Overcoming Memory Limitations in Rule Learning

Overcoming memory limitations in rule learning

Michael C. Frank
Department of Psychology, Stanford University

Edward Gibson
Department of Brain and Cognitive Sciences, MIT

Adults, infants, and other species are able to learn and generalize abstract patterns from sequentially-presented stimuli. Rule learning of this type may be involved in children's acquisition of linguistic structure, but the nature of the mechanisms underlying these abilities is unknown. While inferences regarding the capabilities of these mechanisms are commonly made based on the pattern of successes and failures in simple artificial-language rule-learning tasks, failures may be driven by memory limitations rather than intrinsic limitations on the kinds of computations that learners can perform. Here we show that alleviating memory constraints on adult learners through concurrent visual presentation of stimuli allowed them to succeed in learning regularities in three difficult artificial rule-learning experiments where participants had previously failed to learn via sequential auditory presentation. These results suggest that memory constraints, rather than intrinsic limitations on learning, may be a parsimonious explanation for many previously reported failures. We argue that future work should attempt to characterize the role of memory constraints in natural and artificial language learning.

Infants and adults are able to learn a surprising amount from even a short exposure to novel language stimuli (Gómez & Gerken, 2000). They are able to segment words from fluent speech (Saffran, Aslin, & Newport, 1996b, 1996a); to learn word-referent pairings (L. Smith & Yu, 2008; Yu & Smith, 2007; Vouloumanos, 2008); and to discover distributional categories (Mintz, 2002), non-adjacent dependencies (Gómez, 2002), and even more abstract regularities (Gómez & Gerken, 1999; Marcus, Vijayan, Bandi Rao, & Vishton, 1999). These results support the existence of powerful learning mechanisms which are likely conserved across development. These mechanisms are also likely involved in—and perhaps even at the core of—children's ability to learn their native language (Bates & Elman, 1996; Gómez & Gerken, 2000; Kuhl, 2004). Nevertheless, their number and nature is still largely unknown, and an important goal for research in this area is the characterization of these mechanisms and their relationship to other cognitive systems.

Here we examine a group of phenomena which we refer to collectively as rule learning phenomena (after Marcus et al., 1999). The signature of these phenomena is that they involve learning a regularity that can easily be described symbolically (though its mental representation might have some other form).[1] For instance, in the experiments of Marcus et al. (1999), infants exposed to a set of novel syllable strings of the form *ABB* (where *A* and *B* represent syllables like wo or fe) were able to discriminate strings of the form *AAB* from the *ABB* strings they were trained on, even when the syllables instantiating these rules differed from training to test. Rule learning experiments have provoked interest in the language acquisition community because of their resemblance to the tasks involved in learning structural aspects of language such as syntax and morphology. An understanding of the mechanisms underlying rule learning may therefore help in identifying whether these mechanisms are in fact useful for or involved in natural language acquisition.

An important method in characterizing the mechanisms of rule learning has been comparison of the difficulty that participants (usually adult learners but sometimes infants) have in acquiring rules of different kinds (Endress, Dehaene-Lambertz, & Mehler, 2007; Endress & Bonatti, 2007; Endress, Scholl, & Mehler, 2005; Frank, Slemmer, Marcus, & Johnson, 2009; Gómez, 2002; Johnson et al., 2009; Marcus, Fernandes, & Johnson, 2007; K. Smith, 1966). The strategy in these investigations is to assess the ability of participants to learn particular regularities; these regularities often vary along some dimension such as variability (Gómez, 2002) or position (Endress et al., 2005).

Although researchers intend to investigate the mechanisms of rule learning by seeing where people fail in these tasks, there is often another potential source of failure: memory demands. In the following paragraphs, we describe three examples of this covariation in some depth, in order to illustrate this point.

We gratefully acknowledge Denise Ichinco and Kelly Drinkwater for help with data collection, and Charles Kemp, Talia Konkle, LouAnn Gerken, Sharon Goldwater, Noah Goodman, Gary Marcus, Rebecca Saxe, Josh Tenenbaum, Ed Vul, and three anonymous reviewers for valuable discussion. This research was supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall, Building 420 (Jordan Hall), Stanford, CA 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu.

[1] Because the test items in rule learning tasks are usually novel stimuli that participants have never seen before, this kind of task contrasts with artificial segmentation (Saffran et al., 1996b, 1996a) and word learning (Vouloumanos, 2008; Yu & Smith, 2007; L. Smith & Yu, 2008) tasks in which learners are generally tested on items or pairings that are present in the familiarization stimuli.

In one recent example, experiments by Endress et al. (2005) investigated whether participants were better able to extract a repetition-based regularity of the form *ABCDEFF* or *ABCDDEF*. They found that the repetition-based regularity was better extracted in the final position and attributed this to a specific limitation on detection of repetitions in medial positions. As the authors acknowledge, however, this limitation could be due to more general serial position effects, since final elements of strings are recalled with greater frequency and accuracy than medial elements (Murdock, 1962). The computational demands of success in this task are extremely small: success merely requires recognizing a repetition—something that even newborns can do (Gervain, Macagno, Cogoi, Peña, & Mehler, 2008)—and maintaining a basic representation of its position in the string. Thus, the Endress task is a good candidate for investigations of the role of memory demands in creating a particular pattern of successes and failures, since alleviating memory demands should allow participants to succeed in the medial repetition condition.

Another example of the covariation of memory demands with other dimensions of interest comes from the experiments of Gómez (2002). Gómez investigated the ability of learners to extract an invariant regularity between the initial and final elements of a string of the form *aXb* or *cXd*. The experiments manipulated the number of exemplars of this regularity by changing the variability of the middle element of the string (the number of *X*s that were observed). They found that only when *X* was highly variable were participants able to learn that particular initial and final elements were paired, even though success in this task simply requires recognizing that every time a string begins with a particular initial element (e.g. *a*), it ends with the corresponding final element (e.g. *b*).

In Gómez's experiment, the more internal *X* elements there are—and hence the more meaningful variability there is in the *aXb* pattern—the more total string types are part of the experimental language. Though these two factors naturally co-occur they can be dissociated.[2] Hence, although seeing 12 *X*s may give more evidence for the *a_b* dependency than does seeing two *X*s, in order to appreciate that evidence a learner must be able to remember those 12 strings. We hypothesized that human learners might need even more variability than is strictly necessary from an informational standpoint to be able to remember enough strings to learn the regularities in Gómez's experiment. A test of this hypothesis would be to alleviate memory demands on human learners in the same paradigm. If memory demands cause human learners to need more variability than is informationally necessary for generalization, then reducing memory demands should allow participants to succeed in conditions with less variability than was necessary in the original paradigm. Note that this hypothesis differs in important ways from Gómez's own interpretation, which posited that generalization happened *because* of memory demands, not in spite of them. We return to this issue at length in the discussion of our own experimental results.

A final example of the phenomenon of memory demands co-varying with dimensions of interest comes from an experiment by K. Smith (1966). In this experiment, participants were presented with strings of the form *MN* or *PQ*. *M*s, *N*s, *P*s, and *Q*s each represented distinct and arbitrary sets of letters, leading to e.g. nine possible *MN* strings, of which six were presented. When participants were tested on what they had learned, however, they made systematic errors indicating that they had not distinguished the *M* class from the *P* subclass and would endorse strings of the form *MQ* and *PN* as coming from the language they had heard as readily as they endorsed previously-unheard *MN* or *PQ* strings. In this task, success requires at least two steps: first, using distributional evidence to discover the abstract classes (e.g., *M* or *N*) and second, learning the relationships between these classes. This experiment has had considerable influence on theorizing about human learning capacities, at least in part because it has been taken to represent a plausible approximation of the task faced by children in learning syntactic categories at the same time as they learn the way these categories interact in the grammar of their language (Braine, 1987).

As in the Endress and Gómez experiments, our hypothesis for why participants failed to learn in the Smith experiment is that the use of arbitrary stimuli makes it too difficult for participants to maintain a large enough number of exemplars in memory, and hence they are unable to perform any kind of comparison or grouping. Indirect support for this hypothesis comes from experiments in which experimental materials have provided extra semantic cues for category membership: in general, when semantic categories support category extraction, participants learn successfully (Morgan & Newport, 1981; Braine, 1987; Brooks, Braine, Cajalano, & Brody, 1993). Further evidence regarding this hypothesis comes from more recent work in which the addition of multiple, correlated linguistic cues to category structure also allows learners to extract categories and distinguish new legal category members from illegal members (Mintz, 2002; Gerken, Wilson, & Lewis, 2005). However, more direct evidence for the role of memory in the pattern of successes and failures in experiments of this form would come via direct manipulation of the memory demands of the paradigm.

In each of these experiments (and for the rest of the paper), when we refer to memory demands, we are referring to a specific, pre-theoretic conception of memory: the ability to retain the stimulus materials for long enough to learn. In the case of the Endress experiment, this retention might only be for the duration of a single string, such that once the string is heard, the learner can infer that the fourth and fifth elements were repeated and the others were unique. In the case of the Gómez experiment, some information about a particular

[2] Imagine a language of the form *aXbY* with a *Y* element that had a variability equivalent to the highest variability of the *X* element. This manipulation would equate the number of string types across conditions with different variability of the *X* element, dissociating memory and variability. (Thanks to Charles Kemp for this observation.) Note that covariation of memory demands and variability does not compromise Gómez's result; if anything the result is strengthened, since even under conditions with high memory demands, variability still leads to the extraction of meaningful regularities.

string would likely need to be maintained from one string to the next so that they could be compared to one another. Likewise in the case of the Smith experiment: several strings would likely need to be represented to make the appropriate comparisons to learn the class structure. Although there is a rich literature addressing human memory which we believe should be linked to the literature on artificial language learning in greater detail (for some important attempts, see e.g. Endress & Mehler, 2009; Perruchet & Vinter, 1998; Servan-Schreiber & Anderson, 1990), here we only address the question of whether stimulus retention—the most basic sense of memory—is a factor in previous artificial rule learning failures.

The goal of the current paper is to test the general hypothesis proposed above: that the pattern of successes and failures in artificial language learning experiments is affected by the ability of participants to retain the stimuli in these tasks. Put more simply: we speculate that participants may sometimes fail to learn for no reason deeper than that they cannot remember enough of the training set to make the appropriate inference. We make use of the three experimental paradigms we have already discussed (Endress et al., 2005; Gómez, 2002; K. Smith, 1966) to test this hypothesis.

To make a strong test of our hypothesis, we selected conditions from these experiments in which participants in previous experiments had failed to learn the correct rule. We then manipulated the mode of presentation of the stimuli in each of these tasks: participants either heard exemplars on an iPod, saw a printed list, or received exemplars on separate index cards that they could arrange as they pleased. We predicted that, if limitations on stimulus retention were the crucial bottleneck in extracting the appropriate regularity from the training materials, participants should be more likely to succeed in the conditions in which they received either the printed list or the index cards. We additionally predicted that receiving the index cards should be especially useful for tasks in which there was an element of clustering. Performing clustering tasks requires retaining the association of elements to clusters; in the index card condition this demand could be alleviated by arranging cards containing particular exemplars into spatial clusters.

The concurrent presentation of stimuli introduced a design challenge. There was no clear way to equate the amount of training participants received, creating a situation where differences in performance could potentially be mediated by differences in exposure. For example, if we equated the amount of exposure time, participants could potentially be reading visually presented sentences faster or slower than sentences were presented aurally. We resolved this problem by giving participants both the training and test materials at the same time, and allowing them both unlimited access to the training materials in all conditions and unlimited time to deliberate on their answers.

From an informational standpoint, allowing more exposure to evidence and more time for computation can only make a task easier; however, this manipulation also introduced another change in the demands of our task, encouraging participants to adopt an active, problem-solving or hypothesis-testing strategy rather than a passive, implicit strategy. To compensate for this issue, we made sure that the task was the same across all methods of exposure to the familiarization materials.

The structure of our argument is as follows. If the memory demands of a particular task are too great, participants should not be able to succeed even when they have explicit access to the test materials, unlimited exposure to the training set, and unlimited time to deliberate. And if participants do not succeed under these conditions, then the task is even more likely to be too demanding for learners to succeed under more stringent conditions. Further, if relieving the memory demands of the task allows adults to succeed in a task (even in the same explicit form), this result should provide further support for the claim that a crucial barrier to success is the memory requirement of that task.[3]

Methods

Participants

Forty-eight students and members of the MIT community participated in the study as part of a larger group of studies performed in exchange for payment.

Stimuli

Appendix A gives the full text of the survey we administered in the list condition; sentences were the same in the other two conditions, though they were presented via different modalities. Materials from three artificial languages were included in the experiment. These languages were titled "Flargian," "Gizalld," and "Zeepers," and they corresponded to the artificial languages used by Gómez (2002), Smith (1966), and Endress, Scholl, & Mehler (2005). For clarity we refer to these languages by the corresponding reference rather than by their invented names. Each language contained a set of training sentences and a set of test items.

The Gómez (2002) language was adopted directly from Gómez's design, using the $|X| = 6$ condition. The training set contained sentences of three forms, aXb, cXd, and eXf, where lower-case letters stand for the monosyllables "pel," "rud," "vot," "jic," "dak," and "tood" and X stands for the bisyllables "puser," "wadim," "kicey," "fengle," "coomo," and "loga." When each X was presented in each context, this created a total of 18 sentences. There were six test items for this language, testing three known sentences ("memory") and three novel sentences ("generalization"). Each test item began with a partial phrase from the language; known sentences contained X elements that were part of the training set and novel sentences contained the novel X elements "malsig," "skiger," and "hiftam." Four possible continuations were given for each test sentence; three were the elements b, d, and f (of which one was appropriate), and one was a novel bisyllable.

[3] Note however that this result would not support the contention that retention demands are the only barrier to success; as in the case of the Gómez (2002) experiment, we expect that other factors will also play a role.

The second language was based on the *MNPQ* language described in Smith (1966). It contained sentences of two forms, *MN* and *PQ*, where each letter denoted a set of three unrelated monosyllables. *M*s were "trund," "hram," and "zheep"; *N*s were "lipf," "frunt," and "klard"; *P*s were "narb," "qwun," and "junt"; and *Q*s were "ninz," "omf," and "shamp." Thus, a sentence like "trund lipf" was legal, while a sentence like "hram omf" was not. Fully expanded, this language contains 18 sentences. Our training set consisted of 14 of these 18 sentences, with the test set containing the other 4. For example, possible continuations for an *M* test item were another *M*, a *P*, a *Q*, and an *N* (the correct answer).

The third language was based on the languages used by Endress, Scholl, and Mehler (2005). All sentences had the form *ABCDDEF*, where each letter was an arbitrary CV monosyllable. Thus, the only regularity in each sentence was that the fourth and fifth syllables were the same. The training set consisted of 14 random sentences of this form. The test set consisted of 4 sentences with the fourth and fifth syllables omitted; of the four continuations, two were repetitions (both were correct) and two were not. We added two repetitions so as not to draw attention to one of the four options purely because it was a repetition.

Stimuli were presented via one of three methods. In the cards condition, all sentences in each training set were written on separate index cards. For each language, participants were encouraged to arrange the training set in whatever way was convenient for them. In the list condition, sentences in the training set were presented in a random order on the same page as the continuations (as in Appendix A). In the iPod condition, participants were given headphones and an Apple iPod Nano containing three playlists, each one corresponding to a language. Each playlist contained audio files in the WAV format of each sentence in the training set for that language. The iPod was set to shuffle (randomize) the songs within each playlist, thus playlists were presented in random order each time they were presented. All playlists were under 30s in length. Participants were encouraged to listen to each playlist as many times as they wished in order to make their bets, but not to pause the iPod in the middle of a playlist. Sentences were synthesized using the AT&T text-to-speech engine.[4]

Procedure

Participants were randomly assigned to one of three presentation conditions and asked to bet on the continuations presented in the test sets. Test materials were identical across presentation conditions. Participants were instructed to spread $100 between continuations to express their level of certainty that each was a possible continuation of the test item. We used the betting procedure to provide a detailed, explicit measure of participants' confidence in particular continuations. The betting response procedure was explained via the instructions in Appendix A. All participants were exposed to the three languages in the same order: Gómez, Smith, and then Endress.

Results

We summarized participants' responses by the average bets they placed on the correct continuation or continuations. These averages are shown in Figure 1. While bets were variable, there were systematic differences in responses across languages and conditions, suggesting that the betting method was successful in assessing participants' knowledge.

The basic pattern of results was similar across all three languages. The highest performance (highest mean bet on the correct continuations) was in the cards condition, followed by the list condition, and finally the iPod condition. The successes in the card condition in all three languages suggests that failures to learn these same languages—in two of the three languages by participants in our iPod condition and in all cases by participants in the original experiments—were likely caused by an inability to remember exemplars or intermediate steps in computations.

The magnitude of the differences between presentation modalities varied across languages, however. In the following paragraphs we discuss the results from each language in detail. Because of the wide dispersion of participants' responses as well as the fact that they fell on a fixed interval from $0 to $100, the distribution of means was not appropriately modeled using a Gaussian distribution. Thus, throughout our discussion we use non-parametric statistics which do not rely on facts about the underlying form of the response distribution. These tests are more conservative than their parametric equivalents; in all cases, the corresponding parametric test gives a higher level of significance than the test we report.

In the Gómez (2002) language, participants' mean bets in the cards condition were close to ceiling (93.4), slightly lower in the list condition (80.9), and considerably lower in the iPod condition (58.4). This difference between conditions was statistically significant in a Kruskal-Wallis test (a non-parametric one-way ANOVA, $\chi^2 = 11.97$, $p = .003$). The planned comparison of the cards and list condition trended towards statistical significance in a Wilcoxon rank sum test (a non-parametric test for equivalence of medians, $z = 1.73$, $p = .08$) and the contrast of the list and iPod conditions was statistically significant ($z = 2.08$, $p = .04$).[5] The iPod condition differed significantly from chance in a sign-rank test ($z = -2.95$, $p = .003$). We computed a non-parametric measure of effect size for these contrasts, Cliff's *d* (Cliff, 1993). Cliff's *d* estimates the probability that a sample from one distribution dominates (is greater than) a sample from another (*d* varies from 0 to 1). The d values for the cards/list contrast and list/iPod contrasts were .34 and .43, respectively, indicating relatively large effect sizes, despite the wide variation in responses. In our manipulation of memory (familiar intermediate elements) and generalization (novel intermediate elements), there were small but consistent decrements in performance in the generalization condition (95.2 vs. 91.5, 85.6 vs. 76.3, and 59.9 vs. 56.9, in each

[4] Previously available at `http://www.research.att.com`.

[5] We use z as a test statistic for the Wilcoxon rank-sum test; this approximation is appropriate when both samples have $N > 10$.

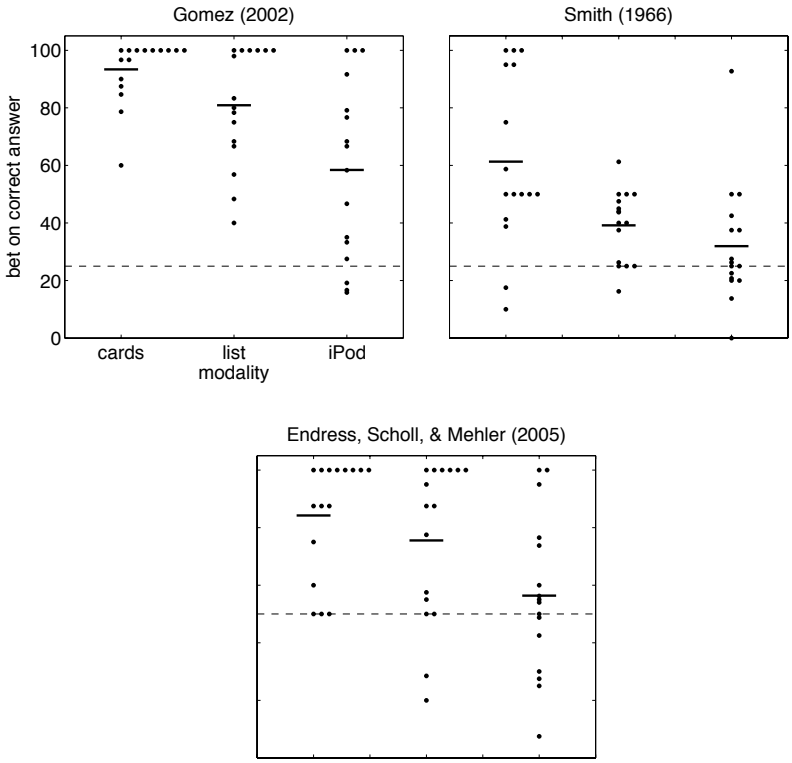

Figure 1. Participants' mean bets on correct answer(s), plotted by the modality of presentation. Subplots show results for the three languages tested. Dotted lines indicate chance levels of betting. Horizontal marks indicate the mean for participants in a particular condition. Points are stacked horizontally to avoid overlap when more than one participant had the same mean for the same condition.

of the three modality conditions, respectively), but this difference only showed a trend towards statistical significance, and only in the list condition (in a paired sign rank test, $z = 1.68$, $p = .09$). Because both kinds of test trials can be solved via simple memorization of the relationship between *a* and *b* elements (as noted above), this result provides only weak evidence that participants truly generalized a novel regularity.

In the Smith (1966) language, the overall pattern was similar but the list condition grouped with the iPod, rather than the cards condition. Means were 61.3, 39.2, and 31.9, respectively. The manipulation of condition was again statistically significant ($\chi^2 = 11.99$, $p = 0.003$), with the cards/list contrast significant ($z = 2.47$, $p = .01$, $d = .51$) and the list/iPod contrast now trending towards significance ($z = 1.74$, $p = .08$, $d = .37$). The iPod condition did not differ significantly from chance. There were two notable patterns in the cards condition. First, five participants were very close to ceiling, indicating that they had definitely learned the correct regularity (the probability of 5 of 16 participants correctly answering 4 of 4 four-alternative forced choice judgments is astronomically low). Second, another grouping of five participants found the positional regularity reported by Smith (1966): they hypothesized that Ms could be followed by Ns and Qs, but not Ps (this strategy would result in betting equally on the N and Q options, hence the bet of 50 on the correct answer). Thus, this task was considerably more difficult than the other two languages. Even in the cards condition, the majority of participants were not at ceiling.

In the third language, based on Endress, Scholl, & Mehler (2005), we found the same basic pattern. Performance was highest in the cards condition; in this language as in the Gómez language, the list condition grouped with the cards condition. Means were 84.2, 75.5, and 56.4, respectively. Because there were two correct answers for each question, chance betting was 50, rather than 25 as in the other conditions. The manipulation of condition was again statistically significant ($\chi^2 = 8.24$, $p = .01$), with the cards/list contrast not significant ($z = 0.81$, $p = .42$, $d = .16$) and the list/iPod contrast trending towards significance ($z = 1.88$, $p = .06$, $d = .39$). Performance in the iPod condition did not differ significantly from chance.

To summarize the results from all three languages: increasing the memory resources available to participants by allowing them to visualize and manipulate all the sentences of the language concurrently led to better extraction of the target rules for each language. Participants were able to succeed in extracting rules using this method even in cases that had been difficult in previous investigations of similar languages.

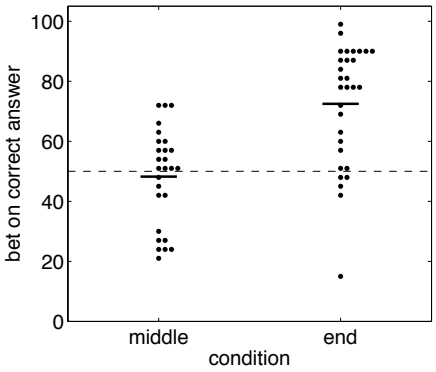

Figure 2. Participants' mean bets on correct answers for our Amazon Mechanical Turk replication of Endress, Scholl, and Mehler (2005), Experiment 1. Dotted lines indicate chance and horizontal marks indicate the mean for participants in a condition.

Discussion

In the discussion of our results we focus on two particular issues: first, the possibility that our results were caused by modality differences; and second, the computational demands of success in the Gómez task.

Modality differences

One possible concern about our results is that the differences we observed could have been caused by differences in presentation modality instead of by true differences in memory demands across the stimuli. For instance, the difficulty of hearing auditory training stimuli while being tested with written materials might have caused the failures we observed in the iPod condition. To control for this possibility we conducted a separate experiment in which we used auditory stimuli and an explicit, written test method to replicate the pattern of success and failure observed by previous work conducted solely in the auditory domain. Because there have not been successes reported using the unmodified Smith (1966) language and because our exposure corpora would need to differ considerably in length for the Gómez (2002) language (due to our stimulus set, which contained no repeated tokens), we decided to replicate the results of Endress et al. (2005).

For this replication we used Amazon Mechanical Turk, an online crowd-sourcing website, to recruit 64 participants. We gave participants the same instructions and test sheet as in the original in-lab study (although with only the Endress language test questions on it) and asked them to listen to a WAV file containing either the same set of strings that participants heard in our main experiment or a comparable set with the repeated element at the end of the string. Strings were played in random order. To ensure that all participants listened attentively to the full WAV file, we embedded three common English words in the stream and asked participants to select the words that were embedded in three 3AFC catch trials. We excluded 8 participants on the basis of one or more incorrect responses to the catch trials; all other participants made correct responses on all three trials, indicating that they were listening to the sound file.

Participants in the middle position condition (the same condition which we ran in the lab in our main experiment) were at chance as before ($M = 48.0$), while participants in the end-position condition correctly learned the repetition regularity ($M = 72.3$). This difference was statistically significant ($z = 4.12$, $p < .0001$). Data are plotted in Figure 2. This control experiment replicates Endress, Scholl, & Mehler's original result and suggests that the differences between the iPod condition and the other two conditions we observed in our experiment were not due to issues with our auditory stimuli or difficulties at test because of the difference in modalities between training and test materials.

There is a more general issue, however, which our control experiment does not address: while our visual stimuli were presented concurrently, our auditory stimuli are sequential. There are therefore two contrasts between the auditory condition and the other two conditions. Might it be the case that the differences we observed have to do with different specialized pattern recognition mechanisms in vision rather than audition, instead of memory demands caused by sequential presentation (as we have argued)? Although we cannot rule out this hypothesis completely, we believe that our data (and the broader set of rule learning findings) do not support an explanation on the basis of modality-specialized pattern recognition mechanisms such as differential generalization ability for visual rather than auditory stimuli for at least two reasons.

First, we observed significant differences even within the two visual presentation methods in the Smith language (where the spatial arrangement of exemplars likely relieved memory demands). A simple main effect of modality does not explain the data in the Smith language. Second, the general picture that has emerged from the infant rule learning literature is that young infants are able to learn simple regularities of the form *ABB* or *AAB* across a wide range of modalities, including sequentially presented speech (Marcus et al., 1999) and musical stimuli (Dawson & Gerken, 2009), as well as sequentially- (Johnson et al., 2009) and simultaneously-presented (Saffran, Pollak, Seibel, & Shkolnik, 2007) visual stimuli. There are complex asymmetries across modalities (for example, musical stimuli seem to be relatively easier for younger infants to process, perhaps because infants are learning expectations the structure of different domains, e.g. Dawson and Gerken, 2009) but in general there seem not to be restrictions on the kinds of rules that can be learned in particular domains (but c.f. Marcus et al., 2007)

Nevertheless, the visual and auditory modalities are suited to different styles of stimulus presentation. Although sequential visual processing is supported by visual short-term memory, details of the retinal image are only stored in a quickly-fading iconic memory. In contrast, the auditory-phonological loop stores stimuli relatively veridically for a period of up to several seconds (Baddeley, 1987). This difference has important consequences for the memory of stimuli in those modalities (Boutla, Supalla, Newport, & Bavelier, 2004). As a

thought-experiment, imagine sequential, orthographic visual presentation of the Endress language we used, where syllables in a string are flashed one at a time on the screen. It is easy to see that there would be no way to maintain the image of seven nonce words in visual short-term or iconic memory. Instead, orthographic forms would be converted to phonological forms and stored in the phonological loop, exactly what is assumed to happen in digit span tasks with sequential orthographic stimuli (Baddeley, Thomson, & Buchanan, 1975).

Our thought experiment suggests that it is unlikely that it is the visual presentation *per se* that allowed our participants to succeed in the two visual/simultaneous conditions. Instead, it is the simultaneous presentations that allowed repetitions and dependencies to become visible. (This is after all why good scientific practice involves plotting data, rather than dealing purely with numerical representations). It is not a simple issue whether the ease of finding relationships in simultaneous presentation is a fact about memory or a fact about visual processing, since simultaneous presentation in the auditory domain is not possible in the same way. Nevertheless, the key limiting factor in the current case seems to be the accessibility of the stimuli for processes of comparison and generalization—what we have referred to here as memory limitations.

Variability and memory in the Gómez language

The next question we address is about the relationship between memory demands and variability in the Gómez language. This language differs in an important respect from the other two languages we studied: it involves recognizing a dependency at test that is also present in the training items, rather than generalizing a regularity that does not share all of its parts with the test items. (In fact, in Gómez's original study, no test items included novel material, so they were identical to the training stimuli and could be remembered verbatim). Thus, a crucial part of understanding human performance in this language is understanding why participants failed, even under conditions of low variability where they potentially could have memorized all the training sentences. Gómez explained this result in terms of a shift in attention from adjacent dependencies between syllables—which do not distinguish legal from illegal sequences in this language—to non-adjacent dependencies. On this account, the cause of the shift is the decrease in element-to-element predictability that comes when the set of intermediate elements is large.

In our work here we have described a different explanation of this phenomenon that appears—at least superficially—to be at odds with Gómez's original explanation. We posited that participants' performance in the task is due to two factors: (1) the necessity of a certain minimal set of examples to support the correct generalization, and (2) the difficulty of remembering larger sets of examples. With only a single example from each category in the language, learners would have no reason to suspect that the non-adjacent *a_b* dependency was definitional. With two examples, it would seem possible, and with six it would seem certain. On the other hand, remembering one or two examples seems easy, but remembering all six might be more challenging. Thus, these two factors are at odds with one another.

In a computational study providing an explicit version of this explanation, Frank and Tenenbaum (in press) created a model that implemented an ideal observer for rule learning.[6] Their model evaluated the relative simplicity and fit to data of different hypothesized regularities. They found that the model showed the same effects as human learners in the Gómez (2002) paradigm: with greater variability, the rule learning model was able to generalize the correct regularity. With no restrictions on memory, their model was able to generalize successfully with two *X* elements and showed perfect generalization with six *X*s (as in our example). One goal of that modeling work, however, was to investigate the effects of memory on performance in rule learning tasks. The simple memory model used by Frank and Tenenbaum assumed that learners have a constant probability of misremembering any given training example. In the rule learning simulations using this memory model, as the probability of misremembering individual examples increased, the number of examples that were necessary for successful generalization also went up. At a relatively high level of misremembering, model performance matched adult performance in the Gómez experiment. Thus, in this model, the conflict between the value of extra evidence and its memory cost resulted in a pattern of performance similar to the human data.

The current experiments test a prediction of the memory-model account of the Gómez results: alleviating memory limitations should decrease the number of examples necessary for successful generalization. Our data suggest that this prediction was confirmed. However, we only partially replicated Gómez's original results: unlike the original pattern of data, our participants' performance was above chance even in the iPod condition. One possible reason for participants' performance in the iPod condition is that, as noted above, the demands of success in the Gómez task are so low: all that is required is to fill in the same final element in a test string as has been observed in the most similar training string (e.g., when hearing aX_1b, match the blank in $aX_7\_$ to *b* rather than to *d*. This matching would not require remembering more than one example and hence could certainly be carried out even in the iPod condition. (It also would not have been possible in Gómez's original experiment, because the test materials in that study were not present at the same time as the training stimuli). Nevertheless, we did observe significant differences between the iPod condition and other conditions, suggesting that our results are likely due to both an effect of template-matching and an effect of generalization in the list and cards conditions but not the iPod condition.[7]

While our current results provide one datapoint in favor of some kind of memory-limitations account (with the

[6] A previous version of this model was reported in Frank, Ichinco, and Tenenbaum (2008).

[7] Thanks to an anonymous reviewer for careful discussion of this issue.

template-matching caveat mentioned above), the issue is still not resolved. In particular, the model of memory described in Frank and Tenenbaum (in press) was designed to be the simplest possible memory model that could capture the relevant phenomena in rule learning; it does not capture many empirical facts about memory. Particularly relevant to the current issue, human memory does not show a constant probability of forgetting examples, irrespective of quantity. Taken to its extreme this view makes predictions that are almost certainly untrue (e.g., that you are equally likely to remember a particular string when it is the only string you hear as when you hear it in a context of ten others). Thus we believe that exploring more realistic models of memory is a crucial part of distinguishing between these competing accounts.

General Discussion

We alleviated the memory retention demands in three separate artificial rule learning tasks by presenting exemplars from the languages concurrently rather than sequentially. This manipulation allowed participants to succeed in circumstances when they would otherwise have failed. The success of participants in the index card and list conditions was not caused by differences in explicitness of task between our experiment and previous work: participants still performed more poorly in the iPod condition, where instructions were identical but stimuli were presented sequentially. It was also not caused by idiosyncrasies of our auditory stimuli or method, as demonstrated by a control experiment in which we successfully replicated previous work. We conclude that the decreased retention demands in these conditions allowed participants to succeed, suggesting that the mechanisms the participants were using depended crucially on memory resources (and hence that previous failures may have been due to the retention demands of the tasks). Future research will be needed to establish whether our conclusions can be extended to rule learning in infancy.

What aspect of the index card condition made performance in this condition consistently higher than in the list condition (especially in the challenging Smith language)? We speculate that the ability to arrange the cards spatially into clusters and groups allowed participants not only to offload the storage of exemplars but also to store the products of the intermediate computations necessary in the Smith language. Extracting the structure of this language requires finding four clusters of words (or two clusters of sentences) based on their distributional properties. Remembering these clusters requires a prohibitive amount of memory resources even if the structure is known in advance. For a simple demonstration, try to find the correct answers for the language presented in Appendix A. Without drawing lines, pointing, or mumbling to oneself, it is very difficult to identify the correct continuations even if one knows what to look for. Using the spatial arrangement of cards to represent clusters, however, makes the task far easier. While we did not record the spatial arrangements used by participants in the current experiment, in two pilot tests we observed that solvers of the Smith and Gómez languages used spatial arrangement to represent clusters of sentences or words.

Though relatively novel in artificial language studies, the general method of distributing cognitive load onto a physical system is well-known in other parts of cognitive science. For example, successful interfaces for high-risk applications like plane flight use many related methods to decrease the possibility of error. In Hutchins' classic study of an airplane cockpit as a cognitive system, he notes that one of the primary goals of the design of the cockpit and the landing procedure is to reduce demands on the memory and attention of the pilots while allowing them to perform a complex series of actions (Hutchins, 1995). In the same way, learners in the index card condition of our study are able to focus on one aspect of the learning task at a time (e.g., what words cluster with *M* words) without danger that they will lose track of other aspects (what words are in the *P* cluster).

How do our results (and those reported in the broader artificial language learning literature) bear on the issue of natural language acquisition? At their best, artificial language studies can be highly informative about the fundamental mechanisms of learning that are continuous across development and even across species. Work on statistical segmentation and grouping has exemplified this description (Saffran et al., 1996b, 1996a; Aslin, Saffran, & Newport, 1998; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Fiser & Aslin, 2002). Although it is not always clear how this work connects with particular tasks in language acquisition, the identification and characterization of basic learning mechanisms is in itself an important task. We hope that our work here falls within this broad project by helping to characterize interactions between learning and memory that may limit learning in a broad range of natural situations.

On the other hand, a separate line of argument in this literature has attempted to construct analogues to particular situations in language acquisition and to argue for *limitations* on learning in these situations. Although this work can be informative, its interpretation often rests on untested hypotheses linking the experimental situation to the task faced by language learners. We give two examples of this phenomenon. First, the Smith language used here has been taken to be representative of the joint task of learning syntactic categories and learning the rules linking these categories (Braine, 1987). Under alternative characterizations of the acquisition of syntax (e.g. Tomasello, 2003; Goldberg, 2003; Bannard, Lieven, & Tomasello, 2009), however, the Smith task is not a valid representation of the task faced by children. Second, although various linking hypotheses have been proposed for the original "rule learning" experiments by Marcus et al. (1999), the identity regularity used in these experiments is at most a minor part of the regularities found in natural languages, and attempts to generalize to other types of regularities have not always been successful (Gomez, Gerken, & Schvaneveldt, 2000; Endress et al., 2007). In both cases, a considerable amount of future work in both natural and artificial domains is necessary in order to understand how the initial results should be interpreted.

Experiments in artificial language learning have often

drawn conclusions from the failures of participants to learn from input sentences with particular characteristics. Our results suggest that this kind of negative finding may sometimes reflect basic limitations on learners' memory rather than limitations on the kinds of computation that learners can carry out. Because of this possibility, failures alone cannot be used to argue for intrinsic limitations on the learning mechanisms available in artificial—or natural—language learning tasks without appropriate controls for memory effects. Our conclusion does not imply that memory effects are merely a roadblock to progress towards understanding more fundamental learning mechanisms, however. Instead, in shaping the overall progress of acquisition, limits on what can be remembered may be just as important as limits on what can be learned (Newport, 1990; Perruchet & Vinter, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, in press). Thus, we hope that our characterization of memory effects in rule learning will lead to further investigation of interactions between the architecture of human memory and the mechanisms of language acquisition.

References

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 321–324.

Baddeley, A. (1987). *Working memory*. Oxford, UK: Oxford University Press.

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575–589.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284.

Bates, E., & Elman, J. (1996). Learning rediscovered. *Science*, *274*, 1849–1850.

Boutla, M., Supalla, T., Newport, E., & Bavelier, D. (2004). Short-term memory span: insights from sign language. *Nature Neuroscience*, *7*(9), 997–1002.

Braine, M. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. *Mechanisms of language acquisition*, 65–87.

Brooks, P., Braine, M., Cajalano, L., & Brody, R. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of memory and language*, *32*, 76–76.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494–494.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378–382.

Endress, A., & Bonatti, L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*, 247–299.

Endress, A., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351–367.

Endress, A., Scholl, B., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*, 406–419.

Fiser, J., & Aslin, R. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822.

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (in press). Modeling human performance in statistical word segmentation. *Cognition*.

Frank, M. C., Ichinco, D., & Tenenbaum, J. (2008). Principles of generalization for learning sequential structure in language. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society.*

Frank, M. C., Slemmer, J., Marcus, G., & Johnson, S. (2009). Information from multiple modalities helps five-month-olds learn abstract rules. *Developmental Science*, *12*, 504.

Frank, M. C., & Tenenbaum, J. B. (in press). Three ideal observer models of rule learning in simple languages. *Cognition*.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(02), 249–268.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, *105*(37), 14222.

Goldberg, A. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, *7*(5), 219–224.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 431–436.

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Gómez, R., & Gerken, L. (2000). Infant artificial language learning and acquisition learning. *Trends in Cognitive Science*, *4*, 178–186.

Gomez, R. L., Gerken, L. A., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory and Cognition*, *28*, 253–263.

Hauser, M., Newport, E., & Aslin, R. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), 53–64.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265–288.

Johnson, S., Fernandas, K., Frank, M., Kirkham, N., Marcus, G., Rabagliati, H., et al. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*, 2.

Kirkham, N., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), 35–42.

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

Marcus, G., Fernandes, K., & Johnson, S. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*, 387.

Marcus, G., Vijayan, S., Bandi Rao, S., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77.

Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*, 678–686.

Morgan, J., & Newport, E. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning & Verbal Behavior. Vol*, *20*(1), 67–85.

Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *10*, 20–21.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.
Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(246-263).
Saffran, J., Aslin, R., & Newport, E. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926.
Saffran, J., Aslin, R., & Newport, E. (1996b). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*, 376.
Saffran, J., Pollak, S., Seibel, R., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680.
Servan-Schreiber, E., & Anderson, J. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16*, 592–608.
Smith, K. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.
Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.
Tomasello, M. (2003). *Constructing a language: A usage-based account of language acquisition*. Cambridge, MA: Harvard University Press.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742.
Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414.

Appendix A: Experimental Materials

Instructions

Welcome! In this experiment, you will be asked to predict the next word in a phrase from one of several alien languages. To help you in this task, we will give you a set of similar sentences in the same language. For each language, try to read over the sentences carefully and then make your predictions.

Your predictions should take the form of bets. Imagine that you have $100 to spend. You should divide it up among the possible completions of the phrase so that you place more money on each option corresponding to how confident you are that it correctly completes the phrase.

Language #1: Flargian

```
pel puser rud
dak wadim tood
pel kicey rud
dak kicey tood
vot wadim jic
pel fengle rud
vot kicey jic
pel coomo rud
vot loga jic
dak coomo tood
vot coomo jic
dak fengle tood
pel wadim rud
pel loga rud
vot puser jic
dak loga tood
vot fengle jic
dak puser tood
```

Test for Language #1

```
Phrase Bet 1 Bet 2 Bet 3 Bet 4
vot puser...
deecha ___ tood ___ jic ___ rud ___
dak coomo...
rud ___ gensim ___ tood ___ jic ___
pel skiger...
tood ___ jic ___ rud ___ roosa ___
dak hiftam...
rud ___ fengle ___ tood ___ jic ___
vot malsig...
fengle ___ tood ___ jic ___ rud ___
pel kicey...
gople ___ rud ___ tood ___ jic ___
```

Language #2: Gizalld

```
trund lipf
hram frunt
trund klard
narb ninz
zheep lipf
narb shamp
qwun ninz
junt shamp
trund frunt
hram klard
zheep frunt
narb omf
qwun omf
junt ninz
```

Test for Language #2

```
junt...
omf ___ narb ___ zheep ___ klard ___
qwun...
narb ___ trund ___ frunt ___ shamp ___
zheep...
hram ___ junt ___ klard ___ ninz ___
hram...
qwun ___ lipf ___ shamp ___ trund ___
```

Language #3: Zeepers

```
zu du ga za za gu zi
mu li zi ru ru ku ki
ri gu gi ta ta ga ni
lu ru di gi gi fi ti
du ka ki gi gi nu ma
ga fu di nu nu gu za
mi li ki ku ku na zu
ga ma mi fu fu lu fi
na la ra gi gi ti da
```

fi ma ra ti ti fa ga
la ga ku li li gu zu
ta zi ku mi mi ri ti
ri ra ku fu fu zu fa
ri li tu ni ni ka zi
ri za zi di di du gi

Test for Language #3

li fa mi...

ma na zi ki ___ fa fa ___ ku ku ___ ra ti ___
du fa lu...
mi gi ta la ___ gi gi ___ za gu ___ ta ta ___
mu di ta...
ma ri ku mi ___ za za ___ na la ___ ki ki ___
da ri nu...
zi ni li li ___ ru ru ___ li ki ___ fi ma ___