory appear unrelated to differences in strategy use (Craig & Lewandowsky, 2012) or attention (Sewell & Lewandowsky, 2012). Another possibility is that children's limited capacity means they are not capable of rapidly *learning* a prior bias, while adults can; since regularization does not occur in the LEARNED BIAS condition, this might explain the difference between adult and child performance. These options are all speculative; much more work in this area needs to be done.

Overall, this paper does not argue against the Less is More hypothesis in general. These results are irrelevant to the version of the hypothesis that is not focused on regularization; in fact, there is a great deal of converging evidence supporting the idea that "starting small" can help a learner to isolate and analyze the separate components of a linguistic stimulus. The results also do not argue against the idea that memory limitations in the form of retrieval affect regularization. What this work *does* suggest, based on converging evidence from computational modeling and seven experiments, is that cognitive limitations during learning do not result in regularization in the absence of a prior bias unless they distort the input in a particular way. Although much work remains to be done, these findings place key empirical and theoretical constraints on how and why cognitive limitations are predicted to affect regularization, and thus have important implications for understanding the difference between child and adult language acquisition.

Acknowledgments

I thank Natalie May, my lab manager, for her tireless enthusiasm and dedication to running many versions of the same experiment, as well as the members of CLCL and our many participants. I would also like to thank Daniel Navarro, Michael Frank, John Dunn, and Stephan Lewandowsky for useful discussions. This research was funded by a University of Adelaide Establishment Grant and Discovery Early Career Researcher Award DE120102378.

References

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science, 2*.

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. Spence & T. Spence (Eds.). *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York: Academic Press.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower (Ed.). *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*, 83–100.

Bartlett, F. (1932). *Remembering*. Cambridge: Cambridge University Press.

Batchelder, W., & Riefer, D. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences, 7*, 173–221.

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*, 9–49.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis, 20*, 3–49.

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review, 114*, 38–70.

Brainerd, C., & Reyna, V. (2005). *The science of false memory*. Oxford University Press.

Brown, G., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539–576.

Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review, 97*, 404–431.

Case, R., Kurland, D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386–404.

Chambers, J., Trudgill, P., & Schilling-Estes, N. (2003). *The handbook of language variation and change*. Malden, MA: Blackwell.

Chin, S., & Kersten, A. (2010). The application of the less is more hypothesis in foreign language learning. In *Proceedings of the 32nd annual conference of the cognitive science society*.

Chun, M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 738–755.

Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences, 10*, 564–570.

Cochran, B., McDonald, J., & Parault, S. (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.

Conway, A., Jarrold, C., Kane, M., Miyake, A., & Towse, J. (2007). *Variation in working memory*. NY: Oxford Univ. Press.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological overview and user's guide. *Psychological Bulletin and Review, 12*, 769–786.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *Quarterly Journal of Experimental Psychology, 65*, 439–464.

Craik, F., Govoni, R., Naveh-Benjamin, M., & Anderson, N. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125*, 159–180.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19*, 450–466.

Davidson, M., Amso, D., Anderson, L., & 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.

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

Dunn, J. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review, 115*, 426–446.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Estes, W. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37–64.

Estes, W. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review, 104*, 148–169.

Fedorenko, E., Gibson, E., & Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language, 56*, 246–269.

Fenn, K., Nusbaum, H., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 435*, 614–616.

Fernald, A., & Simon, T. (1984). Expanded information contours in mothers' speech to newborns. *Developmental Psychology, 20*, 104–113.

Finn, A., Sheridan, M., Kam, C. H., Hinshaw, S., & D'Esposito, M. (2010). Longitudinal evidence for functional specialization of the neural circuit supporting working memory in the human brain. *The Journal of Neuroscience, 30*, 11062–11067.

Flavell, J., Green, F., Flavell, E., Harris, P., & Astington, J. W. (1995). Children's knowledge about thinking. *Monographs of the SRCD, 60*.

Fry, A., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7*, 237–241.

Gais, S., & Born, J. (2004). Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences, 101*, 2140–2144.

Gallo, D. (2006). *Associative illusions of memory*. Taylor and Francis.

Gathercole, S. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences, 3*, 410–419.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis* (2nd ed.). Chapman & Hall.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research, 12*, 2335–2382.

Gomez, R., Bootzin, R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science, 17*, 670–674.

Gordon, P., Hendrick, R., & Johnson, M. (2002). Memory-load interference in syntactic processing. *Psychological Science, 13*, 425–430.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences, 9*, 219–224.

Hitch, G., Burgess, N., Towse, J., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49*, (116–139).

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, and Cognition, 35*, 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 and Development, 1*, 151–195.

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

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tokura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties with non-native phonemes. *Cognition, 87*, B47–B57.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60–99.

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.

Kane, M., Hambrick, D., Tuholski, S., Wilhelm, O., Payne, T., & Engle, R. (2004). The generality of working-memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189–217.

Kanwisher, N. (1987). Repetition blindness: Type recognition without token individuation. *Cognition, 27*.

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307–321.

Kersten, A., & Earles, J. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language, 44*, 250–273.

Klein, K., & Fiss, W. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavioral Research Methods, Instruments, and Computers, 31*, 429–432.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5*, 831–843.

Lai, J., Poletiek, F., 2010. The impact of starting small on the learnability of recursion. In: Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 720–738.

Lewandowsky, S., Oberauer, K., & Brown, G. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences, 13*, 120–126.

Lichtenstein, E., & Brewer, W. (1979). Memory for goal-directed events. *Cognitive Psychology, 12*, 412–445.

Loftus, E. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning and Memory, 12*, 361–366.

MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 49–67). Berlin: Oxford Univ. Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Company.

Mayberry, R. (1993). First-language acquisition after childhood differs from second-language acquisition: The case of american sign language. *Journal of Speech and Hearing Research, 36*, 1258–1270.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419–457.

Mühlhäusler, P. (1986). *Pidgin and creole linguistics*. Oxford: Basil Blackwell.

Murdoch, B. (1960). The distinctiveness of stimuli. *Psychological Review, 67*, 16–31.

Myers, J. (1976). Probability learning and sequence learning. In W. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Newport, E. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American sign language. *Language Sciences, 10*, 147–172.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11–28.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. (2003). The multiple faces of working memory – storage, processing, supervision, and coordination. *Intelligence, 31*, 167–193.

Pecan, E., & Schvaneveldt, R. (1970). Probability learning as a function of age, sex, and type of constraint. *Developmental Psychology, 2*, 384–388.

Perfors, A. (2012). Probability matching vs over-regularization in language: Participant behavior depends on their interpretation of the task. In *34th Annual conference of the cognitive science society*.

Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37*, 607–642.

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.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Rohde, D., & Plaut, D. (1999). Language acquisition in the absence of negative evidence: How important is starting small? *Cognition, 72*, 67–109.

Rosen, V., Bergeson, J., Putnam, K., Harwel, A., & Sunderland, T. (2002). Working memory and apolipoprotein E: What's the connection? *Neuropsychologia, 40*, 425–443.

Salthouse, T. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science, 2*, 179–183.

Salthouse, T., & Babcock, R. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27*, 763–777.

Schacter, D., Guerin, S., & St. Jacques, P. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*, 467–474.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77*, 427–442.

Senghas, A., & Coppola, M. (2001). The creation of Nicaraguan Sign Language by children: Language genesis as language acquisition. *Psychological Science, 12*, 323–328.

Sewell, D., & Lewandowsky, S. (2012). Attention and working memory capacity: Insights from blocking, highlighting, and knowledge restructuring. *Journal of Experimental Psychology: General, 141*, 444–469.

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, 116*, 444–449.

Snow, C. (1999). Social perspectives on the emergence of language. In B. MacWhinney (Ed.), *The emergence of language* (pp. 257–276). Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, M., Gabrieli, J., Stebbins, G., & Sullivan, E. (1998). Working and strategic memory deficits in schizophrenia. *Neuropsychology, 12*, 278–288.

Swets, J. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*, 100–117.

Tan, L. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping, 18*, 158–166.

Thompson-Schill, S., Ramscar, M., & Chrysikou, E. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Sciences, 18*, 259–263.

Turner, M., & Engle, R. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

Ullman, M. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*, 231–270.

Unsworth, N., & Engle, R. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132.

Unsworth, N., Redick, T., Heitz, R., Broadway, J., & Engle, R. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*, 635–654.

Vouloumanos, A. (2007). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729–742.

Walker, M. (2009). The role of sleep in cognition and emotion. *The year in cognitive neuroscience 2009: Annals of the NY Academy of Sciences, 1156*, 168–197.

Weber, A., & Cutler, A. (2003). Lexical competition in non-native spoken word recognition. *Journal of Memory and Language, 50*, 1–25.

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

Werker, J., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*, 672–683.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Wheeler, M., Petersen, S., & Buckner, R. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, 97*, 11125–11129.

Wixted, J., & Mickes, L. (2010). A continuous dual-process model of Remember/Know judgments. *Psychological Review, 117*, 1025–1054.

Wolfram, W. (1985). Variability in tense marking: A case for the obvious. *Language Learning, 35*, 229–253.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language, 65*, 1–14.

Yonelinas, A., Dobbins, I., Szymanski, M., Dhaliwal, H., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition, 5*, 418–441.