Statistical learning in reading:
Variability in irrelevant letters helps children learn phonics skills

Keith S. Apfelbaum
Bob McMurray
and
Eliot Hazeltine

Dept. of Psychology and Delta Center
University of Iowa

Running Head: Statistical Learning and Reading

Corresponding Author:
Keith S. Apfelbaum
E11 SSH
Dept. of Psychology
University of Iowa
Iowa City, IA 52240
keith-apfelbaum@uiowa.edu
319-335-0692

Abstract

Early reading abilities are widely considered to derive in part from statistical learning of regularities between letters and sounds. Although there is substantial evidence from laboratory work to support this, how it occurs in the classroom setting has not been extensively explored; there are few investigations of how statistics among letters and sounds influence how children actually learn to read, or what principles of statistical learning may improve learning. We examined two conflicting principles that may apply to learning grapheme-phoneme-correspondence (GPC) regularities for vowels: 1) variability in irrelevant units may help children derive invariant relationships and 2) similarity between words may force children to use a deeper analysis of lexical structure. We trained 224 first-grade students on a small set of GPC regularities for vowels, embedded in words with either high or low consonant similarity, and tested their generalization to novel tasks and words. Variability offered a consistent benefit over similarity for trained and new words in both trained and new tasks.

Statistical learning in reading: Variability over similarity

Substantial research attests to the efficacy of focusing on decoding for both reading instruction and intervention (Ehri, Dryer, Flugman, & Gross, 2007; Ehri, Nunes, Stahl, & Willows, 2001; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Torgeson et al., 2001). Learning the mappings between orthography and phonology enables students to read words they haven't seen before, and eventually recognize many words by sight. These mappings are typically described as grapheme-phoneme-correspondence (GPC) regularities (or consistencies). For example, the vowel E is pronounced as /ɛ/ in a word like BED, but as /i/ when paired with an A as in BEAD. Acquiring these regularities in English is difficult, as they are only quasi-regular (Seidenberg, 2005): there are exceptions (e.g., BEAR) and sub-regularities within the exceptions (e.g., DEAD, THREAT, LEAD).

Typically, many GPC regularities are taught by explicitly describing them, and then providing examples and activities as reinforcers. This is practical and successful for many children. However, it is unlikely that all GPC regularities can be taught this way. The Dual-Route Cascade (DRC; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) model of reading uses over 1000 GPC rules to simulate English reading, far more than can reasonably be taught explicitly. Thus, there is likely substantial learning that must occur via implicit mechanisms.

Given the complexity of these regularities and the scope of the learning problem, it is unsurprising that many students fail to acquire basic reading skills (U.S. Dept. of Education, 2010). This suggests a need for instructional innovation. Yet despite decades of cognitive science research on visual word recognition, decoding and their acquisition (e.g., Coltheart, Curtis, Atkins, & Haller, 1993; Glushko, 1979; Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989), few approaches to *teaching* decoding are based on this research (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

Contemporary learning theories hold that many behaviors that appear rule-governed, like reading, may be driven by implicitly-learned statistical regularities (Elman, 1990; McClelland & Patterson, 2002; Seidenberg & McClelland, 1989). Statistical learning need not be pre-tuned to specific statistics (Gómez, 2002); rather, learners can harness statistics over seemingly irrelevant elements (e.g., the consonant frame in a vowel GPC regularity) to find the right statistics. Although not uncontroversial, such mechanisms can explain rule-like behavior in many domains, including verb morphology (McClelland & Patterson, 2002; Plunkett & Marchman, 1991; though see, Pinker & Ullman, 2002), relational semantics (McClelland & Rogers, 2003), and, most pertinently, GPC regularities for reading (Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989; though see, Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

Connectionist models of reading (Harm, McCandless, & Seidenberg, 2003; Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989; Seva, Monaghan, & Arciuli, 2009) encode statistical regularities across graded connections between representations, yet their emergent behavior appears rule-like. These models demonstrate many behaviors akin to those of developing readers. For example, they over-generalize regularities to novel words, as seen in both typically-developing and struggling readers (Harm, et al., 2003; Harm & Seidenberg, 1999), and show sensitivity to subtle orthographic cues to lexical stress (Seva, et al., 2009).

Laboratory learning paradigms have been used to examine statistical learning more generally, by manipulating statistics over a small set of items and giving participants short-term (usually passive) exposure to these items. Such studies demonstrate that adults and children encode the statistical patterns across a range of domains (for a revew, see Saffran & Thiessen, 2007), although reading has not been examined in this way. For example, this type of statistical learning can be seen in word segmentation (Saffran, Aslin, & Newport, 1996), cross-word regularities (Gómez, 2002), auditory tones (Creel, Newport, & Aslin, 2004), visual scenes (Fiser & Aslin, 2002) and motor sequences (Hunt & Aslin, 2001). Moreover, participants can learn a range of statistical relationships including contingencies between sequential (Saffran, et al., 1996) and non-adjacent items (Gómez, 2002), statistical distributions (Maye, Werker, & Gerken, 2002), and associations between words and objects (Yu & Smith, 2007). Because this is accomplished without conscious access to abstract rules, statistical learning is closer to procedural or implicit learning than explicit learning (Cleeremans, 1997); However, the breadth of domains suggests that statistical learning may not be a single monolithic process, but rather a description of a set of learning mechanisms. As a result, the specific statistics that may apply to reading must still be determined.

Statistical Learning and Reading. Despite the aforementioned studies demonstrating statistical learning in complex learning tasks, there are few analogues of this type of work in reading. As a result it is unclear how GPC regularities are learned to support reading. There is substantial, albeit indirect, evidence that implicitly-learned regularities underlie reading and spelling. The logic here is that statistical regularities between letters (e.g. transitional probabilities between letters) and between phonemes and letters predict a variety of measures of performance in both adults and children, suggesting that the necessary skills and/or representations may have been acquired in this way. For example, spelling and word recognition are sensitive to statistical regularities between letters and sounds (Andrews & Scarratt, 1998; Glushko, 1979; Näslund, 1999; Treiman & Kessler, 2006; Treiman, Kessler, & Bick, 2003). Arciuli and colleagues have shown that readers are also sensitive to regularities between a word’s orthography and its grammatical form (Arciuli & Monaghan, 2009), and between a word’s orthography and its stress pattern (Arciuli, Monaghan, & Seva, 2010; Seva, et al., 2009). Further, children with no explicit decoding training make spelling errors that reflect the statistics of the texts used in their classes (Thompson, Fletcher-Flinn, & Cottrell, 1999). Thus, the traces of statistical learning are observable in reading performance.

There are also clear links between reading ability and statistical learning more broadly. Arciuli and Simpson (2012) tested children and adults in a visual statistical learning task and collected standardized reading measures. Participants who were more sensitive to statistical patterns in the visual task also showed better reading abilities. This suggests a link between a domain-general statistical learning capacity and reading outcomes, thus providing an important source of evidence for the statistical nature of early reading acquisition.

While these investigations show that learners encode statistics from input, and that their ability to do so more generally is related to reading outcomes, we know little about the more mechanistic question of how children engage statistical learning to acquire decoding skills. Few studies address which statistics are most effective or how to use these principles in curriculum development. In this regard, one notable study by Arciuli and colleagues (2010), analyzed which portions of words carried the most information about stress. They performed corpus analyses of the reading materials that children of different ages are exposed to, and measured children’s sensitivity to statistical patterns at different ages. This showed that children are attuned to the statistical structure of the material they are exposed to at a given age during learning. This sort of “natural experiment” shows how differing classes of statistics affect behavior. However, it remains unclear how to harness our understanding of statistical learning to promote learning.

Crucially, none of the aforementioned studies manipulate the statistics provided during learning, so most evidence for statistical learning is correlational. There are no findings to guide the manipulation of statistics in the learning environment to optimize the development of reading skills. Such data could form an essential empirical foundation for effective decoding instruction. This requires *in situ* demonstrations that statistical principles describe learning to read in the classroom and investigations of which principles can most effectively promote learning.

If acquisition of reading skills is in part statistical, this suggests that in standard classroom approaches, reinforcing activities that emphasize the statistics across words are as important as explicit description of GPC regularities, as these activities provide the basis of implicit learning. Moreover, if mappings are encoded in terms of probabilistic relationships between letters and sounds (rather than rules), practice with the regularities should offer a better platform for learning than the explicit rules that are given during instruction. Thus, the statistics over the words used in reinforcement and training activities—even statistics over seemingly irrelevant elements—could have important effects on learning outcomes. But what sort of statistics? The principles that may be relevant to reading derive from laboratory studies and computational models. However, while applying these principles to classroom instruction could benefit students, such extensions are rare. This is in part because transferring learning principles to complex domains like reading is not straightforward, and the practical implications are sometimes contradictory.

Variability and similarity. A prime example of this (and the focus of our study), is the contrast between variability and similarity among elements that are seemingly irrelevant to the regularities being learned. These conflicting principles can both improve learning in some tasks. However, this work has largely been conducted in simple domains, and it remains unclear how these principles scale up to complex tasks like learning to read.

Numerous studies have shown that *variability* in seemingly irrelevant elements helps learners identify relevant information. Gómez (2002) showed that learning dependencies between words that span an intervening word can be improved by increased variability in the intervening word. Similarly, Rost and McMurray (2009, 2010) showed that variability in talker voice improves early word learning (see also Lively, Logan, & Pisoni, 1993). Variation in irrelevant perceptual cues even helps pilots learn to land planes (Huet et al., 2011). According to these studies, when trying to teach children a GPC regularity involving vowels (e.g., A is pronounced /æ/ in a CVC frame), variation in the (mostly irrelevant) consonant frames may help.

Yet variability is not always beneficial; studies on learning second languages (Perrachione, Lee, Ha, & Wong, 2011) and motor skills (Wulf & Shea, 2002) suggest that variability is less effective for complex skills or for novice learners. Indeed, other studies have found benefits for *similarity*, where overlapping stimuli lead learners to identify minute features that differentiate stimulus classes (Schyns, Goldstone, & Thibaut, 1998). Similarly, work with infants suggests that 4 month-olds learn categories better with highly-similar training exemplars (Oakes, Coppage, & Dingel, 1997; Quinn, Eimas, & Rosenkrantz, 1993). Low variability forces learners to locate the few features that differentiate similar stimuli (Ahissar & Hochstein, 2004) and thus extract a more invariant structure. This is quite similar to some theories of early phonological development (Charles-Luce & Luce, 1990; Metsala & Walley, 1998). Finally, work in categorization suggests that comparison may help locate diagnostic features, and that comparison is most effective when items have similar non-diagnostic features (Goldstone, 1996). The similarity principle thus predicts that to teach the vowel GPC regularities, one should use words that have highly similar consonant frames.

In part this debate may boil down to how we frame the problem of discovering GPC regularities. Variability is often invoked when learners must detect invariant mappings among noisy irrelevant elements, whereas similarity is invoked in categorization and discrimination where participants must discover what separates an often small number of stimulus classes.

It is unclear which better characterizes decoding. However, determining how these principles operate in decoding could guide us to more effective pedagogical decisions and inform our understanding of reading development. Many standard phonics approaches seem to embrace similarity, using tasks like "word families" that highlight similarity between words (Baumann, Hoffman, Ro, & Duffy-Hester, 1998); and McCandliss, Beck, Sandak and Perfetti (2003) also showed improvements in reading after the *Word Building* intervention, which embraces similarity and word-family principles. However, *Word Building* and word families tasks have not been compared to a set of similar tasks using variability. Conversely, Gibson's (1970) early work on reading suggests that children can learn orthographic regularities in the context of variable stimuli; however, this too was not contrasted against similarity. Thus, the existing work on in reading offers no better resolution of this debate than the aforementioned laboratory studies.

The Laboratory and the Classroom. To compare these principles and to address the need for experimental studies of statistical learning in reading requires a short-term laboratory-learning paradigm in which stimuli and tasks can be controlled precisely, feedback administered consistently, and learning tested in a uniform way. However, extant simple laboratory paradigms for examining statistical learning may be insufficient to study decoding acquisition *in situ,* particularly if one hopes to scale up to pedagogically useful principles. First, as a whole, decoding is acquired slowly (over several years), so a single session may not elicit meaningful gains. Simplifying the learning goals for study in the lab, perhaps teaching only one or two GPC regularities, is also problematic: decoding is a system and a crucial skill is discriminating among multiple rules for a given letter string (e.g., MAT vs. MATE vs. MEAT). Second, the wide variation in children's initial abilities demands a large sample—we cannot work with an artificial language in which children have no initial knowledge. The scope of learning and the scale of the research may not be most efficiently conducted in one-at-a-time laboratory study. Finally and most importantly, a history of education research suggests that simple principles derived in optimal settings do not always yield gains when brought into a real classroom (e.g., Lundberg & Fox, 1991). Thus, we studied statistical learning processes in a real classroom using teaching tools that are similar to the instructional media children encounter in their schools. This may enable us to apply our findings to education more quickly.

We repurposed an existing computer based reading intervention, *Access Code* (Foundations in Learning Inc., 2010), to use as a platform for training. *Access Code* is part of the early reading curriculum in several school districts across the United States. This program uses a variety of multi-media tasks to teach children many GPC regularities (mostly vowels) over about 16 weeks. It precisely controls the stimuli, tasks, and measures, and its current usage (and similarity to many other computer-based interventions) makes this a somewhat ecologically-relevant basis for teaching for many students.

Using *Access Code* as a base, we taught first-grade students six GPC regularities. This kept training to about four days, while still emphasizing the contrast between multiple regularities. The computer-based intervention allowed us to control the quality of the stimuli, the structure and timing of the tasks and the delivery of feedback to create a fairly well-controlled laboratory-like learning task. However, as an existing classroom activity, this also mirrors what children are likely to encounter. This combined the laboratory precision of fine-grained control of learning parameters with a set of natural tasks that are already in use pedagogically.

Using this paradigm, we asked whether statistical regularities drive the acquisition of decoding, and whether variability or similarity in irrelevant units (the consonant frame for vowel GPC regularities) is more beneficial. First-grade children learned six GPC regularities for vowels over a few days. Children were not explicitly taught the regularities, but performed a series of tasks using words that embodied them. Half the children learned over a set of words with variable consonant frames; the other half learned over words with similar frames. After training, we gauged improvement on trained and untrained words, and on old and new tasks. The critical questions were whether variable or similar words led to better learning, how this generalized, and how different groups of students learned from variable and similar stimuli.

Previous work on statistical learning and reading does not offer clear predictions; however, work on variability in motor skill learning may (Del Rey, Whitehurst, & Wood, 1983; Magill & Hall, 1990). In this light, we expected variability to improve students' abilities to generalize GPC regularities to new tasks and new words, as students are exposed to regularities in a variety of contexts. However, we expected similarity to improve learning for specific training words and perhaps words that are highly similar to them. Variability may also be more effective in learning less consistent regularities, which require greater encoding of surrounding contexts. Although our stimuli all use dominant regularities (for that letter string), the less consistent monograph regularities may require greater variability (see Appendix C). Meanwhile, highly consistent regularities, such as the digraph regularities used here, may be better learned with highly similar words, as these stimuli offer less distracting information about the GPC regularities. Given previous findings of differential benefits of variability and similarity, it seemed unlikely that either one of these principles would dominate learning across all levels of the study. However, understanding the exact contributions of each to learning and generalization is vital for understanding how children acquire decoding skills, and how we may structure materials to improve this.

Methods

Students were randomly assigned to either the variable or similar group. Each student performed a pre-test, three to five days of training and a post-test. Post-test and pre-test used a set of words and tasks that partially overlapped with those used at training, to test generalization. However, the post-test was identical to the pre-test, including the same words and tasks, and identical between the two groups. During training, groups used different word-lists but were otherwise treated identically, receiving the same tasks.

Participants

Two-hundred sixty-four first-grade students (average age 7:0) from the West Des Moines, Iowa (USA), Community Schools participated in this study. Students were recruited from 15 classrooms in five elementary schools. In these schools, all first-grade students were invited to participate, except those students with individualized educational programs (IEPs: programs developed for students with specific disabilities including diagnosed learning, language and developmental disabilities). Parents of eligible students were first sent a letter detailing the study; subsequently consent forms were sent home with the students. Of the eligible students, approximately 75% participated, yielding a wide array of abilities. Two-hundred twenty-four students completed the entire study; 32 left the study after missing more than two sessions (typically due to illness); and eight left for other reasons. Of those completing the study, 119 were girls and 105 were boys; other demographic data is shown in Table 1.

Table 1. Demographic Breakdown of the two training groups

		Variable	Similar	Total
Gender	Male	49	56	105
	Female	62	57	119
First Language	English	103	102	205
	Other	8	11	19
SES	Free/Reduced Lunch	26	25	51
	Not eligible	75	81	156
	Unknown	10	7	17
Race	Caucasian	85	89	174
	African-American	5	4	9
	Asian	8	4	12
	Native American	0	1	1
	Multiple Races	5	4	9
	Unknown	1	1	2

Design

Children learned six GPC regularities: three short vowels (e.g., A as in BAT; I as in BIT; and O as in BOT) and three digraphs (e.g., AI as in BAIT; EA as in BEAT; and OA as in BOAT; Table 2). The first two days consisted of a pre-test with both types of vowels without feedback. Over the next 3-5 days, children were trained on three blocks of trials: short-vowels, digraphs and mixed. A subset of the words used in training was also present in the pre- and post-tests, while the majority of the training words were unique from the test words. Finally, children underwent post-testing, which was identical to the pre-test in both tasks and words.

Table 2. Vowels and digraphs used in the study.

Spelling	Pronunciation	Example Words
A	æ	FAT, PAD
O	ɑ	BOG, TOP
I	ɪ	RIM, SIT
AI	eɪ	BAIL, RAIN
EA	i	LEAP, MEAT
OA	oʊ	COAT, ROAD

Pre- and post-test did not differ between groups and consisted of two cycles of six tasks. Each cycle consisted of 48 trials (8 trials/task), with no error feedback. After pre-test, children were randomly assigned to a training group using pre-test scores to balance groups on initial performance (Table 3), and several demographic variables (Table 1).

Training consisted of six cycles of six tasks (6 tasks × 8 words/task × 6 cycles = 288 total trials), using either similar or variable words, depending on group. Tasks (Table 4) were based on tasks in *Access Code* (Foundations in Learning Inc., 2010) and emphasized different ways of using the GPC regularities. Feedback was given on each trial.

The primary manipulation was variability of the training words. Its effects on learning were assessed for words and tasks used in training, and for generalization to new words and tasks. To assess generalization across *tasks*, four tasks were used in both pre-/post-testing and training, and two were unique to testing (Table 4). For generalization across *words*, testing word-lists included four levels: trained words (e.g. PAT), close words (the same GPC regularities in similar consonant frames, e.g. FAT), far words (the same GPC regularities in more dissimilar frames, e.g. GAS), and alternative-rule words (untrained GPC regularities, e.g. PEN). Close, far and alternative-rule words were not used during training and were identical across groups.

Table 3. Pre-test scores and standard deviations for the two training groups by gender and native language

	Variable Words		**Similar Words**	
Group	**% Correct**	**SD**	**% Correct**	**SD**
Male	69.8	12.9	69.4	13.7
Female	70.9	14.7	71.2	12.8
Not ELL	70.9	13.4	70.7	12.9
ELL	63.7	18.4	66.5	15.8

Table 4. Descriptions of tasks

Task	**Description**	**Training**	**Testing**	**Stimuli**
Change the word (vowel)	See a consonant frame and eight vowel options. Asked aurally to change one word to another ("change the vowel in *cat* to make *coat*)	✓		Word
Change the word to non-word (initial)	See vowel and offset consonant and eight onset consonant options. Asked to change word to non-word ("change *meat* to make *geat*")	✓		Non-word
Find the word	Hear a word played and find that word from eight displayed alternatives	✓	✓	Word
Families	Hear the vowel and coda consonant of a word, and find the word that contains those sounds from eight alternatives	✓	✓	Word
Make the non-word	Hear a non-word and choose the letters to spell it from eight displayed alternatives for each position	✓	✓	Non-word
Fill in the blank	Hear a non-word and see a consonant frame and eight vowel options. Choose which vowel completes the played word	✓	✓	Non-word
Verify word/sound	Hear a word and see one printed on the screen. Determine whether they match		✓	Word

Change the non-word (final)	See onset consonant and vowel and eight offset consonants. Asked to change one non-word to another ("Change *geam* to make *geap*")	✓	Non-word

Word Lists

The training-groups received different word-lists emphasizing variability or similarity in the consonant frames (Appendix A). A common word-list was used for pre- and post-testing (Appendix B).

Training Words. Training words instantiated six GPC regularities, including three monograph vowels and three digraphs (Table 2). Each of these regularities is the dominant pronunciation for the given rule (see Table 5), although the monographs were somewhat less consistent. For each regularity, 5-6 words and 5-6 non-words were selected. Word-lists were reviewed by an expert in early reading to ensure the words were appropriate for and recognizable by first-grade children.

Table 5. Consistency of the GPC regularities trained. Consistency was quantified as the proportion of words in which that letter string was observed with the corresponding pronunciation. See Appendix C for details on these calculations and alternative conceptualizations of regularities

	Orthography	**Pronunciation**	**Example**	**Consistency**
Short Vowels	A	/æ/	HAT	52.1%
	I	/ɪ/	HIT	82.1%
	O	/ɑ/	HOT	52.3%
Digraphs	EA	/i/	HEAT	75.0%
	OA	/ou/	HOE	84.6%
	AI	/ɑɪ/	HIGH	76.0%

Our primary manipulation was the variability of the consonants surrounding the critical vowels (see Table 6). The ***similar word-list*** used items with overlapping consonant frames (e.g. COAT, CAT, and the non-word CAIT). Every item shared onset and coda consonants with at least one other item in the set (*M*=2.4 words shared entire frames; *M*=21.8 words shared a single consonant). The ***variable word-list*** minimized consonant overlap. No item in this list shared both consonants with more than one other word in the set (*M*=0.2 words with shared frames; *M*=10.4 words shared one consonant). The similar word-list only used 21 consonant frames (for 64 items), and for each frame there was an average of 3 items (words and non-words) instantiating different GPC regularities. In contrast, the variable list used 57 frames and each frame appeared in 1.1 items. As a result, the similar group saw many items which only differed in vowels (e.g., BAT, BIT, BAIT, BEAT), creating an ideal situation for contrast/comparison learning.

The selected words were balanced on other factors (Table 6). There was no difference between word-lists in log-frequency (Brysbaert & New, 2009) ($M_{similar}$=7.09, $M_{variable}$=7.49; *t*(62)=-.86, *p*=.391). Lists were also balanced for imageability (based on the MRC Psycholinguistic Database: Coltheart, 1981), for the words for which imageability was available

(26/32 similar words, 23/32 variable words; $M_{similar}$=499.0, $M_{variable}$=503.6; $t(47)=-.17$, $p=.87$). The words and non-words in the similar word-list had significantly more phonological neighbors (words made by a one-phoneme addition, subtraction or deletion, estimated with Vaden, Halpin, & Hickok, 2009) than those in the variable list ($M_{similar}$=36.1, $M_{variable}$=30.1; $F(1,124)=18.6$, $p<.001$). This is unsurprising, as the words in the similar word-list were selected based on the existence of other words with similar frames. Words also had more neighbors than non-words (M_{words}=37.8, $M_{non\text{-}words}$=28.5; $F(1,124)=44.2$, $p<.001$; although this did not interact with list, $F(1,124)=2.1$, $p=.15$), suggesting that caution is necessary when comparing learning of words and non-words. Finally, we computed positional probability for each word as the product of the probability of each letter appearing in its position. The word-lists did not differ in positional probability ($M_{similar}$=.056, $M_{variable}$=.054; $F(1,124)=1.04$, $p=.31$). While words had higher positional probabilities than the non-words (M_{words}=.059, $M_{non\text{-}words}$=.051; $F(1,124)=16.02$, $p<.001$), this did not interact with word-list ($F(1,124)=1.40$, $p=.24$),

Table 6. Properties of the items assigned to similar and variable word-lists.

	Property	Similar	Variable	Sig.
Experimentally Manipulated	*Number of other items sharing both consonants*	2.4	.2	<.0001
	Number of other items sharing one consonant	21.8	10.4	<.0001
	Number unique consonant frames	21	57	N/A[1]
	Items / consonant frame	3.04	1.1	N/A
Other factors	*Log Frequency (words only)*	7.09	7.49	.39
	Imageability (words only)	499	503.6	.87
	Lexical Neighbors[2]	36.1	30.1	<.001
	Positional Probability[3]	.056	.054	.31

[1] No statistics could be computed for count values
[2] When non-words have available values, they are included in the means reported here
[3] Positional probability is the product of the individual probabilities of each letter in each position and weighted by the log-frequency weighting of items

While there were no overall differences between lists on the positional probability of the letters, the variable list qualitatively appeared to include more difficult letters. We were not able to fully balance the particular letters used in each word-list. In order to find sufficient similar words that first-graders were likely to know, we over-sampled common letters; and to construct the variable list we tried to maximize the number of consonants and frames across words, forcing us to use many lower-frequency letters. This predicts an advantage for the children learning similar words and will be addressed in the discussion.

Two words and two non-words from each GPC regularity were shared across word-lists. These words had low similarity scores in the similar word-list. These allowed us to test learning using identical stimuli between groups and were the only training items used at test.

Testing Stimuli. The word-list for testing was the same for both training-groups. For each

GPC regularity, eight words and eight non-words were selected: two were the shared items from the training word-lists; two were minimal pairs with these shared items (close words) to measure generalization to similar-sounding words; two were not closely related to either word-list by the metrics described above (far words) but employed the same GPC regularities to measure more distal generalization; and two were words using different vowels (alternative-rule words; three words each using E and U for monographs, two each using OO, OU and EE for digraphs) to differentiate task-specific learning from GPC learning.

Procedures

For each student, the experiment lasted approximately 1.5 weeks. One or two schools were run at a time, and the entire study was run from January to March, 2011.

Students participated during the school day. They were removed from class in small groups whose size depended on the school and the availability of computers. Each student used a unique login to track his/her progress. During pre- and post-testing students completed one cycle (48 trials) each day for two days[1]. During training, the students worked for 20 minutes per day, completing as many tasks as possible, after which they were logged out (after completing the current task). On the next day they began where they had left off. If students completed a full cycle within 20 minutes, they proceeded to the next cycle. Students took as many days as needed to complete all training cycles, usually between three and five.

Upon entering a cycle of tasks, students saw icons representing the six tasks in that cycle and they selected the order to complete them. After completing the eight trials in the task, students returned to the task selection screen, where a checkmark signified which were complete. Each cycle was presented in a new color to reinforce students' advancement.

Feedback. During training, to keep students motivated and promote learning, feedback and scaffolding were given on each trial, and a score accumulated across trials. Students had two attempts to select the correct answer from a small number of alternatives. If they responded incorrectly on their first attempt, a buzz sounded, the incorrect response was removed and they tried again. After two incorrect responses, the correct answer was revealed and no points were awarded. After a correct response (on either attempt), a ding sounded, and points were awarded. Within a task, the score was displayed at the bottom of the screen, and the task selection screen showed the score for each task and the full point total.

During pre- and post-test, students had only one chance to respond and received no feedback or points. Students received neutral reinforcement (e.g., "Thanks for working so hard") approximately every fourth trial to keep them engaged.

Tasks. Eight tasks were used across the experiment. Four were used in both testing and training, two exclusively in training and two exclusively in testing. Half of the tasks used real words, and half used non-words. Each task was run for eight trials. Each task included detailed auditory instructions before the trials began. Each trial presented shorter spoken instructions, accompanied by a target stimulus and a small number of responses. For example, in *Fill in the Blank*, children heard "Make the non-word GAT," accompanied by G_T on the screen. They could then chose from eight response options (selected from among A, E, I, O, U, AI, EA, EE, OA, OO and OU). Children could repeat the auditory stimulus for the trial or the task

[1] Most students completed the pre-test on Thursday and Friday and then were assigned to their training group over the weekend. A few students were absent during one day of pre-testing and finished the pre-test on Monday.

instructions at any time by clicking on icons on the screen. For a summary of all tasks, see Table 4. Auditory materials (stimuli, carrier phrases and instructions) were recorded by a phonetically-trained female, talking at a slow, clear rate of speech in a sound-proof room using a Kay Elemetrics CSL 4300b at 44,100 Hz sampling rate. Stimuli were presented over headphones to minimize disruption from other students undergoing training/testing at the same time.

Results

Data were analyzed with logistic mixed effects models using the LME4 package (Bates & Sarkar, 2011) of R (version 2.13.1). Each model considered every trial individually, using a binary dependent variable (1=correct) indicating accuracy on each trial. The primary factors of interest were test (pre-/post-, within-participant) and training-group (similar/variable, between-participant). We were also interested in how training-group affected generalization to different word-types (trained, close, far and alternative-rule) and task-types (trained vs. test-only), whether tasks including words differed from those including non-words, whether results differed by gender, how students at different levels of initial performance benefited from variability or similarity[2]. A single model examining all factors would entail many fixed effects and numerous interactions, leaving too few trials in each cell for the model to converge. Thus, we instead ran a series of models with group and test as factors, plus one additional factor.

All models included participant and word as random intercepts, as these improved model fit over using participants alone (all p<.001 using χ^2 test of model fit), while further adding school did not improve fit (all p>.05). Correlation among fixed effects (R_{max}) did not exceed .017 in all models. Only pre- and post-test data were analyzed, as the training-groups used different word-lists during training. We were concerned that the ELL students may have responded differently, so analyses were run both with and without these students. As there were no differences in the patterns significance, we report the analyses with all students.

There is no widely agreed upon measure of effect size like R^2 of Cohen's *D* for logistic models, and estimating such factors is even more difficult in mixed-designs such as this one. We report log odds ratios (*LOR*) as an estimate of how much more likely the data derived from a model with a specific factor and without it. These were estimated as the difference in log-likelihood for models with and without each factor.

Effect of Variability and Generalization across Tasks. We first asked if training-group affected learning (the test x group interaction) and whether its effects were moderated by familiarity with the tasks (the three-way interaction with task-type). There was a significant effect of test (*B*=.24, *SE*=.03, *LOR*=61, *Z*=8.1, *p*<.0001), indicating significant learning between pre- and post-test. There was no effect of training-group (*B*=.12, *SE*=.14, *LOR*=12, *Z*=.8, *p*=.41). However, the test x training-group interaction was highly significant (*B*=.29, *SE*=.06, *LOR*=12, *Z*=4.8, *p*<.0001), as improvement for the variable group (*M*=5.2%, *SD*=12.9%) exceeded the similar group (*M*=1.9%, *SD*=14.9%; Figure 1A).

There was a significant effect of task-type (*B*=1.23, *SE*=.24, *LOR*=14, *Z*=5.1, *p*<.0001); participants performed *worse* on repeated tasks (*M*=65.6%, SD=18.0%) than test-only tasks (*M*=84.6%, *SD*=13.4%). This difference appeared at both pre- and post-tests, so it was likely due

[2] While we wanted to examine native language and SES, our final sample included too few students in ELL and low SES group for this analysis to be feasible.

to differences in the overall difficulty of the tasks we chose to be repeated or test-only. Task-type did not interact with test (*B*=-.08, *SE*=.06, *LOR*=1, *Z*=-1.4, *p*=.17) or training-group (*B*=.08, SE=.06, *LOR*=1, *Z*=1.29, *p*=.20) nor was the three-way interaction significant (*B*=.15, *SE*=.12, *LOR*=0, *Z*=1.27, *p*=.21). These results indicate that the variable group consistently outgained the similar group across tasks (Figure 1B).

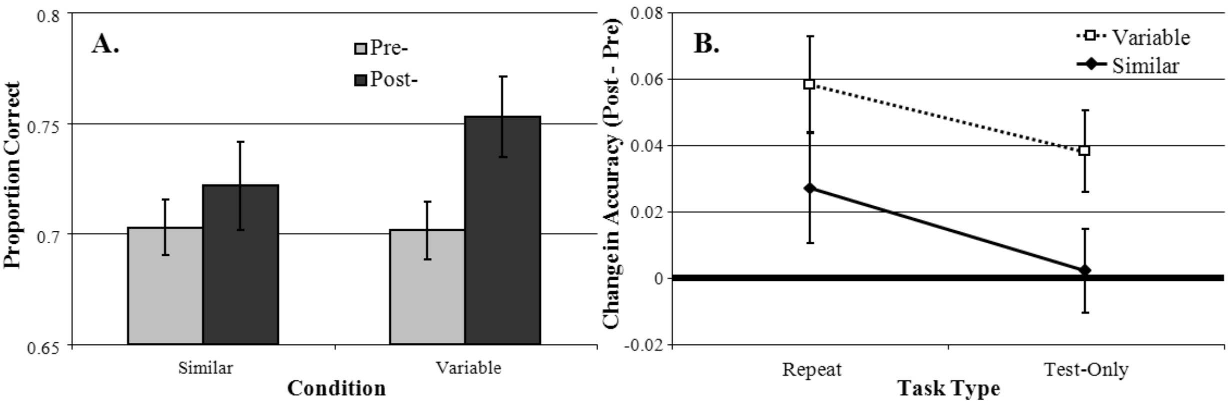

Figure 1. A) Performance as a function of test (pre-/post-) and condition. B) Learning (change in accuracy from pre- to post-test) as a function of task-type and condition. Note, in this and future figures, we use bar graphs whenever we report raw accuracy, and line graphs to report change scores. Error bars in both reflect standard error of the mean.

In fact, the similar group showed little improvement in the test-only tasks. Planned comparisons were conducted by examining a subset of the data (each of the four task-type by group cells) using a similar model. This revealed significant learning for both types of tasks in the variable group (Repeat: *B*=.40, *SE*=.04, *LOR*=42, *Z*=9.10, *p*<.0001; Unique: *B*=.38, *SE*=.07, *LOR*=13, *Z*=5.0, *p*<.0001); however, the similar group improved in repeated tasks (*B*=.18, *SE*=.04, *LOR*=9, *Z*=4.26, *p*<.0001) but not for test-only tasks (*B*=.02, *SE*=.07, *LOR*=0, *Z*=.28, *p*=.78). Thus, while variability enhances learning overall, it was essential for generalization to new tasks.

Generalization across Words. We next asked if the benefit of variability generalized across the four types of words and non-words: trained, close, far and alternative-rule. Word type was treated as a linear factor[3].

Again, there was a significant effect of test (*B*=.26, *SE*=.026, *LOR*=82, *Z*=9.95, *p*<.0001), no effect of training-group (*B*=.09, *SE*=.14, *LOR*=21, *Z*=.69, *p*=.49) and a test x training-group interaction (*B*=.24, *SE*=.05, *LOR*=12, *Z*=4.75, *p*<.0001). There was no effect of word-type (*B*=.001, *SE*=.17, *LOR*=22, *Z*=.007, *p*=.99); however, there was a word-type x test interaction (*B*=.23, *SE*=.035, *LOR*=23, *Z*=6.72, *p*<.0001), as strong learning for the trained and close words was balanced by higher initial performance for far and alternative-rule words (Figure 2A). Most

[3] We considered two ways to represent word-type: as a linear trend (a generalization gradient) or a factor with four levels. For the linear trend, we set trained words to 1, close words to 1/3, far words to -1/3 and alternative-rule words to -1. For the factor, we used three centered dummy codes to represent the levels. The dummy coded model did not offer any benefit over the linear model ($\chi^2(8)$=5.1, *p*=.74; $\text{BIC}_{\text{dummy}}$=37585, $\text{BIC}_{\text{linear}}$=37505).

importantly, the training-group x word-type interaction was not significant (*B*=.017, *SE*=.035, *LOR*=1, *Z*=.51, *p*=.61, Figure 2B), nor was the three-way interaction (*B*=.07, *SE*=.07, *LOR*=1, *Z*=1.0, *p*=.30). Planned comparisons showed significant learning in all conditions except the alternative-rule word-types in both training-groups (Table 7). Thus, the benefits of variability extend to untrained words using the same phonics rules. While students performed differently across word-types, the variable group exhibited greater performance than the similar group; variability appears to offer an across-the-board learning benefit over similarity, even for words dissimilar to those trained.

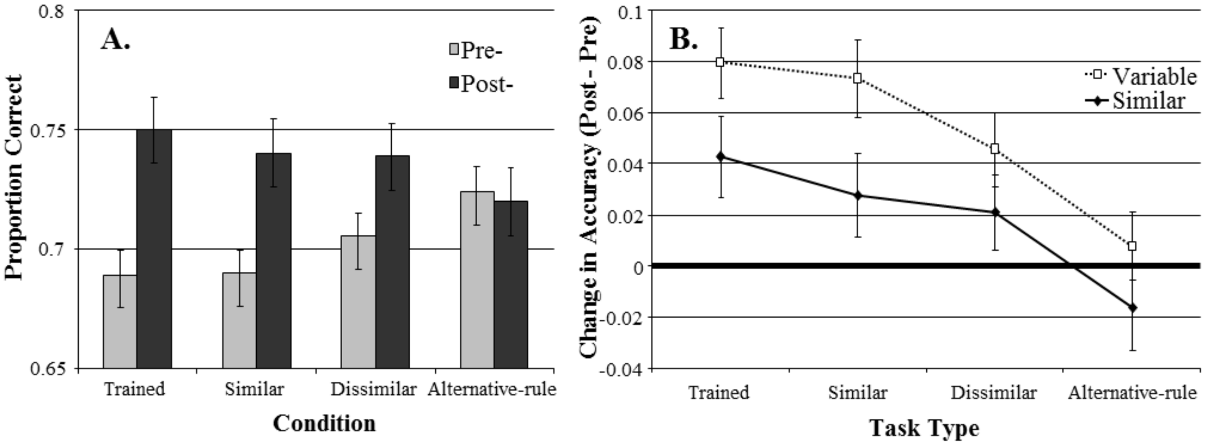

Figure 2. Effect of word-type. A) Learning across conditions as a function of word-type. B) The amount of learning as a function of condition and word-type. Error bars indicate standard error of the mean.

Table 7. Results of mixed models examining the effect of test (pre-/post-) in individual training-group x word-type cells

Group	Word-Type	B	SE	Z	p
Similar	Trained	.31	.07	4.26	<.0001
	Similar	.19	.07	2.71	.0067
	Dissimilar	.15	.07	2.12	.034
	Alternative-rule	-.12	.07	1.64	.101
Variable	Trained	.59	.08	7.90	<.0001
	Similar	.55	.08	7.31	<.0001
	Dissimilar	.35	.08	4.68	<.0001
	Alternative-rule	.05	.07	0.76	.446

Generalization across GPC Regularities. We next asked if the effect of variability differed between short-vowels and digraphs. Differences in learning found herein could result from the complexity of the regularities (digraphs involved multiple letters) or to differences in consistency of the regularities (the digraph regularities were more consistent, as they are more likely to have only one pronunciation).

As in the prior analyses, this analysis found a significant effect of test (*B*=.25, *SE*=.02, *LOR*=139, *Z*=9.72, *p*<.0001), no effect of training-group (*B*=.09, *SE*=.14, *LOR*=11, *Z*=.67, *p*=.50), and a test × training-group interaction (*B*=.24, *SE*=.05, *LOR*=11, *Z*=4.71, *p*<.0001). There was also a main effect of GPC regularity (*B*=-.54, *SE*=.25, *LOR*=81, *Z*=2.16, *p*=.030): digraphs had a lower accuracy (*M*=67.2%, *SD*=16.4%) than short vowels (*M*=76.7%, *SD*=15.7%). There was also a test × regularity interaction (*B*=.67, *SE*=.05, *LOR*=79, *Z*=12.62, *p*<.0001): The short vowels showed little improvement (across both similar and variable groups), while digraphs showed substantial gains (Figure 3A). This did not differ by group: the training-group × rule interaction was non-significant (*B*=.03, *SE*=.05, *LOR*=0, *Z*=.66, *p*=.51) as was the three-way interaction (*B*=.04, *SE*=.11, *LOR*=0, *Z*=.41, *p*=.68).

However, whether or not learning was observed depended on the group and the rule (Figure 3). For digraphs, there was a highly significant interaction between test and training-group (*B*=.27, *SE*=.07, *LOR*=7, *Z*=3.67, *p*=.0002). As sub-analyses showed significant learning in both training-groups (Similar: *B*=.46, *SE*=.05, *LOR*=40, *Z*=8.87, *p*<.0001; Variable: *B*=.75, *SE*=.54, *LOR*=96, *Z*=13.79, *p*<.0001), the larger training effect in the variability group was responsible for this interaction. For short vowels, there was also an interaction (*B*=.22, *SE*=.07, *LOR*=4, *Z*=2.98, *p*=.0028), but this was driven by a significant *decrement* in the similar group (*B*=-.18, *SE*=.05, *LOR*=6, *Z*=3.58, *p*=.00034), and no gains by the variable group (*B*=.04, *SE*=.05, *LOR*=0, *Z*=.68, *p*=.50). This small decline may reflect some form of catastrophic interference. That is, as many children knew short-vowels at the onset of the study (but had not been exposed to digraphs), the digraph training may have interfered with their earlier learning, an effect that was moderated by variability.

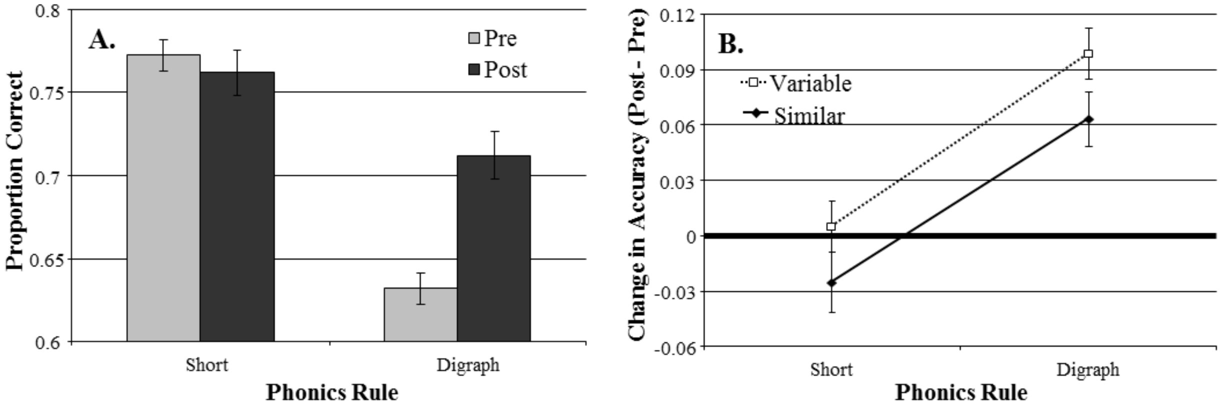

Figure 3. Learning as a function of GPC regularity. A) Pre- and Post-test performance averaged across groups as a function of GPC regularity. B) Change in performance for each group and regularity. Error bars indicate standard error of the mean.

Learning as a Function of Gender. We analyzed student performance to determine whether the effects of similarity/variability are similar across both genders. Our analysis of gender included test, training-group and gender (contrast coded).

As before, we found a significant main effect of test (*B*=.38, SE=.04, *LOR*=73, *Z*=10.54, *p*<.0001), no effect of training-group (*B*=.09, *SE*=.19, *LOR*=12, *Z*=.48, *p*=.62), and a test × training-group interaction (*B*=.32, *SE*=.07, *LOR*=12, *Z*=4.47, *p*<.0001). The overall main effect of gender was not significant (*B*=-.21, *SE*=.14, *LOR*=15, *Z*=-1.50, *p*=.13). However, the gender × test interaction was significant (*B*=-.26, *SE*=.05, *LOR*=14, *Z*=-4.96, *p*<.0001), as females showed greater learning than males (Figure 4A). Moreover, while the training-group × gender interaction was not significant (*B*=-.009, *SE*=.28, *LOR*=2, *Z*=.034, *p*=.97), the three-way interaction was marginally significant (*B*=-.17, *SE*=.11, *LOR*=2, *Z*=1.71, *p*=.087; Figure 4B).

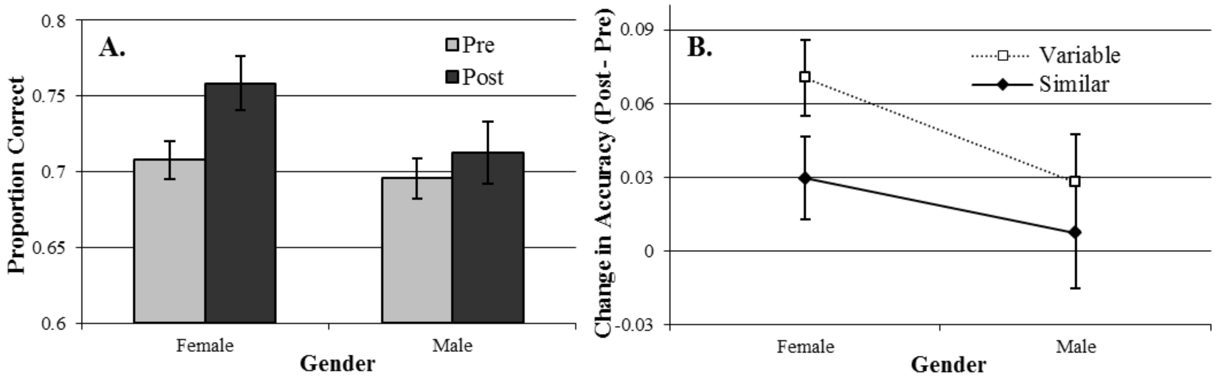

Figure 4. A) Performance as a function of test (pre-/post-) and gender. B) Learning as a function of gender and training group. Error bars indicate standard error of the mean.

We further examined learning in each of the four gender × training-group cells. Girls showed significant learning in both training-groups (Similar: *B*=.23, *SE*=.05, *LOR*=10, *Z*=4.36, *p*<.0001; Variable: *B*=.55, *SE*=.05, *LOR*=58, *Z*=10.68, *p*<.0001), although learning in the variable group was greater (training-group x test: *B*=.33, *SE*=.07, *LOR*=10, *Z*=4.50, *p*<.0001). However, boys only showed significant learning in the variable group (*B*=.20, *SE*=.05, *LOR*=6, *Z*=3.65, *p*=.0003), not in the similar group (*B*=.05, *SE*=.05, *LOR*=0, *Z*=.99, *p*=.32). For boys, who overall learned less than girls, variability may have been particularly essential for learning.

Word vs. Non-word Tasks. The next set of analyses examined whether effects differed for words and non-words. With the present design, we cannot make strong conclusions about this as it was confounded with task (words were tested on different tasks than non-words) and neighborhood density (words had more neighbors than non-words). Nonetheless, this analysis was done for two reasons. First, tasks involving non-words were more difficult than those involving words (at pre-test, $M_{\text{non-words}}$=58.8%, *SD*=13.7%; M_{words}= 81.6%, *SD*=15.7). Thus, this comparison offers another opportunity to examine items or situations in which learning or performance may have been more difficult. Second, in the connectionist Triangle model of reading, (Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989), non-words uniquely tap mappings between orthography and phonology, whereas known words can be recognized via this pathway (e.g., sounding it out) or via directly mapping orthography to meaning (or a combination). Thus, learning could differ depending on which pathways are available, and the subset of non-word items may uniquely tap a single pathway. While we cannot make strong conclusions as to whether variability influences word or non-word processing differently, determining where learning is strongest, and whether some tasks or types of words do not elicit learning can further our understanding of variability effects, particularly in the context of connectionist/statistical learning models.

This model was the same logistic mixed effects model as in the primary analyses and included word/non-word (contrast coded) as a fixed effect along with training-group and test.

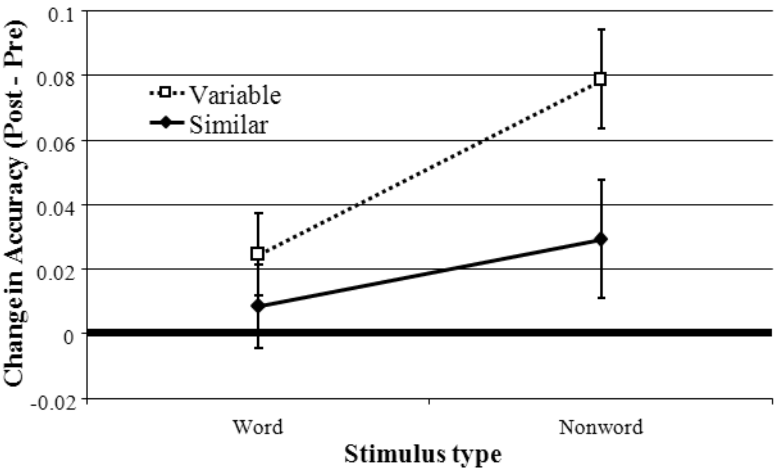

Figure 5. Learning as a function of word-type and training group. Error bars indicate standard error of the mean.

As in previous analyses, there was a main effect of test (*B*=.24, *SE*=.03, *LOR*=67, *Z*=9.25, *p*<.0001), no effect of training-group (*B*=.10, *SE*=.14, *LOR*=15, *Z*=.78, *p*=.44), and a test × training-group interaction (*B*=.23, *SE*=.05, *LOR*=13, *Z*=4.43, *p*<.0001), indicating better learning in the variable group. The effect of word/non-word was highly significant (*B*=-1.31, *SE*=.21, *LOR*=26, *Z*=5.99, *p*<.0001) with participants responding more accurately to words (*M*=82.5%, *SD*=16.5%) than non-words (*M*=61.4%, *SD*=16.4%). The training-group × word/non-word interaction was significant (*B*=-.13, *SE*=.05, *LOR*=4, *Z*=2.4, *p*=.015): averaging across both pre- and post-test, the variable group performed slightly better overall than the similar group on words, whereas the groups performed the same with non-words. There was also a test × word/non-word interaction (*B*=.20, *SE*=.05, *LOR*=7, *Z*=3.69, *p*=.0002) with less learning across both groups for words (M_Δ=1.6%, *SD*=13.4%) than non-words (M_Δ=5.3%, *SD*=17.8%). Finally, the three-way interaction was not significant (*B*=.16, *SE*=.10, *LOR*=1, *Z*=1.47, *p*=.14), indicating that the variable group showed greater learning than the similar group for both word and non-word stimuli, and the amount of improvement did not differ between the two types of stimuli.

While both groups showed some learning, the similar group seemed to show little improvement for words (Figure 5). Thus, follow-up tests were conducted to determine if there was significant learning in each cell. These revealed significant learning for both types of stimuli in the variable group (Words: *B*=.23, *SE*=.06, *LOR*=8, *Z*=3.96, *p*< .0001; Non-words: *B*=.49, *SE*=.05, *LOR*=51, *Z*=10.11, *p*<.0001). However, the similar group showed evidence of learning only for non-words (*B*=.18, *SE*=.05, *LOR*=7, *Z*=3.84, *p*=.0002) but not for words (*B*=.07, *SE*=.06, *LOR*=1, *Z*=1.29, *p*=.197). When performing tasks with real words (or performing these specific tasks), learners appear to require variability. In a triangle-model formulation, this may suggest that variability helps focus children on the orthography→phonology mappings, avoiding the lexical route (which is less relevant for specifically learning decoding skills).

Learning as a Function of Initial Ability. Finally, we asked whether variability had a differential effect for learners with different starting levels. We did not have an independent assessment of students' decoding skills, as standardized testing is not performed on this age group in West Des Moines. Thus, a median split of their pre-test was used (Median=79.16%). This prevented us from using performance-group as a factor (since it would be correlated with pre-test). Thus, we examined test and training-group in the low-performer and high-performer groups separately.

Both high- and low-initial-performers showed a main effect of test (Low: *B*=.18, *SE*=.03, *LOR*=18, *Z*=5.22, *p*<.001; High: *B*=.37, *SE*=.04, *LOR*=48, *Z*=9.22, *p*<.0001), no main effect of training-group (Low: *B*=.08, *SE*=.16, *LOR*=5, *Z*=.49, *p*=.62; High: *B*=.13, *SE*=.12, *LOR*=7, *Z*=1.02, *p*=.30), and a test × training-group interaction (Low: *B*=.21, *SE*=.07, *LOR*=5, *Z*=3.09, *p*=.0020; High: *B*=.30, *SE*=.08, *LOR*=7, *Z*=3.74, *p*=.00018; Figure 6). Thus, both low and high performers learned more from variable words. Simple effects analyses showed that variability led to significant learning for both initial-performance levels in (Low: *B*=.29, *SE*=.05, *LOR*=17, *Z*=5.87, *p*<.0001; High: *B*=.53, *SE*=.06, *LOR*=41, *Z*=9.01, *p*<.0001), but that similarity only led to significant learning for high-performers (Low: *B*=.07, *SE*=.05, *LOR*=1, *Z*=1.51, *p*=.13; High: *B*=.21, *SE*=.05, *LOR*=8, *Z*=3.96, *p*<.0001).

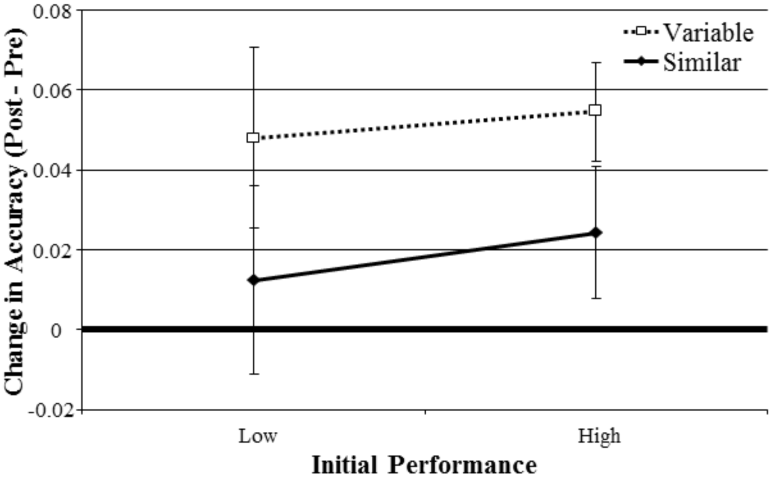

Figure 6. Change in performance as a function of initial performance level and condition. Error bars indicate standard error of the mean.

Discussion

Building on the literature implicating statistical learning in the acquisition of decoding, we asked how similarity among the words encountered during learning affects children's ability to acquire and generalize GPC regularities. We trained students on several GPC regularities using a modified classroom-implemented reading software package. Students learned these regularities via training on either highly similar words or more variable items. We then gauged how much students improved after several days of training using their given word-list.

Our word lists were designed to maximize the kind of learning proposed by both similarity and variability accounts. Each consonant frame in the similar group was used with an average of 3.04 different GPC regularities (pronunciations), compared to only 1.1 in the variable group. This should have helped children focus on the highly discriminative vowels, and set up a situation supportive of direct comparisons/contrasts across words (a hypothesized benefit of similarity). However, in the similar lists, the same consonant appears with many vowels and many pronunciations – if children do not know to ignore the consonants during learning, these could be incorrectly associated with the pronunciation, degrading performance. In contrast, the variability word-list used over twice as many consonant frames (57 vs. 21) and many more total consonants (19 vs. 10), which may minimize the formation of associations between consonants and vowel pronunciations by implicitly focusing participants' attention and/or associative linkages on the vowels.

Despite theoretical support for both similarity and variability under our task conditions, results were unequivocal. Children exposed to items with greater variability in consonant frames learned vowel GPC regularities better than children exposed to items with similar frames. This benefit extended to novel tasks and words, showing that variability significantly improved generalization. It also held for both students who entered the study with greater reading abilities and those who began at a lower level, showing that variability can help both early acquisition of regularities and later refinement of learned categories. This improvement was present for simpler GPC regularities involving monograph short vowels and for more complex regularities involving digraph long vowels, and it occurred for both girls and boys, and in both the easier tasks using real words and the harder tasks using non-words. Variability in irrelevant elements consistently facilitated GPC learning and transfer, and we did not see a benefit for similarity under any combination of conditions.

Not only did stimulus variability increase learning, but in some cases it was essential for learning to occur at all: similar words produced little measurable benefit for low-initial-performers, while both performance-levels exhibited learning when trained with variable stimuli. Girls learned in both conditions, (though more so with variability), whereas boys only learned with variable stimuli. This accords with past research showing that boys are delayed in learning to read relative to girls (e.g. Wolf & Gow, 1986). Indeed, we found that boys in our study began at a lower performance level (though this difference was not significant); as with the analysis of initial-low-performers, this performance level may have necessitated variable training for learning to occur. Generalizing learning to novel tasks required variable training stimuli, although some generalization across words could be seen in both groups. There were also indications that when children were performing very well, variability (but not similarity) was necessary to improve performance. For example, in the short-vowels, children in the similar group appeared to lose ground, while this effect was arrested for students receiving variable words. This may occur because first-grade students in both groups already knew these short-vowel regularities before the study, and training with similar words led them to modify an effective strategy already in place, (i.e. catastrophic interference). Similarly, we found that tasks employing real words (which also showed better performance overall) only showed learning with variable stimuli, while non-word tasks showed learning in both cases. This may be because with real words students had multiple routes to the correct answer, which slowed learning, while non-words permitted only the decoding route, making for a clearer learning situation, and one more likely to benefit from variability, which emphasizes the invariant mappings.

Our results contrast with our predictions that similarity and variability would each be helpful in certain circumstances. Instead, we found that variability consistently yielded better learning than similarity, suggesting that all the GPC regularities and generalization conditions in this study benefit from a similar learning mechanism. Also, while variability showed the predicted improvement for generalization to new words and tasks, it also improved word- and task-specific learning. Perhaps most importantly, despite more training on specific consonant frames for students receiving the similar word-list, these students did not show greater learning for these words.

Limitations. While these findings implicate a crucial role for variability in learning to read, there are a number of limitations that should be addressed. First, unavoidably, the similar word list employed more frequent (and likely familiar) letters than the variable list. This should have enhanced learning, but variability nonetheless led to greater gains, suggesting that letter frequency was a small component of performance.

Second, and more importantly, this study used only a small set of vowel GPC regularities. While these regularities spanned different degrees of consistency, all involved dominant GPC mappings. It remains to be seen how variability would affect less consistent regularities or subregularities within groups of exceptions (e.g. exceptions like LEAD, HEAD, DEAD, for the dominant EA→/i/ regularity). In learning more complex regularities, similarity may prove more helpful, helping to elucidate the specific contexts in which a regularity holds, or pointing learners toward fine-grained levels of analysis.

Third, it is unclear how variability and similarity may play out in languages with a more transparent orthography, like German or Greek (Protopapas & Vlahou, 2009), wherein GPC regularities more closely resemble rules. Children learning to read these languages may benefit from other classes of statistics. It is also possible (though we feel unlikely) that they may engage entirely different, rule-based learning mechanisms that are insensitive to statistical patterns in orthographic forms of their language. It is equally possible, however, that variability can play an even greater role here by highlighting the much stronger invariances in this language. Examining the role of variability in such languages can offer a more thorough picture of the role of statistical learning in learning to read more generally by allowing us to test proposed statistical mechanism against the backdrop of orthographic systems with different statistical properties.

Finally, it remains to be seen how our laboratory manipulation extends to other forms of learning, such as classroom instruction, extension activities like worksheets, and children's own exposure to text when they read on their own. Although one could simply manipulate the lists of items that are embedded in these materials to maximize variability, other properties of the learning system are quite different. For example, feedback may be delayed or non-existent. Here the extant literature may offer some insight. Variability is often implicated in unsupervised forms of learning (Gómez, 2002; Rost & McMurray, 2009, 2010), whereas similarity-based learning seems to thrive more often in error-driven situations (Schyns & Rodet, 1997). However, our task showed variability benefits despite employing error-driven training, suggesting that the variability effect should be robust in other forms of instruction.

Bringing Variability to the Classroom. The findings from this study have important implications for designing reading curricula. Tasks like word families are popular in pedagogy. These tasks emphasize the similarity between different words employing the same regularity by showing, for example, several words that end in –AT followed by several that end in –OB, and so forth. However, our work suggests that this is not the right framing and that these tasks may be less effective than tasks which highlight the range of different contexts in which regularities hold. Rather than showing several words ending in -AT that produce the /æ/ sound, children may benefit more greatly from learning that emphasizes how a wide array of consonant frames elicit this pronunciation. At a broader level of application, we see that these principles are potentially wide-ranging and can easily be applied to early readers, worksheets and other activities.

At a more fundamental level, the results from this study further the literature showing the statistical nature of learning to read (e.g., Arciuli & Monaghan, 2009; Arciuli, et al., 2010; Arciuli & Simpson, 2012; Treiman & Kessler, 2006; Treiman, et al., 2003), by identifying more precisely the class of statistics that children may harness in the service of learning at least some GPC regularities. To learn to read, children must encode probabilistic links between orthographic and phonological or semantic forms. While children become exposed to a wide variety of words as they read more, the distribution of words used during early learning may determine how quickly children learn GPC regularities. If children are faced with primarily overlapping words, from tasks like word families or beginner's books emphasizing rhymes, this may not benefit their acquisition of crucial reading skills to the same extent as a much more variable word list.

At a finer level, this work also reveals something important about the nature of GPC learning. As we've argued, variability and similarity have each been used to frame the learning problem in different ways. Similarity appears to help when the goal is classification (e.g., discriminating the words with an /æ/ sound from those with an /i/ sound); while variability helps when the goal is obtaining more invariant mappings. Our work suggests that during the acquisition of these GPC regularities, children need to detect invariant mappings amidst a number of irrelevant elements; variability seems to help children identify the criterial aspects of orthography to master GPC regularities. Variability has often been invoked in situations in which children do not appear to know what elements of the stimulus contain relevant statistics and which do not (Gómez, 2002; Rost & McMurray, 2009, 2010). Variability in irrelevant elements helps children avoid making spurious associations with variable elements (since they never appear frequently enough to form associations) and can thus help children identify the right statistics. This suggests that learning to read may also involve some aspect of dimensional attention (either explicitly, as in attentional accounts, or implicitly via associations) in which children must learn to pay attention to specific classes of letters/sounds for particular purposes. While we have illustrated how variability can help children narrow in on the right dimensions, other exercises and training tasks may serve this same purpose. The present study also introduces a platform with which to investigate how best to structure reading education to help children quickly and effectively learn GPC regularities. By studying statistical learning within a classroom setting, findings can easily be incorporated into existing curricula.

Implications for (and from) the Literature on Skill Learning. At a broader theoretical level, our findings fit nicely into current research on variability. Variability benefits are seen in numerous domains, including phonological development (Rost & McMurray, 2009, 2010), lexical dependencies (Gómez, 2002), L2 acquisition (Lively, et al., 1993; Perrachione, et al., 2011), motor skills (Kerr & Booth, 1978), and even learning to land planes (Huet, et al., 2011). Numerous mechanisms have been proposed for these effects, and such mechanisms may be relevant to reading. Broadly, our work shows variability in irrelevant elements can enhance learning when the regularities lie within a high-dimensional mapping. Under associative accounts, this can be explained if learners do not know which elements (types of letters) should be associated with responses (phonology). As a result, with limited variability, learners partially associate all letters in a word with the sound. For example, if all training words for the short vowel /æ/ began with the consonant T, students may falsely learn that the letter T was essential to predicting the sound /æ/. Variability may help students learn which elements are relevant by blocking the formation of spurious associations with non-criterial elements (c.f., Apfelbaum & McMurray, 2011). Our data are also consistent with dimensional attention or Bayesian accounts, that suggest people use variability to weight whole dimensions, with variable dimensions receiving less weight (Ernst & Banks, 2002; Toscano & McMurray, 2010). This suggests an attentional process by which children treat consonants and vowels differently in the context of particular GPC rules. It is still, however, an open question as to whether the idea of dimensional attention applies to reading.

Given the wealth of prior work in other domains and the strong theoretical accounts for it, a benefit of variability in reading may seem unsurprising. However, other work suggests a cost to variability, and we examined (within our context) many situations in which this could have appeared. Studies with learners of a second language (Perrachione, et al., 2011) and in motor skill learning (Wulf & Shea, 2002) suggest that variability can be detrimental for complex skills, or for learners with poor initial abilities. However, we found *better* learning with variability for the more complex digraphs, the more difficult non-words, and even for students who were initially low-performing. Similarly, the items in the similar group were designed to promote comparison and force children to make fine-grained distinctions as has been shown in prior work to promote learning (Namy & Gentner, 2002; Schyns & Rodet, 1997), yet there did not appear to be any benefit to this. Future work must examine why similarity mechanisms appear operative in some problems or domains, while variability is more valued in others. It may be helpful to develop tasks that incorporate similar statistics/mappings in non-reading domains to isolate the effect of domain and/or background knowledge (c.f., Wifall, McMurray, & Hazeltine, submitted)

Despite the consistent variability benefits here, it is not obvious that every form of variability will be beneficial for reading. It is important to consider why variability sometimes impedes learning. For example, similarity and variability along category-relevant dimensions play a complex role in the formation of cohesive categories, depending on whether the variability is observed within- or between categories (Palmeri, 1997), or in relevant or irrelevant dimensions (Rost & McMurray, 2010). Here we only examined variability in category-irrelevant dimensions. Beyond stimulus variability, a number of other dimensions may be important. Trial-by-trial variability (e.g., whether items are blocked or mixed during training) may matter, as work in L2 speech perception suggests different learners may benefit from mixed or blocked training (Perrachione, et al., 2011). Similarly, in motor skill learning, variation among tasks may promote generalization in some tasks (Wulf & Shea, 2002). Developing a more nuanced view of the interplay of task and stimulus variability will facilitate the application of learning phenomena to the classroom, and clearly work is needed in reading investigating these factors. The paradigm introduced here provides a means to examine the roles of various forms of variability in a domain of fundamental importance—learning to read—using methods that are ecological, friendly to application, and well controlled.

Conclusions. At the broadest level, reading is not simply an instructional process, but rather a complex developmental one. Children aren't simply taught to decode – they learn to do it in a complex environment in which the instructional experience at school plays an important role, but their exposure to text plays an equally important one. Our work adds to substantial prior work suggesting that this is largely a statistical or associative process. However, it also argues that specific types of statistics matter – children do not come to the table knowing what letters to ignore for a given GPC regularity and the frequency of "spurious" correlations between letters that are irrelevant for a given GPC regularity and phonemes may be quite important. Corpus work may help explain if and when this occurs and to characterize the natural types of variability that children encounter. A powerful approach to teaching children to read may thus be through framing the problem in terms of building more invariant mappings, rather than classifying word-types or families. This makes practical sense: when children read, they are not consciously trying to decide what family a word is, or work out some explicit rules; rather, they are trying to arrive at a pronunciation for it, and over time, they are trying to determine mappings that will consistently support the right pronunciations over many contexts and words.

However, to the extent that the instructional environment is a crucial aspect of reading development, the immediate application of our findings is clear: by applying the methods of laboratory learning to classroom settings we have demonstrated that using similar stimuli as a scaffold for GPC learning may not be the most effective pedagogical choice for teaching reading in English. Crucially, irrelevant variation can easily be applied to existing curricula, materials (e.g., worksheets) and interventions by manipulating the word lists to enhance variability in the graphemes (or phonemes) that are not directly relevant to GPC regularities. More importantly, it suggests that the variation among words and texts that children encounter while learning to read may be a critical developmental determiner of later outcomes (beyond the mere quantity of exposure) and may serve as a predictor of individual differences in outcomes.

Acknowledgements

Thanks to Donna Gaunt and Laura Porter for assistance with data collection, Dr. Bob Davis and the principals of the West Des Moines School District for help coordinating the study across schools, Foundations in Learning, Inc, for repurposing *Access Code* for this research, and Carolyn Brown and Jerry Zimmermann for helpful discussions and ongoing support throughout this project. This project was supported by NSF Grant BCS 1026794 awarded to BM and EH.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457-464. doi:10.1016/j.tics.2004.08.011

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? . *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1052-1086. doi:10.1037/0096-1523.24.4.1052

Apfelbaum, K., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science, 35*(6), 1105-1138. doi:10.1111/j.1551-6709.2011.01181.x

Arciuli, J., & Monaghan, P. (2009). Probabilistic cues to grammatical category in English orthography and their influence during reading. *Scientific Studies of Reading, 13*(1), 73-93. doi:10.1080/10888430802633508

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language, 63*, 180-196. doi:10.1016/j.jml.2010.03.005

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304. doi: 10.1111/j.1551-6709.2011.01200.x

Bates, D., & Sarkar, D. (2011). lme4: Linear mixed-effects models using S4 classes.

Baumann, J. F., Hoffman, J. V., Ro, J. M., & Duffy-Hester, A. M. (1998). Where are the teachers' voices in the phonics/whole language debate? Results from a survey of U.S. elementary classroom teachers. *The Reading Teacher, 51*(8), 636-650.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments and Computers, 41*, 977– 990. doi:10.3758/BRM.41.4.977

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language, 17*(1), 205–215. doi:10.1017/S0305000900013180

Cleeremans, A. (1997). Principles for implicit learning. In D. Berry (Ed.), How implicit is implicit learning? (pp. 195-234). Oxford, UK: Oxford University Press.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505. doi:10.1080/14640748108400805

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*, 589-608. doi:10.1037/0033-295X.100.4.589

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204-256. doi:10.1037/0033-295X.108.1.204

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of non-adjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1119-1130. doi:10.1037/0278-7393.30.5.1119

Del Rey, P., Whitehurst, M., & Wood, J. (1983). Effects of experience and contextual interference on learning and transfer. *Perceptual and Motor Skills, 56*, 581-582.

Ehri, L. C., Dryer, L., Flugman, B., & Gross, A. (2007). Reading rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. . *American Educational Research Journal, 44*(2), 414-448. doi:10.3102/0002831207302175

Ehri, L. C., Nunes, S. R., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*(3), 393-447. doi:10.3102/00346543071003393

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211. doi:10.1207/s15516709cog1402_1

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429-433. doi:10.1038/415429a

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*, 15822-15826. doi:10.1073/pnas.232472899

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37-55. doi:10.1037/0022-0663.90.1.37

Foundations in Learning Inc. (2010). Access Code. Iowa City, IA.

Gibson, E. J. (1970). The ontogeny of reading. *American Psychologist, 25*(2), 136-143. doi:10.1037/h0029419

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*(4), 674-691. doi:10.1037/0096-1523.5.4.674

Goldstone, R. (1996). Isolated and interrelated concepts. *Memory and Cognition, 24*, 608-628. doi:10.3758/BF03201087

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436. doi:10.1111/1467-9280.00476

Harm, M. W., McCandless, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading, 7*(2), 155-182. doi:10.1207/S1532799XSSR0702_3

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*(3), 491-528. doi:10.1037/0033-295X.106.3.491

Huet, M., Camachon, C., Gray, R., Jacobs, D. M., Missenard, O., & Montagne, G. (2011). The education of attention as explanation of variability of practice effects: Learning the final approach phase in a flight simulator. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1841-1854. doi:10.1037/a0024386

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*(4), 658-690. doi:10.1037/0096-3445.130.4.658

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills, 46*(2), 395-401.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/and/l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242-1255. doi:10.1121/1.408177

Lundberg, M. A., & Fox, P. W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research, 61*(1), 94-106. doi:10.2307/1170668

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*, 241-289. doi:10.1016/0167-9457(90)90005-X

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, 101-111. doi:10.1016/S0010-0277(01)00157-3

McCandliss, B., Beck, I., Sandak, R., & Perfetti, C. (2003). Focusing attention on decoding for children with poor reading skills: Design and preliminary tests of the Word Building Intervention. *Scientific Studies of Reading, 7*(1), 75-104. doi:10.1207/S1532799XSSR0701_05

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences, 6*(11), 465-472. doi:10.1016/S1364-6613(02)01993-9

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews: Neuroscience, 4*(1). doi:10.1038/nrn1076

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89-120). Mahwah, NJ: Lawrence Erlbaum Associates.

Namy, L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 13*(1), 5-15. doi:10.1037/0096-3445.131.1.5

Näslund, J. (1999). Phonemic and graphemic consistency: Effects on decoding for German and American children. *Reading and Writing, 11*(2), 129-152. doi:10.1023/A:1008090007198

Oakes, L., Coppage, D. J., & Dingel, A. (1997). By land or by sea: the role of perceptual similarity in infants' categorization of animals. *Developmental Psychology, 33*(3), 396-407. doi:10.1037/0012-1649.33.3.396

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory & Cognition, 23*, 324-354. doi:10.1037/0278-7393.23.2.324

Perrachione, T., Lee, J., Ha, L., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. . *Journal of the Acoustical Society of America, 130* (1), 461-472. doi:10.1121/1.3593366

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

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multilayered perceptron: Implications for child language acquisition. *Cognition, 38*, 1-60. doi:10.1016/0010-0277(91)90022-V

Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavioral Research Methods, 41*(4), 991-1008. doi:10.3758/BRM.41.4.991

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception, 22*(4), 463-475. doi:10.1068/p220463

Rayner, K., Foorman, B. R., Perfetti, C., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31-71. doi:10.1111/1529-1006.00004

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339-349. doi:10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning. *Infancy, 15*(6), 608. doi:10.1111/j.1532-7078.2010.00033.x

Saffran, J. R., Aslin, R. N., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., & Thiessen, E. D. (2007). Domain-general learning capacities. In E. Hoff & M. Shatz (Eds.), Handbook of Language Development (pp. 68-86). Cambridge, UK: Blackwell.

Schyns, P. G., Goldstone, R., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*(1), 1-54. doi:10.1017/S0140525X98000107

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning Memory and Cognition, 23*(3), 681-696. doi:10.1037/0278-7393.23.3.681

Seidenberg, M. S. (2005). Connectionist models of word reading. *Current Directions in Psychological Science, 14*, 238-242. doi:10.1111/j.0963-7214.2005.00372.x

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of visual word recognition and naming. *Psychological Review, 96*, 523-568. doi:10.1037/0033-295X.96.4.523

Seva, N., Monaghan, P., & Arciuli, J. (2009). Stressing what is important: Orthographic cues and lexical stress assignment. *Journal of Neurolinguistics, 22*, 237-249. doi:10.1016/j.jneuroling.2008.09.002

Thompson, G., Fletcher-Flinn, C., & Cottrell, D. (1999). Learning correspondences between letters and phonemes without explicit instruction. *Applied Psycholinguistics, 20*, 21-50. doi:10.1017/S0142716499001022

Torgeson, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K., Conway, T., & Rose, E. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*, 33-58. doi:10.1177/002221940103400104

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: A statistical approach to cue weighting and combination in speech perception. *Cognitive Science, 34*(3), 436-464. doi:10.1111/j.1551-6709.2009.01077.x

Treiman, R., & Kessler, B. (2006). Spelling as statistical learning: Using consonantal context to spell vowels. *Journal of Educational Psychology, 98*, 642-652. doi:10.1037/0022-0663.98.3.642

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88*(1), 49-78. doi:10.1016/S0010-0277(03)00003-9

U.S. Dept. of Education. (2010). *The nation's report card: Reading, 2009* (NCES 2010-458). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. http://www.iphod.com.

Wifall, T., McMurray, B., & Hazeltine, R. E. (submitted). Similarity impairs motor learning: Implications for the power law of learning?

Wolf, M., & Gow, D. (1986). A longitudinal investigation of gender differences in language and reading development. *First Language, 6*(81), 81-110. doi:10.1177/014272378600601701

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review, 9*, 185-211. doi:10.3758/BF03196276

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414-420. doi:10.1111/j.1467-9280.2007.01915.x

Appendix A. Words and non–words used in training

Vowel	SIMILAR WORD-LIST		VARIABLE WORD-LIST	
	Words	Non-words	Words	Nonwords
A	bat	cal	fan	zam
	hat	gat	pat	cag
	pat	hap	pal	dap
	cat	gad	lap	gad
	pal	lat	ram	gax
	bad	ral	cab	ral
I	bit	pid	sit	pid
	hit	git	hit	bip
	pit	gip	bid	fid
	lid	rit	wig	zib
	bid	mip	his	mip
O	rod	lod	hop	fob
	hop	pol	got	pol
	pot	mot	lot	wot
	cot	gop	mom	gop
	got	rol	rot	yom
AI	bait	laip	bait	Taib
	rail	cait	fair	cait
	raid	cail	rail	cail
	hail	haip	pain	vaid
	pail	pait	maid	raif
EA	heap	geat	bean	Feap
	meat	reat	meat	meab
	beat	gead	seal	gead
	bead	meap	gear	meap
	heal	leat	heal	seaf
	real	geap	weak	veam
OA	coat	boad	coat	Boam
	moat	poat	foam	voaf
	load	loap	load	zoal
	goat	hoat	toad	hoat
	boat	poad	soap	poad

Appendix B. Word lists used in testing

WORDS				NON-WORDS			
Trained Word	**Close Word**	**Far Word**	**Alternative-rule Word**	**Trained Word**	**Close Word**	**Far Word**	**Alternative-rule Word**
pat	fat	gas	pen	gad	yad	waz	bep
pal	pad	sag	bed	ral	rav	vab	med
hit	sit	rim	peg	pid	pim	yim	fen
bid	big	tip	mud	mip	fip	tiv	vub
hop	top	fox	hug	pol	pon	rog	sut
got	god	bog	bus	gop	vop	zon	hup
bait	bail	hail	boot	cait	caif	naig	noop
rail	rain	vain	room	cail	zail	jaif	soom
meat	mean	leap	sour	gead	geam	rean	boud
heal	meal	dear	pout	meap	meag	heak	foup
coat	coal	roam	beef	hoat	yoat	roan	beel
load	road	boar	reed	poad	poam	goaf	meef

Appendix C. Measuring consistency of GPC regularities

To determine how consistent the GPC regularities used in this study were, we consulted the orthographic and phonological forms present in the MRC Psycholinguistic Database. Although these measures are somewhat coarse, the MRC is a fairly clear standard in the field of visual word recognition, and there are few comparison databases as large with both orthographic and phonological information. Measures from this database should offer a relative measure of consistency of given orthographic-phonological mappings.

There are multiple ways that consistency can be measured in a language. The traditional measure of consistency of a GPC regularity asks how often a given letter (or set of letters) yields a certain pronunciation given a highly constrained context. That is, we can ask how often the letter A yields an /æ/ sound when it is the only vowel in a word. A more conservative method presumes that the child does not yet know that vowels and consonants should be treated differently. As such, we can ask how often an A in the middle of the word elicits the /æ/ sound, regardless of the class of surrounding letters. We present both measures, and discuss potential differences in interpretation between the two.

Using the traditional measure, we searched the MRC Psycholinguistic Database for all single-syllable words with the vowel (or vowels) of interest anywhere medially in their orthographic forms, and with no other vowels in the word. We excluded all words without a phonological transcription. Next, we determined what proportion of the candidate items had the pronunciation predicted by the GPC regularity. All six regularities included in our study involved the dominant pronunciation by this measure (dominance measured as more than half of orthographic forms produced the predicted pronunciation; Table A1).

For the more conservative measure, we again searched the MRC Database for all single-syllable words containing specific vowels or vowel pairs word medially that had a phonological transcription. However, now we did not consider the presence of other vowels in the words. This analysis gave a very different picture of the consistency of the regularities: the monograph regularities were much less consistent, while the digraphs were still highly consistent (Table A1).

This second measure suggests that early in learning to read, educating children on the difference between consonants and vowels can make it much easier for them to learn certain regularities. Without this knowledge, they must learn that the A in BAT is fundamentally different than the A in BOAT on their own; under such a system, learning digraph vowel regularities may prove easier than learning monographs.

Table C1. Consistency of the GPC regularities trained by two metrics

	Orthography	**Pronunciation**	**Example**	**Consistency: Measure 1**	**Consistency: Measure 2**
Short Vowels	A	/æ/	HAT	52.1%	26.7%
	I	/ɪ/	HIT	82.1%	56.7%
	O	/ɑ/	HOT	52.3%	23.8%
Digraphs	EA	/i/	HEAT	75.0%	76.5%
	OA	/ou/	HOE	84.6%	81.8%
	AI	/ɑɪ/	HIGH	76.0%	78.2%