Simultaneous training on overlapping grapheme phoneme correspondences augments learning and retention

Tanja C. Roembke a,⇑, Michael V. Freedberg b, Eliot Hazeltine c, Bob McMurray c,d,e

a Institute of Psychology, RWTH Aachen University, D-52066 Aachen, Germany
b Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD 20817, USA
c Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, USA
d Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242, USA
e Department of Linguistics, University of Iowa, Iowa City, IA 52242, USA

Article Info
Article history:
Received 10 April 2019
Revised 8 October 2019

Keywords:
Reading
Statistical learning
Vowel overlap
Variability
Reading acquisition
Grapheme-phoneme-correspondence (GPC) regularities

Abstract
An important component of learning to read is the acquisition of letter-to-sound mappings. The sheer quantity of mappings and many exceptions in opaque languages such as English suggests that children may use a form of statistical learning to acquire them. However, whereas statistical models of reading are item-based, reading instruction typically focuses on rule-based approaches involving small sets of regularities. This discrepancy poses the question of how different groupings of regularities, an unexamined factor of most reading curricula, may affect learning. Exploring the interplay between item statistics and rules, this study investigated how consonant variability, an item-level factor, and the degree of overlap among the to-be-trained vowel strings, a group-level factor, influence learning. English-speaking first graders (N = 361) were randomly assigned to be trained on vowel sets with high overlap (e.g., EA, AI) or low overlap (e.g., EE, AI); this was crossed with a manipulation of consonant frame variability. Whereas high vowel overlap led to poorer initial performance, it resulted in more learning when tested immediately and after a 2-week-delay. There was little beneficial effect of consonant variability. These findings indicate that online letter/sound processing affects how new...
knowledge is integrated into existing information. Moreover, they suggest that vowel overlap should be considered when designing reading curricula.

© 2019 Elsevier Inc. All rights reserved.

Introduction

The complexity of learning to read

Literacy is correlated with academic success (National Reading Panel, 2000) and life outcomes (e.g., Beck, McKeown, & Kucan, 2002). However, 60% of students in the United States lack basic reading skills (National Center for Education Statistics, 2010). This may in part be due to the complexity of the problem children face; reading a sentence requires word recognition, use of text conventions, and syntactic and semantic processing. Even the most elemental aspect of reading—isolated word recognition—is complex given that the sound–spelling system of many languages is irregular and children must sort through a large number of lexical candidates.

There are too many words for children to memorize their forms. As a result, in alphabetic languages phonics-based curricula teach children to map letters to sounds. This mapping offers children generalizable regularities—so-called grapheme–phoneme correspondence (GPC) regularities—that allow a child to access the sound of an unfamiliar word from its spelling. For example, once a child learns that EA makes the /i/ sound, he or she can read a variety of words such as MEAT and GLEAM. Thus, one critical aspect of learning to read is acquiring GPC regularities.1

There has been growing interest in applying principles of learning from cognitive science to education (e.g., (Apfelbaum, Hazeltine, & McMurray, 2013; Kellman, Massey, & Son, 2010; Rohrer, Dedrick, & Stershic, 2015)). The current study builds on this effort by examining the acquisition of GPC mappings to support decoding skills. We focused on a disconnect between statistical learning theories of reading acquisition that emphasize exposure to large numbers of items and common classroom approaches that emphasize teaching rules.

Statistical learning

In English and other orthographically deep languages, GPC regularities are only quasi-regular. For any “rule” (e.g., EA → /i/), there are many exceptions (e.g., STEAK) and sub-regularities (e.g., HEAD, DEAD) (Seidenberg, 2005). Thus, acquiring these regularities can be challenging.

This quasi-regularity of many languages motivates the hypothesis that GPC mappings are acquired via statistical learning mechanisms that track predictive relationships among letters and between letters and sounds (Seidenberg, 2005). Supporting this, statistical learning ability (in visual and auditory domains) is correlated with reading outcomes (Arciuli & Simpson, 2012; Qi, Sanchez Araujo, Georgan, Gabrieli, & Arciuli, 2019; Raviv & Arnon, 2018; Spencer, Kaschak, Jones, & Lonigan, 2015). Moreover, the pattern of children’s spelling errors often conforms to the statistical structure of the input (Pollo, Kessler, & Treiman, 2009; Thompson, Fletcher-Flinn, & Cottrell, 1999; Treiman, Gordon, Boada, Peterson, & Pennington, 2014; but see Sénéchal, Gingras, & L’Heureux, 2016), and children are sensitive to consonantal context when spelling vowels in nonwords (Treiman & Kessler, 2006). Together, these data suggest that the acquisition of sound/spelling correspondence may derive from statistical learning. Although the general relationship between statistical learning and reading is well supported, it is not clear how it develops and/or what other factors influence the extent to which

1 We use this term to describe regularities in the orthographic/phonological language system. We do not use it to imply a specific rule-based representation as in models like Coltheart, Rastle, Perry, Langdon, and Ziegler (2001).
statistical learning occurs (see Alt, 2018, Elleman, Steacy, & Compton, 2019, and Schmalz, Moll, Mulatti, & Schulte-Körne, 2019, for discussions).

In models of statistical learning, learning is fundamentally item-based and GPC regularities are not explicitly stored: Learning a GPC regularity is conceptualized as emergent from experience with individual words. In this case, children are not acquiring individual regularities but rather are acquiring a system of mappings that “works” across all words. Supporting this, Armstrong, Dumay, Kim, and Pitt (2017) showed that after adults are taught a small set of nonwords with irregular pronunciations, they spontaneously generalize this to a new set of untrained nonwords.

Although this documents a potential role of statistical learning, it is unclear how to use this to improve instruction or remediation. (Apfelbaum, Hazeltine, & McMurray, 2013) offered a first step by manipulating item-level statistics (e.g., within the set of words used in training) to promote acquisition of GPC regularities for vowels. In English, vowels are most challenging for early readers because they are less regular than consonants (Fowler, Liberman, & Shankweiler, 1977; Näslund, 1999) and may be harder to process (New, Araújo, & Nazzi, 2008).

When confronted with a whole word (e.g., TEAM), children must learn that the vowel letters (EA) are most predictive of the vowel pronunciation, not the consonants. However, if children initially weighted all letters equally, this could create problems. If children were trained on words with similar consonants (as is typical in many curricula), they might acquire spurious associations between the vowel pronunciation and the consonants. This could impair learning and hinder generalization (see Juel & Roper-Schneider, 1985). For example, after exposure to HAT, CAT, and MAT, children could learn that the /æ/ sound is predicted by AT, not by A. The solution to this problem is to train on items that are more variable on irrelevant dimensions (Gómez, 2002; Rost & McMurray, 2009) so that only the vowels consistently predict pronunciation.

To test this, (Apfelbaum, Hazeltine, & McMurray, 2013) trained children on six GPC regularities (three digraphs and three short vowels) using a computer-based reading intervention that implemented a variety of sound/spelling tasks (Foundations in Learning, 2010). Target vowels were embedded in training word lists where consonants were either similar (e.g., MET, PET, MOAT, POT) or variable (e.g., MET, LEG, LOAF, TOP). Children in the variable condition showed larger gains at posttest for trained and novel words. Thus, manipulations at the item level—motivated by a learning theory—can benefit instruction.

Apfelbaum, Hazeltine, & McMurray, (2013) findings are consistent with a naïve associative account of statistical learning (i.e., the blind tracking of items’ co-occurrence); variability impedes the formation of incorrect associations between consonant letters and the vowel phonemes, whereas similarity in the consonant frames reinforces them.

Explicit teaching of regularities

Where statistical learning focuses on statistics across items, reading instruction is typically structured around regularities. The emphasis on regularities offers practical benefits; explicit regularities can be easily described and conceptualized explicitly, allowing students to bootstrap learning. However, there are too many GPC regularities to teach all of them. Thus, curricula often focus on GPC regularities that generalize to a large number of words (cf. Fry, 1964), permitting children to learn many new words independently (Share, 1995). Further constraining the space, typical curricula focus on only a subset of these regularities at a time (e.g., short vowels) to make it easier to conceptualize the regularities.

In contrast, statistical learning emphasizes properties of individual items and only treats regularities as emergent. In this case, the optimal training set for instruction should mirror the distribution of letters in the language (e.g., in the extreme, all possible items). Consequently, the groupings imposed by typical curricula limit the range of statistics that children are exposed to and may bias learning. Importantly, because these groupings are brief and change over the course of the curricula (e.g., children are exposed to short vowels for 1 or 2 weeks and then long vowels), this can create further local biases in the statistics.
Little research has examined how the grouping of GPC regularities affects learning. For example, the short vowels E (as in MET) and A (as in CAP) are typically grouped in most curricula. However, discriminating different GPCs within a class (e.g., short vowels) may be a trivial memorization process because they have no letters in common. In contrast, a harder problem may be discriminating a word that includes an E and makes the /e/ sound (e.g., MEN) from a word that includes an E and makes the /i/ sound (e.g., MEAN). Thus, grouping GPC regularities that have no letters in common might not offer any benefits. As this example illustrates, understanding how learning is affected by structure at the level of the generalizable regularities (in addition to structure at the item level) may have consequences for the basic understanding of how children learn to read as well as implications for structuring curricula.

The current study examined this issue, focusing on whether or not instruction should group GPC regularities whose orthographic forms “overlap.” In digraph vowels, letters can combine to represent a phoneme(s) (as in EA, AI, and OI). Unlike the consonant manipulations of (Apfelbaum, Hazeltine, & McMurray, 2013), where the consonants need to be ignored, here the conjunction of letters is critical and the challenge is to map the combination of letters to a sound that might not be associated with either individual letter. For example, both OA and EA contain an A, and A can also function in isolation. All three strings, however, map to different phonemes, making the interpretation of A contextually dependent. Most curricula do not systematically consider this type of overlap, instead grouping regularities by conceptually related regularities (e.g., long vowels).

We contrast two predictions. First, the naïve associative account (Gough & Juel, 1991; Juel & Roper-Schneider, 1985) suggests that overlapping vowels may be more difficult to learn because the same individual letters must be associated with different sounds. However, variability in consonant frames could highlight the overlapping vowels as a unit, potentially mitigating the effect of overlap. Alternatively, statistical learning has also been conceptualized in terms of acquiring a system of mappings (e.g., as in connectionist models). Here, overlap may be advantageous; connectionist modeling suggests that new information that is consistent with established knowledge may be more easily integrated with existing schemas (McClelland, 2013). This “schema-based” account predicts that simultaneous training on overlapping GPC should benefit learning. It is unclear whether consonant variability will interact with overlap under this view.

To distinguish these hypotheses, we taught first graders a small number of GPC regularities for short vowels and digraphs over several days. Overlap was manipulated via the sets of vowels on which children were trained and tested. These either maximized shared letters across the set (EA, OA, OI, AI, E, and O) or minimized them (AI, OO, EE, OU, I, and U). To relate this regularity-based manipulation to an item-based factor, we further asked whether the role of overlap is moderated by consonant variability. Consonant variability was manipulated by selecting words that had either variable or similar consonant frames. This led to a $2 \times 2$ between-participants design.

To deliver training and testing, we repurposed an internet-based reading intervention program (Foundations in Learning, 2010). We tested first graders for two reasons. First, they match the children tested in (Apfelbaum, Hazeltine, & McMurray, 2013), allowing us to follow up on the effect of variability. Second, they are still in the process of acquiring these GPC mappings, so they were unlikely to perform at ceiling but had sufficient skills to perform the tasks.

Students’ baseline ability was first assessed with a pretest, incorporating a variety of simple tasks tapping the six GPC regularities. Students then completed training with feedback on each trial over approximately 1 week, followed by a posttest. We also included a third testing point 1 to 2 weeks later to assess retention. Unlike (Apfelbaum, Hazeltine, & McMurray, 2013), GPCs were fully interleaved across training to improve learning and maximize the effect of overlap (Carvalho & Goldstone, 2014).
Method

Overview

The experiment included four phases: pretest, training, posttest, and retention. Pretest, posttest, and retention had an identical design. There were four experimental conditions (High Overlap/Low Overlap GPCs × Similar/Variable consonants). These were crossed with two conditions for counterbalancing task order (A/B). Overlap was manipulated throughout the experiment, whereas variability was manipulated only during training.

Participants

Participants were 316 first-grade students (average age = 6.46 years) from eight elementary schools in the West Des Moines School District, Iowa. All first-grade students without an individualized education plan for a learning disability were invited to participate. Approximately 50% of first graders in the eight schools participated, yielding a wide distribution of reading abilities. Of the 316 students enrolled in the study, 281 completed pretest, training, and posttest. The remaining 35 withdrew during pretest or training or were excluded based on behavioral issues reported by the research staff members (who were not aware of the hypotheses or students’ conditions). Two children did not complete posttest immediately after training but did complete retention; their data were analyzed for the retention phase only. Two additional children completed posttest data 5 days after the end of training (and they completed retention shortly thereafter). We excluded their posttest data but kept their retention data in the analysis. Thus, data from 277 children were analyzed at posttest, and data from 267 children were analyzed at retention.

The final sample included 32 non-native English speakers, roughly equally distributed across conditions. These children’s data were included because English language learner status did not interact with experimental factors in similar work (Apfelbaum, Hazeltine, & McMurray, 2013). Table 1 shows the demographic breakdown of the groups.

We obtained children’s reading assessment scores on the Fountas & Pinnell Benchmark Assessment System from the school district. Letter scores (A–Y) were converted to an ordinal scale ranging from 0 to 23 with an average of 6.74 (SD = 4.08). A score of 6 or 7 (equivalent to letter scores F and G) roughly maps onto the upper level of what is expected for first graders (Fountas & Pinnell, 2011). Moreover, the range and standard deviation of scores suggest a fair amount of variability in students’ reading ability. Participants’ reading scores were used as a between-participant factor for moderator analyses (see Supplement S2 in online supplementary material).

Design

Overlap

Children learned one of two sets of six GPC regularities (high overlap/low overlap). These GPC regularities captured common digraphs and monographs of English that have high spelling–sound consistency. Each set included two short vowels (e.g., E as in MEN, O as in POT) and four digraphs (e.g., EA as in MEAN, AI as in MAIL) (Fig. 1 and Table 2). One vowel (AI) was used in both overlap

| Table 1 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Demographic breakdown of participants across conditions for the final dataset. |
| High overlap | Low overlap | Total |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Variable | Similar | Variable | Similar |                |                |                |
| Gender         | 35          | 33          | 33          | 38            | 139          |
| Boys           |              |              |              |                |                |                |
| Girls          | 37          | 32          | 38          | 35            | 142          |
| First language | 67          | 57          | 62          | 63            | 249          |
| English        |              |              |              |                |                |                |
| Other          | 5           | 8           | 9           | 10            | 32           |
conditions. Each GPC regularity represented the dominant pronunciation for the given letter string (see Table 2 and Supplement S1 in supplementary material). Thus, our investigation focused on learning the most regular GPCs (not exception or secondary pronunciations).

Given only five vowel letters, it was not possible to construct completely non-overlapping sets. On average, each GPC regularity in the high overlap set shared a letter with 2.67 others (range = 1–4), where each regularity in the low overlap set shared a letter with 1.0 other regularities (range = 0–2). The GPC regularities in each condition were balanced on overall difficulty using procedures outlined in Supplement S1.

Variability

Within each overlap condition, there were two item lists that differed in the amount of variability in the consonants. Similar item lists included words that shared one consonant or both consonants (e.g., MEAL, MOAN, PEN, MEN); words in the variable group were selected to minimize the number of shared consonants. For more information on how variability was manipulated and items were selected, see Supplement S1.

Relation of overlap and variability to phase

The overlap manipulation was based on the set of GPCs on which children were trained (not individual items). Thus, the overlap condition controlled the words used in all experiment phases (training...
and tests). In contrast, variability was an item-level manipulation of the consonants that did not affect the GPCs. Consequently, the variability manipulation affected only the items used in training.

This led to a hybrid random assignment scheme. First, participants were assigned to the overlap condition at the onset of the experiment, balancing gender, school, and language status. After pretest, we performed random assignment to the respective variability condition. Because pretest ability was known at this point, assignment for variability could balance pretest scores in addition to demographic factors and task set. This procedure meant that whereas pretest scores were well matched within variability conditions, pretest scores were not matched across overlap condition as anticipated (see also Results section).

Tasks
To ensure that training and testing tapped a robust ability and to keep students engaged, eight tasks were used (Table 3). For any given participant, two tasks were not included in training and appeared in testing only to assess generalization. We counterbalanced which tasks were held out for generalization testing with two sets (A/B).

Students performed each closed set task in multiple blocks of 10 trials, switching frequently. Tasks were grouped into “cycles,” with each cycle including a task selection screen. After completing the tasks in a cycle, students advanced to the next cycle. Half of the tasks used real words, and the other half used nonwords. Nonwords allowed a more thorough manipulation of the overall amount of consonant variability and a more direct test of grapheme → phoneme knowledge given that children cannot perform well by “memorizing” the words. Each trial included 10 response options (chance = 10%). Foil responses were drawn from other items within that block unless this would have led to repetitions. In the latter case, additional foils were added to create 10 unique options.

Table 3
Tasks used.

<table>
<thead>
<tr>
<th>Experiment phase</th>
<th>Task</th>
<th>Description</th>
<th>Task type</th>
<th>Testing role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing</td>
<td>Find the Word</td>
<td>Hear a word played and find that word among 10 displayed alternatives.</td>
<td>Recognition</td>
<td>Generalization A B Trained</td>
</tr>
<tr>
<td></td>
<td>Fill in the Blank (Nonword)</td>
<td>Hear a nonword and see a consonant frame and 10 vowel options. Choose which vowel completes the played nonword.</td>
<td>Spelling</td>
<td>Generalization A B Trained</td>
</tr>
<tr>
<td></td>
<td>Make the Word</td>
<td>Hear a word and choose the letters to spell it from 10 displayed alternatives for each position.</td>
<td>Spelling</td>
<td>Trained Generalization</td>
</tr>
<tr>
<td></td>
<td>Nonword Family</td>
<td>Hear the vowel and coda consonant of a nonword and find the nonword that contains those sounds among 10 alternatives.</td>
<td>Recognition</td>
<td>Trained Generalization</td>
</tr>
<tr>
<td>Training</td>
<td>Word Family</td>
<td>Hear the vowel and coda consonant of a word and find the word that contains those sounds among 10 alternatives.</td>
<td>Recognition</td>
<td>Not applicable</td>
</tr>
<tr>
<td></td>
<td>Change the Word (Vowel)</td>
<td>See a consonant frame and 10 vowel options. Asked aurally to change one word to another.</td>
<td>Spelling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Find the Nonword</td>
<td>Hear a word played and find that nonword among 10 displayed alternatives.</td>
<td>Recognition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change the Nonword (Initial)</td>
<td>See vowel and offset consonant and 10 onset consonants. Asked to change one nonword to another.</td>
<td>Spelling</td>
<td></td>
</tr>
</tbody>
</table>

Note. Two of the testing tasks could also appear in training; these are marked as “Trained.” Tasks that appeared only in testing are marked as “Generalization.”
Testing
Each test (pretest, posttest, and retention) consisted of three cycles of four tasks, resulting in 120 total trials. Items differed between overlap conditions to reflect the different GPC regularities but were identical between the similar and variable conditions. For each GPC regularity and item type (word/nonword), five types of items were tested: two trained words, two untrained (generalization) words from the same GPC regularity, and a baseline word not from any of the trained GPC regularities. Items were presented three times during test.
Words were randomly assigned to block in a fixed order for each student, roughly balancing the number of items from each GPC regularity within a block. Within a block, item order was random. During testing, students did not receive feedback.
Testing sessions were run over 2 days. Participants were logged out of the program automatically after completing 60 trials (halfway through the second cycle). The following day, they completed the remaining 60 trials. Pretest generally occurred on Thursday and Friday to allow for the second phase of random assignment over the weekend, although some children missed a day and so pretest was extended to the following week. Posttest was conducted on the day immediately after the last day of training (or the following Monday if training finished on a Friday). Retention was conducted at least a week after posttest (typically 2 weeks), depending on the school’s schedule.

Training
Training consisted of five cycles of six tasks for a total of 300 training trials. Training items differed as a function of overlap and similarity. There was an equal number of trials for each GPC regularity, roughly equal within tasks. For each GPC regularity, five words and five nonwords were selected (60 items total). Each item was repeated five times during training.
During training, children always received feedback via three mechanisms. They heard a high tone for correct answers and a low tone for incorrect answers. On incorrect trials, students repeated the trial after hearing “Try again!” On repetitions, the initial choice was removed from the screen, and for spelling tasks buttons were added to allow children to hear the sound corresponding to each response. Use of the buttons was optional and infrequent. If children responded incorrectly twice, they were told the correct answer. The score was accumulated across trials within each task. Children received 10 points for a correct answer on the initial attempt and 5 points for a correct answer on the second attempt.
On each training day, participants ran for a fixed period of 20 min, at which point they were automatically logged out of the system and would continue where they had left off the following day. When participants completed the total number of training trials, they were done.

Items and stimuli
Items were matched on frequency, imageability, concreteness, and neighborhood density across conditions (see Supplement S1). All words, training instructions, letter cues, and carrier sentences were recorded by a phonetically trained native speaker of English. Recordings were conducted in a sound-attenuated room using a Kay Elemetrics CSL 4300B machine (Kay Elemetrics, Lincoln Park, NJ, USA) at a 44,100-Hz sampling rate. For each item, the speaker produced three to five exemplars in a neutral carrier phrase. These recordings underwent noise reduction using Audacity (Audacity Team, 2014), to remove background noise. Next, the clearest exemplar of a given item was selected from the phrase. Then, 50 ms of silence was spliced onto the onset and offset of each word, followed by amplitude normalization using Praat (Boersma & Weenink, 2016).

Procedures
The study was conducted between February and May 2014. Pretest, training, and posttest lasted for about 1.5 weeks at a given school, with an additional 2 days of retention testing. Stimuli were presented over high-quality headphones to minimize disruption from other students taking part in the study at the same time.
After logging in, children saw a task selection screen that displayed icons for all possible tasks. Each completed task was marked with a checkmark and showed how many points were awarded within that task (during training). After completing a cycle, participants were presented with a new task selection screen in a new color.

Each task began with auditory instructions. Trials started with a carrier sentence that included the target stimulus. For instance, in the Find the Word task, students were told to “Find the word MEAT.” This was accompanied by 10 written responses. Children responded using a computer mouse. Students received positive reinforcement (e.g., hearing “Thank you for working so hard!” or “Great job!”) during both testing and training; these were given approximately every five trials and were independent of accuracy on a particular trial.

Analysis

The dependent measure in all analyses was accuracy. Baseline words (untrained words from untrained GPC classes) were not analyzed. There was a large effect of vowel overlap at pretest (Table 4). Thus, to quantify the amount of learning (rather than differences in postlearning performance), we focused on change scores (posttest – pretest). To investigate learning in practical terms, we calculated untransformed change scores between pretest and either posttest or retention. This captures meaningful differences in performance, much as classroom learning would be evaluated. However, this analysis could also reflect ceiling or floor effects (larger gains are possible when pretest scores are low than when they are high). Thus, we also scaled scores using an empirical logit scale using the empirical logit transformation (which stretches the space near 0 and 1) and computed change scores between logit scaled pretest and posttest (or retention) accuracy (Fischer-Baum, 2014). The empirical logit transformation is a standard statistical solution to overcome differences in cases where two groups are in the learning curve because it is easier to improve when one is in a medium accuracy range than when one is close to ceiling. Thus, it allowed us to counteract the anticipated accuracy differences at pretest.

Results

Pretest

We first examined pretest to detect differences in baseline performance due to overlap (Table 4). Participants in the High Overlap × Variable condition and the High Overlap × Similar condition performed lowest, whereas those in the Low Overlap × Variable condition and the Low Overlap × Similar condition did better. There was relatively little variability in how well students performed (SEM always < 3%).

Because there were no main effects or interactions with variability, analyses of pretest were collapsed across this condition. We performed an analysis of variance (ANOVA) examining overlap and vowel-type (monograph vs. digraph, within participants). We found a significant main effect of overlap, $F(1, 279) = 10.53$, $p = .001$, $\eta^2_p = .036$, with lower performance in the high overlap condition. We also found a significant main effect of vowel type, $F(1, 279) = 258.24$, $p < .001$, $\eta^2_p = .481$, supporting overall lower performance for digraphs (Fig. 2A). There was a significant interaction between overlap and vowel type, $F(1, 279) = 31.38$, $p < .001$, $\eta^2_p = .101$. This was driven by a significant effect of overlap for digraphs, $F(1, 279) = 34.79$, $p < .001$, $\eta^2_p = .111$, but not for monographs ($F < 1$).

Whereas GPC regularities were balanced on initial difficulty (see Supplement S1), it is possible that differences between groups were due to the GPC regularities in the high and low overlap conditions. We tested this by examining performance on Al, which appeared in both conditions (Fig. 2B), although the specific items differed. Al showed a similar pattern with about a 10% decrement in performance for the high overlap group, $F(1, 279) = 17.89$, $p < .001$, $\eta^2_p = .060$. This suggests that the differences

---

2 In the Make the Word task, children spelled the target item by essentially making three separate responses: selecting the onset consonant(s), the vowel, and then the coda consonant(s). For this task, we used only vowel accuracy because this was our primary interest.
between conditions were unlikely to have derived from the difficulty of the individual GPC regularities but rather emerged from interrelationships among GPC regularities in the set. To determine whether these effects could have arisen from spurious differences between the students, we conducted a 2 \times 2 (Overlap \times Variability) ANOVA with reading scores (collected by the district) as the dependent variable. We found no significant differences in reading score as a function of overlap, $F(1, 277) = 1.81$, $p = .179$, $g^2_p = .006$, or variability ($F < 1$), and there was no interaction ($F < 1$). Thus, the effect of overlap at pretest is not a consequence of preexisting differences between groups.

**Posttest**

We next examined learning gains (change in performance) at posttest as a function of condition (vowel overlap and consonant variability). Fig. 3A shows accuracy for each group at pretest and posttest. First, we examined posttest change scores using untransformed accuracy with a $2 \times 2$ between-participant (Overlap \times Variability) ANOVA. We found a significant effect of overlap, $F(1, 273) = 7.05$, $p = .008$, $g^2_p = .025$, with larger learning gains in the high overlap condition than in the low overlap condition. Variability was not significant, $F(1, 273) = 1.09$, $p = .297$, $g^2_p = .004$, and did interact with overlap, $F(1, 273) < 1$ (Fig. 3A and 3B). In an item analysis, a mixed $2 \times 2$ (Overlap \times Variability) ANOVA was used (with variability as within items and overlap as between items). Overlap remained significant, $F(1, 94) = 8.09$, $p = .005$, $g^2_p = .079$. There was a significant effect of variability, $F(1, 94) = 9.64$, $p = .003$, $g^2_p = .093$, with more learning after variable training than after similar training, but no interaction, $F(1, 94) < 1$. Note that variability may have been significant by items (but not by participants) because in this analysis variability is within items (whereas it was between participants).

We next applied the empirical logit transformation to accuracy and recalculated the difference score (Fig. 3C and 3D). This showed no significant effect of overlap, $F(1, 273) = 2.44$, $p = .120$, $g^2_p = .009$. The effect of variability, $F(1, 273) = 3.35$, $p = .068$, $g^2_p = .012$, was marginally significant.
showing slight advantages in variable conditions. Overlap and variability did not interact ($F < 1$). The item analysis found significant effects of overlap, $F(1, 94) = 8.18, p = .005, \eta^2_p = .080$, and variability, $F(1, 94) = 6.10, p = .015, \eta^2_p = .061$, but no interaction ($F < 1$). These results are similar to those of the untransformed item analysis; there was a tendency toward better learning with high overlap or variable consonants.

We followed up on these analyses by asking whether a variety of moderating factors influenced any of these factors (see Supplement S2). These included whether tasks and items were trained or generalization, word/nonword, gender, and reading ability. Only a few moderators were identified (vowel type was significant, and there was a marginal interaction with reading ability). This suggests that this pattern held across a variety of conditions.

Finally, given the small learning gains in the low overlap group, untransformed change scores were used to investigate whether there was measurable learning in each sub-condition. We compared the untransformed change scores with zero using a one-sample $t$ test (Table 5). We found significant learning in all groups. Learning gains ranged from 6.5% to 12.9%, which were considerable given the relatively short amount of training.

**Retention**

Finally, we examined retention (Fig. 4). We first examined untransformed retention change scores (retention – pretest) in a $2 \times 2$ ANOVA. A significant effect of overlap was observed, $F(1, 263) = 9.67, p = .002, \eta^2_p = .035$, with better retention in the high overlap condition than in the low overlap condi-
No other effect reached significance [variability: \( F < 1 \); Variability \( \times \) Overlap: \( F(1, 263) = 1.20, p = .275, \eta^2_p = .005 \)].

The item analysis replicated the significant effect of overlap, \( F(1, 94) = 18.87, p < .001, \eta^2_p = .167 \). As with posttest, the item analysis also showed evidence for a variability effect, \( F(1, 94) = 7.51, p = .007, \eta^2_p = .074 \), but this time there was a significant interaction, \( F(1, 94) = 16.38, p < .001, \eta^2_p = .148 \). We examined this by separating the data by overlap condition and comparing similar and variable conditions within each (with a factorial ANOVA adjusted using the Bonferroni correction, \( \alpha = .025 \)). For the low overlap condition, there was no effect of variability (\( F < 1 \)). However, for the high overlap condition, there was a significant effect of variability, \( F(1, 47) = 28.37, p < .001, \eta^2_p = .376 \). Unlike at posttest, retention was better with similar consonants than with variable consonants for high overlap students.

We next evaluated retention with logit transformed scores. We again found a significant effect of overlap, \( F(1, 263) = 4.17, p = .042, \eta^2_p = .016 \), but no effect of variability (\( F < 1 \)) or interaction (\( F < 1 \)).

### Table 5

<table>
<thead>
<tr>
<th>Overlap group</th>
<th>Variability group</th>
<th>Mean ± standard deviation</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>High overlap</td>
<td>Similar</td>
<td>.112 ± .119</td>
<td>( t(64) = 7.57, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td>Variable</td>
<td>.129 ± .133</td>
<td>( t(68) = 8.07, p &lt; .001 )</td>
</tr>
<tr>
<td>Low overlap</td>
<td>Similar</td>
<td>.065 ± .176</td>
<td>( t(72) = 3.16, p = .002 )</td>
</tr>
<tr>
<td></td>
<td>Variable</td>
<td>.084 ± .142</td>
<td>( t(69) = 4.90, p &lt; .001 )</td>
</tr>
</tbody>
</table>

---

**Fig. 4.** (A) Raw proportions correct at pretest and retention across all the Overlap \( \times \) Variability conditions. (B) Untransformed retention change scores for all conditions. (C) Transformed accuracy at pretest and retention across all conditions. (D) Transformed retention change scores across all conditions. All panels include standard errors of the mean as error bars.
item analysis using the transformed change scores also found a significant effect of overlap, $F(1, 94) = 22.97, p < .001, \eta^2_p = .196$. There was also a significant effect of variability, $F(1, 94) = 9.86, p = .002, \eta^2_p = .095$, and a significant interaction, $F(1, 94) = 15.58, p < .001, \eta^2_p = .142$. We split the data by overlap condition and adjusted for repeated tests using the Bonferroni correction ($\alpha = .025$). Again, we found a significant effect of variability in the overlap condition only [high overlap: $F(1, 47) = 34.29, p < .001, \eta^2_p = .422$; low overlap: $F(1, 47) < 1$], with better retention for similar consonants in the high overlap condition.

Finally, we asked whether students in each group showed nonzero retention (Table 6). Only students in the high overlap groups exhibited significant retention. In fact, change scores for overlap were similar to those at posttest, suggesting only minimal loss over the testing delay.

Overall, retention was better with high overlap GPCs than with low overlap GPCs. This was true for both untransformed and transformed scores, suggesting advantages for long-term retention with overlapping materials. There was also an overall benefit of similarity, although item analyses indicated that this benefit may be restricted to the high overlap condition. Moderator analyses (see Supplement S2) indicated that the only significant moderator was word/nonword status. The effect of overlap on retention appeared to be driven by the words. Other task, item, and participant factors showed little effect.

**Discussion**

**Overview**

A marked effect of overlap was evident before training; at pretest, children performed worse on the high overlap GPC than on the low overlap GPC. That pattern was more complex after training (see Table 7 for a summary). At immediate posttest, larger learning gains were observed with high overlap in raw change scores, but evidence was less robust with the empirical logit transformation. More important, retention was consistently better for students in the high overlap condition than in the low overlap condition. This overlap benefit is notable given that initial performance in the high

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Planned comparisons of retention change scores.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap group</td>
<td>Variability group</td>
</tr>
<tr>
<td>High overlap</td>
<td>Similar</td>
</tr>
<tr>
<td></td>
<td>Variable</td>
</tr>
<tr>
<td>Low overlap</td>
<td>Similar</td>
</tr>
<tr>
<td></td>
<td>Variable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Summary of major findings.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap</td>
<td>Posttest</td>
</tr>
<tr>
<td></td>
<td>Participant</td>
</tr>
<tr>
<td>Raw</td>
<td>O</td>
</tr>
<tr>
<td>Logit</td>
<td>–</td>
</tr>
<tr>
<td>Moderators</td>
<td>Digraphs/monographs</td>
</tr>
<tr>
<td>Variability</td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>–</td>
</tr>
<tr>
<td>Logit</td>
<td>v</td>
</tr>
<tr>
<td>Moderators</td>
<td>Spelling/recognition task+</td>
</tr>
</tbody>
</table>

Note. For factors (overlap/variability), an uppercase letter indicates a significant effect and a lowercase letter indicates a marginal effect. O: overlap > no overlap; V, variable > similar; S, similar > variable. An asterisk (*) indicates the presence of an interaction with overlap. A plus sign (+) indicates a marginal interaction with the moderator (see Supplement S2 in supplementary material).
The overlap condition was much lower. Clearly, there is no cost to training overlapping items simultaneously, and students may learn more.

Evidence for the variability effect was mixed. At posttest, the variability benefit was significant by items. By participants, it was not significant with raw scores but was marginally significant with transformed scores. This suggests a small learning advantage for variable consonants, although this is weaker evidence than in prior studies. At retention, the effect was significant but reversed and was restricted to the high overlap condition.

Limitations

Several limitations are worth noting. Many are unavoidable consequences of the structure of English and the items available for first graders. First, overlap was manipulated by training students on distinct sets of English regularities. Thus, we cannot eliminate the possibility that effects were driven by the particular GPC regularities selected for each condition. Pilot data with a large sample (see Supplement S1) do not suggest numerically large differences for four of the six GPC regularities that were tested, and one regularity was shared between conditions. Moreover, items were balanced on frequency and neighborhood density between conditions. Thus, it is unlikely that differences in the inherent difficulty of the items or the GPC regularities were a large factor.

Second, we tested only a small subset of English GPCs. Digraphs were selected because they offer a clear way to instantiate and manipulate overlap. Although there is no reason to believe that these findings are not representative, it is important to replicate them with other classes of GPC regularities to test their general applicability.

Third, it is also possible that these factors work differently or are less critical for other languages that are more orthographically transparent (Nigro, Jiménez-Fernández, Simpson, & Defior, 2015). This should be investigated in future research. However, these factors are likely to be relevant to these languages. For example, in German—a transparent language—the vowel E can be used in isolation (as in ENTE [DUCK]) but can also be combined with U to create the new digraph EU (as in EULE [OWL]). Thus, even in a language with no quasi-regularity, children must still confront overlap (and variability may still be important in isolating the even more regular GPCs). In addition, our study used only the most regular pronunciation of each digraph. Thus, it is reasonable that these findings would generalize to orthographically transparent languages.

Finally, we note that all of our testing used closed set tasks in which children selected the correct choice from a range of options. This contrasts with open set tasks like oral reading that are more standard in classroom and assessment work. Open set responding was not possible with our procedures, and it remains to be seen whether overlap benefits these tasks. However, in other work with older children, we have found strong correlations ($r > .60$) between closed set tasks and oral reading, suggesting that these results would likely hold (Roembke, Hazeltine, Reed, & McMurray, 2019).

Consonant variability

The effect of variability was small (Table 7). At immediate posttest (the closest comparison with Apfelbaum, Hazeltine, & McMurray, 2013), effects were consistent with a variability benefit, but statistical evidence was not robust, significant by item, and marginal by participant with logit transformed scores. These statistically weak effects were surprising given (Apfelbaum, Hazeltine, & McMurray, 2013) results, and evidence from other areas of learning that variability in irrelevant elements leads to robust learning benefits (e.g., Gómez, 2002; Rost & McMurray, 2009).

Several differences between (Apfelbaum, Hazeltine, & McMurray, 2013) and the current study could explain this discrepancy. First is statistical power: (Apfelbaum, Hazeltine, & McMurray, 2013) split their similarly sized sample along only one factor (variability), not two factors (variability and overlap). This may explain why some by-participant analyses were not significant, but by-item analyses were (because variability was manipulated between participants but within items). However, it does not account for the relatively small magnitude of the effect.

Second, our manipulation of variability may have been inadvertently weaker than that of (Apfelbaum, Hazeltine, & McMurray, 2013). Similar items in that study shared the full consonant
frame with 2.40 other items (variable = 0.20) compared with only 1.99 here (variable = 0.31). To ensure overlap between vowels, we used vowels from a small pool of appropriate words, affording less flexibility to maximize variability. Thus, the smaller difference between the variability conditions may have led to a smaller effect.

Third, the (Apfelbaum, Hazeltine, & McMurray, 2013) training regime blocked trials by vowels, whereas we interleaved all vowels during training. The decision to interleave was motivated by findings that interleaving items may maximize learning (Carvalho & Goldstone, 2014; Vlach, Sandhofer, & Kornell, 2008). However, fully interleaved GPC regularities might have helped to eliminate spurious associations between consonant graphemes and vowel pronunciations to isolate the respective vowels without requiring consonant variability.

This was investigated in a follow-up study by (McMurray, Roembke, & Hazeltine, 2019), which simultaneously manipulated consonant variability and training blocked by vowels. They showed a variability benefit when blocking highlighted the overlapping strings but showed no variability benefit for interleaved training. These findings offer a unifying explanation as to why the observed variability effect was smaller than in (Apfelbaum, Hazeltine, & McMurray, 2013) and suggest that how vowels are blocked affects whether variability benefits learning or not. These types of interactions suggest that children’s statistical learning is sensitive to other factors in the input (particularly factors local to a sequence of training) and should not be conceptualized as simple co-occurrence counting across the entire training set.

Pretest
Performance at pretest was lower in the high overlap condition than in the low overlap condition. This was likely due to two factors. First, on a trial-to-trial basis, simply alternating between similar GPC strings could add difficulty, analogous to phenomena like cumulative semantic interference (e.g., Oppenheim, Dell, & Schwartz, 2010). Second, items or letters that were assigned to the same task served as each other’s foils. Consequently, foil items were more likely to overlap with the target item in the high overlap condition than in the low overlap condition. Thus, accurately selecting the target in the high overlap condition required adequate understanding of the relevant GPC regularities but also the ability to inhibit competitors in the moment (see McMurray, Horst, & Samuelson, 2012, for an analogy in word learning).

However, the pretest effect of overlap was observed only for digraphs. This is likely because digraphs are more difficult. Monographs are typically introduced first in reading curricula and, thus, would have been practiced the most. This early—and often isolated—exposure may lead to more robust mappings that are more resilient to interference from overlapping strings later. It is also possible that digraphs are not perceived as a unit: The two letters, therefore, may compete with a larger number of foils on each trial than a monograph with only one letter. This could explain why accuracy for digraph vowels was particularly low in the high overlap condition, where foils were more likely to be closer competitors.

Learning and retention
Even though the high overlap condition was harder at pretest, there were larger learning gains with high overlap training than with low overlap training. This difference was particularly pronounced at retention, where no retention was observed after low overlap training. Overlapping vowels, which introduced difficulties during pretest (and presumably during training), may help in the long run.

3 When teaching a small set of vowels, the only way to avoid this foil overlap would be to introduce new vowel strings that appear only as foils. These, of course, could easily be eliminated because they are never correct, making the task unintentionally easy.
Theoretical accounts

The naïve associative account does not predict these findings. The fact that individual letters must be mapped to more than one sound was expected to impede learning, but this was not observed. So why was vowel overlap helpful for learning and retention?

Connectionist instantiations of statistical learning (McClelland, 2013) show that new materials can be more rapidly and robustly integrated into an existing network if they share schematic structure with already learned items than if they are inconsistent with what is known. In this case, the learning system is not characterized as isolated associations between elements such as letters and sounds; rather, it is characterized as associations among sounds, letters, and (most important) complex intermediate representations. In situations where new material can be better integrated because of similarity to existing items, learning is more efficient. This could also explain the similarity benefit at retention; similar consonants may be an additional way to build schemas (although it is unclear why this would not be observed at immediate posttest).

At a broader level, our work offers clear evidence that statistical learning in reading must be considered as part of a complex network (e.g., Seidenberg, 2005): Students are not learning an entirely novel set of associations from scratch. Rather, new associations are embedded in a rich network of existing knowledge. When we consider statistical or associative accounts in this light—as part of a developmental process, not a laboratory learning exercise—the nature of learning may change. These interactions might also relate to why it has sometimes been difficult to find a relationship between domain-general statistical learning and reading (e.g., Schmalz et al., 2019). That is, statistical learning in reading development may be more than just a matter of whether a child is a good statistical learner. Moreover, learning depends on more than the cumulative statistics to which the child is exposed. Local groupings of items or GPCs likely matter, and this can vary markedly across curricula and the text corpora to which a child is exposed. This may explain why there is no straightforward relationship between, for example, sequence learning and reading ability.

Moderators

Supplement S2 presents extensive analyses on potential factors that may moderate these effects at the level of task, item, and participant. For the most part, few of these were significant. However, there were notable exceptions.

First, the effect of overlap did not differ in generalization tasks and items. Although caution is warranted when interpreting null effects, it suggests that overlap not only improved learning but also that learning gains are not fragile or specific to the items/tasks that were used. Critically, it suggests that the bulk of learning is at the level of generalizable regularities and is not item specific.

Second, at the item level, the word/nonword distinction moderated the effect of overlap on retention, with a larger effect of overlap for words than for nonwords. This could speak to a schema-based account in that the effect of overlap helps the child to not only learn the regularities but also store words in an orthographically organized lexicon.

Finally, at the participant level, we found that reading ability was a marginal moderator of the overlap effect. Here, stronger readers showed a larger effect of overlap. This may also support schema-based learning because these readers would have had a larger and more robust set of GPC schemas at the onset of the experiment. Unfortunately, we did not have access to other person-level factors. It would be particularly important to examine variation in statistical learning (Arciuli & Simpson, 2012; Elleman et al., 2019) or other types of learning. However, there was little variability in learning gains (Tables 5 and 6). This suggests that in highly supervised classroom-style learning, there may be less variability than in unsupervised or naturalistic learning settings. At the broadest level, the lack of strong moderators suggests that over the long haul most students, independent of their individual differences, would benefit similarly from orthographic overlap.

Implications for teaching

Our goal was to identify principles by which materials can be structured to improve learning, considering implications both from statistical learning and from typical reading instruction. This work
was motivated by the fact that current statistical learning theories make competing predictions for what should benefit learning when we consider the real world of grouping GPC regularities for instruction. Rather than testing a full curriculum, we conducted a rapid “field test” of these principles. This type of study may offer an important addition to our understanding of instructional practice given the dearth of work on the microstructure of curricula (e.g., the specific groupings of items or regularities in a curriculum; McMurray, Roembke, & Hazeltine, 2019).

In this light, our findings suggest why vowels may be particularly hard for early readers (e.g., Treiman, 1993). Although clearly part of the difficulty lies in the fact that most digraph rules have several alternative pronunciations (e.g., EA can be /i/ in MEAT, /e/ in THREAT, and /eɪ/ in STEAK), our work indicates that the overlap among digraphs may also contribute to their in-the-moment difficulty (e.g., at pretest).

Despite this, our study suggests that it might not be effective to avoid overlap in reading curricula that target subsets of the GPC regularities. At immediate posttest, high overlap training led to learning gains that were at least as large as, if not larger than, low overlap training, and retention was more robust. This principle can easily be incorporated into curricula by reworking the way in which GPC strings are introduced, although clearly such curricula require randomized control testing for efficacy.

One potential caveat may be that participants in this study already had some knowledge of the GPC regularities on which they were trained (mean pretest accuracy >55% correct). In addition, the interaction between overlap and reading ability was marginally significant at posttest (see Supplement S2), with lower performing children having a tendency to benefit less from high overlap than highly performing children. This could indicate that high overlap groupings might not be effective from the beginning of reading instruction but will be useful later. Future research should investigate in more detail to what extent individual differences in reading ability and statistical learning affect the success of curricular manipulations at the item and grouping levels.

More broadly, the findings highlight the power of systematically manipulating how items are grouped during training. Computer-based training may be helpful in enacting these principles because they can efficiently instantiate controlled manipulations of items, stress different forms of overlap at different times, and provide immediate feedback (which might be important for some forms of implicit learning; Maddox, Ashby, & Bohil, 2003). Although this cannot replace real teachers, computer-based training may supplement the curriculum and traditional school media (e.g., worksheets) by leveraging learning principles for better outcomes. Future research might investigate the extent to which the computer interface facilitated students’ learning: For example, was the provided feedback during training necessary for children’s learning?

Most children completed the training portion of the experiment within 3 to 5 days of daily 20-min sessions. Although improvements were impressive at posttest, only the high overlap group’s performance remained above pretest at retention. This suggests two things: First, relatively little time may be needed to harness the positive effects of item-level manipulations (e.g., one could imagine a child completing a couple of tasks each day after school). Second, these effects are longer-lasting (i.e., at least 1–2 weeks) only if the statistics are well tuned to the child’s needs.

If one takes statistical learning seriously as a mechanism of reading acquisition, it must be considered in the context of how reading instruction is structured. Our findings offer clear recommendations for how curricula could be designed, both on the item level and at the level of grouping regularities that are taught simultaneously, while providing evidence for a network view of statistical learning.

Acknowledgments

The authors thank Steph Wilson in the West Des Moines School District, Iowa, for valuable assistance with the schools; Alicia Emanuel, Shellie Kreps, and Tara Wirta for assistance with data collection; Eric Soride for help in managing and implementing the technology; and Carolyn Brown and Jerry Zimmerman of Foundations in Learning for support and discussions throughout the project. We also thank Ariel Aloe for consultation on the statistical approach. This study was funded by a National Science Foundation (NSF) grant (BCS 1330318) awarded to E.H. and B.M.
Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2019.104731.

References


