Field Tests of Learning Principles to Support Pedagogy: Overlap and Variability Jointly Affect Sound/Letter Acquisition in First Graders

Bob McMurray, Tanja C. Roembke & Eliot Hazeltine


To link to this article: https://doi.org/10.1080/15248372.2018.1526176
Field Tests of Learning Principles to Support Pedagogy:
Overlap and Variability Jointly Affect Sound/Letter Acquisition
in First Graders

Bob McMurray, Tanja C. Roembke, and Eliot Hazeltine
University of Iowa

ABSTRACT
Many details in reading curricula (e.g., the order of materials) have
analogous in laboratory studies of learning (e.g., blocking/interleaving).
Principles of learning from cognitive science could be used to struc-
ture these materials to optimize learning, but they are not commonly
applied. Recent work bridges this gap by “field testing” such prin-
ciples: Rather than testing whole curricula, these studies teach students
a small set of sound-spelling-regularities over a week via an Internet-
delivered program. Training instantiates principles from cognitive
science to test their application to vowel acquisition, a critical part
of reading. The current study follows up on two prior studies, which
found differing effects of consonant variability for learning vowels.
In addition to investigating this discrepancy, this study examined a new
principle: blocking/interleaving. Although interleaved training is typi-
cally beneficial, this is difficult to apply in reading where there are
many regularities. We compared a fully interleaved regime (six
vowels) to two blocked regimes teaching two vowels on each
block. Blocked conditions differed on whether vowels overlapped
(EA with OA) or not (EA with OU). Blocking was crossed with con-
sonant variability. Four hundred seventeen first graders were pre-
tested on six vowels, and underwent three to five days of training,
followed by a posttest and retention test. Blocking had little effect.
However, there was a variability benefit when overlapping vowel
strings were blocked together, and no effect of variability for inter-
leaved training. Thus, benefits may only be observed if blocking
highlights contrast between regularities. When applied to real-world
skills, learning principles from cognitive science may interact in com-
plex ways.

A pressing problem in education is reading: Although as many as 16% of children can be
diagnosed with dyslexia (Pennington & Bishop, 2009) (itself a large number), 60% of
children do not meet grade-level proficiency (NCES, 2013). Further intensifying this
problem, reading is highly stable, and many children do not “grow out” of their deficits
(Svensson & Jacobson, 2006; Wright, Fields, & Newman, 1996; however, see Shaywitz,
Escobar, Shaywitz, Fletcher, & Makuch, 1992). Reading is essential for other academic
subjects, a foundation for job skills, and important for language and cognitive develop-
ment more broadly. Thus, it is not surprising that poor reading in elementary school
predicts poor life outcomes (Blomberg, Bales, Mann, Piquero, & Berk, 2011; Catts, Fey, Zhang, & Tomblin, 2001; Fall & Roberts, 2012; Reed & Wexler, 2014; Wagner, Kutash, Duchnowski, Epstein, & Sumi, 2005).

Education research has responded: There is substantial work on best pedagogical content, such as debates over whole language or phonics (Ehri, Nunes, Stahl, & Willows, 2001; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Liberman & Liberman, 1990; Torgesen et al., 2001), as well as on component skills such as phoneme awareness (Bhattacharya & Ehri, 2004; Ehri et al., 2001). There are many clinical trials of curricula or remediations (e.g., Santa & Høien, 1999; Svensson & Jacobson, 2006). Correlational and longitudinal studies have identified the constellation of skills that are impaired (or preserved) in struggling readers (e.g., Catts, Fey, Zhang, & Tomblin, 1999; Catts et al., 2001; Catts, Gillispie, Leonard, Kail, & Miller, 2002; Cutting & Scarborough, 2006).

Many curricula and remediations—though scientifically tested—are developed on the basis of educational history and best practice. Similarly, the outcomes measured by correlational studies are typically the outcomes targeted by curricula and state standards, outcomes such as comprehension, fluency, or decoding. In this paradigm, insights from cognitive and developmental science on the nature of learning have played only an indirect role in shaping reading practice. The gap is likely the consequence of two factors: First, the cognitive science of learning is often studied in the lab and in domains that appear to have little pedagogical relevance. Second, it can be extremely challenging to scale from basic science that is often designed to test a single theoretical issue in the lab to a complex real-world phenomenon where multiple factors may be operative. In this article, we describe a complementary approach—field testing—which attempts to bridge this gap. Our work illustrates how basic principles of learning from cognitive science can be applied to real-world educational problems, and how we can gain useful theoretical insight about real-world problems with something short of a clinical trial of a complete curriculum. We also demonstrate the limits of applying basic theory such as statistical learning to the real world, which motivates the need for such field testing. We illustrate this approach with a new field test of principles that may support children’s learning of early reading skills.

Field testing principles from cognitive science in reading

Cognitive science has long studied the basic mechanisms of word recognition and learning, understanding which could shape curricula and teaching. For example, any given reading curriculum must make choices such as what letters and sounds are taught together, the structure of the sets of words used in spelling lists, or the ordering of material. These decisions have close analogues in laboratory studies of statistical and motor skill learning, investigating whether material should be blocked or interleaved (Carvalho & Goldstone, 2014, 2015), the spacing of practice across multiple sessions (Kornell & Bjork, 2008), or the role of variability in irrelevant factors (Gómez, 2002; Logan, Lively, & Pisoni, 1991; Rost & McMurray, 2010). Such studies identify principles that could be used to structure curricula and practice to promote learning (Dempster, 1988; Rohrer & Pashler, 2010; Willingham, 2002). However, the application of such principles to reading pedagogy has been sparing at best (however, see Kellman, Massey, & Son, 2010; McDaniel, Anderson, Derbish, & Morrisette, 2007; Rohrer, Dedrick, & Stershic, 2015; for examples from math and science). Instead, most education research
(and much of the funding) has been devoted to developing and testing complete curricula and/or interventions.

One challenge in applying learning principles from cognitive science to reading is the complexity of learning in real-world settings. Learning principles such as the role of spacing are often identified with a small set of highly controlled materials learned in a laboratory environment (e.g., Kornell & Bjork, 2008). However, real materials cannot be as easily controlled: For example, there are constraints on which words are available to train a spelling/sound regularity, or children may have partial reading knowledge from exposure outside of class. Moreover, in the laboratory, learning paradigms can be highly constrained, such as learning to classify stimuli into two categories. However, in the real world, the task can be enormous—in reading, 26 letters must be mapped to 44 phonemes, and the same knowledge may be needed for multiple purposes (e.g., children must learn sound/spelling regularities for both reading and spelling). Thus, it is not a given that a principle can be translated directly from the laboratory to the classroom, and layers of complexity can impede this translation.

In the past five years, our research group has attempted to fill this gap. Our goal is to “field test” learning principles in the context of reading, to ask if a given learning principle works in the context of real (often unbalanced and imperfect) materials in a school setting. Along the way, we have identified places where cognitive theory is too limited to offer clear insight. This is where field tests can help. The field testing approach should not be confused with effectiveness research of an entire curriculum (the norm in education research). Rather, our goal is to capture a small aspect of reading and test it in a short-term controlled way to identify principles that could be applied more broadly, or to resolve conflicting predictions from cognitive theory. This field testing step could also be immediately useful for developing the micro-structure of a curriculum: word lists, the order in which material is presented, and the way practice unfolds. Long term, it may indicate alternative models of instruction and assessment.

**Grapheme-phoneme correspondences**

Our work focuses on the problem of learning the mappings between sounds and letters, so-called grapheme-phoneme correspondences (GPCs). For instance, in English, EA usually makes the /i/ sound (as in MEAT and DREAM), while E makes the /ɛ/ sound (BED, GEM). Acquiring GPC regularities is a large component of phonics curricula in first and second grade.

In the cognitive science of reading, one of the longest running debates is the extent to which GPCs are explicitly represented as rules. Information processing models propose an explicit role for GPC rules (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). This is consistent with the way GPCs are typically taught: GPC rules are explicitly explained to children, and curricula often are organized around sets of rules (e.g., children learn “short vowels” [e.g., A, E] for a few weeks and then “long vowels” [e.g., O_E, A_E]).

In contrast, connectionist models suggest that learners can achieve the same functions by encoding probabilistic or statistical relationships among words, letters, and sounds (Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). Part of the motivation for this approach is that rules do not adequately describe GPC mappings. For example, whereas EA is typically pronounced
/i/ (e.g., MEAT, TEAM), it can also make /ɛ/ as in THREAT, or /eɪ/ as in STEAK. Moreover, the same sound (/i/) can be made by multiple letter strings such as E (BE), EE (BEE) and E_E (CEDE). Models that explicitly represent grapheme-phoneme-mappings require more than 1,200 rules to adequately represent the English system, and they still require exceptions (Coltheart et al., 2001).

Thus, although GPCs are described and taught as rules, they are at best quasi-regularities (Seidenberg, 2005). The notion that GPCs are encoded probabilistically and via a procedural mechanism is supported by a large body of work examining the role of similarity structure in predicting skilled word reading (e.g., for highlights see, Glushko, 1979; Jared, McRae, & Seidenberg, 1990; Zevin & Seidenberg, 2006), as well as developmental work linking individual differences in reading outcomes to statistical learning ability (Arciuli & Simpson, 2011; Spencer, Kaschak, Jones, & Lonigan, 2015).

The quasi-regular nature of the sound/spelling system of English may, at first blush, suggest the futility of teaching explicit rules. However, pedagogically, the utility of teaching GPC regularities is hard to deny: They are easy to explain and offer children meta-linguistic or explicit strategies for “sounding out” a word in the absence of more automatic or implicit processes. Moreover, to the extent that instruction necessarily must focus on a subset of the sound/spelling system at any given time, explicit rules offer convenient ways to structure a curriculum. Perhaps most importantly, for emerging readers, many (if not most) written words they encounter will be novel. Thus, a learning approach focused on learning individual items (sight word methods), or implicit abstraction across items (whole language) may not be successful for all readers. In contrast, rules permit generalization and inference over words that have not been encountered in text before. This benefit of rules is perhaps one of the reasons for the success of phonics (Ehri et al., 2001).

The critical question then is how one can merge the practicality of rules while respecting the statistical or probabilistic nature of the reading system. This issue led us to a series of field tests (Apfelbaum, Hazeltine, & McMurray, 2013; Roembke, Freedberg, Hazeltine, & McMurray, submitted) including a new one reported here. These field tests investigate the kinds of learning paradigms that may help bridge the divide between pedagogical approaches favoring rules, and statistical learning accounts from cognitive science. In particular, this broader line of work was motivated to identify and test ideas about learning that achieve two broader goals relevant to pedagogy. First, we sought to identify and test learning principles that specifically enhance generalization and abstraction, enabling children to use the skills acquired in a short field test to read many (untrained) items. Second, a major focus in reading is the development of automatic word recognition, as this is an important predictor of later outcomes such as fluency (Compton, 2003; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Kirby, Parrila, & Pfeiffer, 2003; LaBerge & Samuels, 1974; Roembke, Reed, Hazeltine, & McMurray, in press). Assessing automaticity is not the goal of the present study; however, when considering the kinds of learning principles that may be relevant to reading, the emphasis in the literature on automaticity suggests we may gain insight from work on the acquisition of motor skills (procedural learning) where automaticity is also an important focus.

A principle of learning: Irrelevant variability

Our initial investigation focused on the role of irrelevant variability. In many domains, variability in specifically irrelevant aspects of the items or task promotes greater...
abstraction and generalization (Braithwaite & Goldstone, 2015; Magill & Hall, 1990; for reviews). For example, Gómez (2002) taught people an artificial grammar in which the first word in a sequence perfectly predicted the third. She found that the number of possible intervening words in the middle position affected learning—when the middle item was highly variable, learning the abstract rule was easier. Similarly, for adults learning novel phonetic contrasts (e.g., Japanese listeners learning to discriminate /l/ and /r/) and infants learning a contrast for the first time, acquisition is better when there is variation in the talker’s voice, which is not relevant to the phonemic discrimination (Lively, Logan, & Pisoni, 1993; Rost & McMurray, 2009, 2010). Similar effects have been observed in skill learning and implicit learning (e.g., Huet et al., 2011; Kerr & Booth, 1978; Rowe & Craske, 1998). Thus, this principle satisfies the functional demands of learning a rule, and the intuition that such learning may be procedural.

Theoretically, variability may benefit learning by helping to highlight relevant information. Variable dimensions or elements of a stimulus or task get down-weighted, allowing learners to focus on the invariant relevant dimensions or elements. This leads to more abstract representations that are not contextually bound to the irrelevant elements (Apfelbaum & McMurray, 2011). In principle, any given dimension or element may be both relevant and irrelevant: for r/l discrimination, talker voice is irrelevant, but for identifying the talker, this information is clearly relevant. However, in an instructional context, relevance can be determined by instructional goals (e.g., what one is trying to teach), making this principle particularly relevant for reading.

In learning GPCs, the irrelevant variability principle makes clear predictions. Consider the GPC regularities for vowels. Typically, such regularities are taught (in part) with “word families,” highly similar sets of words instantiating a GPC regularity. For instance, to teach the EA→/i/ rule, one might use BEAD, READ, LEAD, whereas to teach the AI→/eɪ/ rule, one might use RAID, LAID, MAID. The close similarity in consonant structure was thought to promote deeper analysis and/or inference about how the vowels work. However, if the child does not isolate the vowel from the rest of the word, the inclusion of similar consonants in the training set could also create a situation in which the irrelevant D “comes along for the ride.” That is, she or he learns that EAD makes the /i/ sound and OAD makes the /oʊ/ sound, but is unclear about EAT or OAM. This leads to a narrower contextually dependent mapping. Such training further creates a situation in which the irrelevant cue (D) is equally predictive of both pronunciations, creating ambiguity. In contrast, the use of consonant frames that are more variable (e.g., TEAL, READ, MEAT, LOAF, SOAK, FOAM) could prevent the final consonant from ever forming a strong association with the pronunciation, leading to a more context invariant mapping.

Our initial goal was to test this insight. We therefore conducted a field test of the variability principle by teaching children a small number of rules for vowels over a limited time. As a field test, this was not intended as a test of a complete curriculum, nor was it even sufficient time for children to master the targeted GPC regularities. However, it was a useful model for asking if the principle of irrelevant variability could benefit first graders learning reading skills over realistic sets of items.

A critical challenge was how to deliver training. Three insights led us to a somewhat unorthodox approach. First, traditional education research (e.g., efficacy and effectiveness studies) has long been challenged by issues of fidelity of implementation (O’Donnell, 2008)—whether teachers or experimenters can faithfully implement the intended instruction. Second,
a proper instantiation of variability requires carefully manipulating a large set of items. Third, both the quasi-regular nature of the GPC system, and seminal work on irrelevant variability imply a procedural or implicit mode of learning (Kerr & Booth, 1978) that benefits from rapid feedback after each response (Ashby, Maddox, & Bohil, 2002; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005).

A computer-based training procedure could address all three issues. Computers can maintain fidelity, deliver many items, and provide immediate feedback in a way that cannot always be done with live instructors. Thus, we partnered with an education technology company, Foundations in Learning, Inc (Coralville, IA). Foundations in Learning builds and markets Internet-delivered reading interventions and assessments, and they allowed us to reconfigure their technology platform to deliver short-term learning and testing experiences directly to schools.

Using this platform, Apfelbaum et al. (2013) conducted an initial test of the variability hypothesis. They taught 264 first graders six GPC regularities: three short vowels (A, O, I) and three digraphs (AI, EA, and OA). Training and testing was implemented in a variety of tasks to maintain interest and to force students to acquire the rules, not just to optimize task performance. For example, in the Find the Word task, children heard a word, and selected its written form from among eight choices; in Fill in the Blank, children selected the missing vowel for a word. Children started with two days of pretest, then underwent 288 training trials (across three to five days), followed by a posttest. For half the children, training items were words and nonwords with highly variable consonant frames, whereas for the other half consonant frames were highly similar. Pre- and posttest were identical between groups. The results showed a benefit of consonant variability. Children in the variable group performed better than those in the similar group, and this benefit generalized to words and to tasks that were not used in training.

Broadly speaking, these results clearly fit into a framework of upweighting relevant elements (vowel letters) and downweighting irrelevant ones (consonants). However, what kind of mechanisms can account for such results? One possibility is what Roembke et al. (submitted) term a naïve associative account (see, Apfelbaum & McMurray, 2011). Under this view, learners link all the letters of a word directly to its pronunciation. Here, if irrelevant consonants are constant across trials, they also (erroneously) participate in these associations, impairing learning. This is similar to learning operations posited in the “visual cue” or “paired associate” stage of reading in classic approaches (Ehri, 1995; Gough & Juel, 1991; Juel & Roper-Schneider, 1985). Variability blocks these erroneous associations because individual consonants are not repeated sufficiently to form a strong association with the vowel pronunciation.

**From principles to curricula**

The longer-term goal of our research program is to marry principles such as irrelevant variability to larger issues in pedagogy. Statistical learning more puts the emphasis on extracting regularities across individual items; indeed, at some level the optimal set of items would be one that encompasses the statistics of the entire language. The irrelevant variability principle is consistent with this item-based approach and offers ideas of how to structure items to highlight the best statistics for a targeted instruction of a specific GPC.
However, statistical learning also has its limits, some of which were not apparent until we considered how reading is actually taught. A critical part of our project has been long-term engagement with the education community over a series of informal discussions with teachers, a workshop with teachers and community members, and conversations with the curriculum developers at Foundations in Learning. These interactions highlighted the fact that statistical learning does not offer clear insight into an additional issue faced by teachers and curriculum developers—which GPC regularities to teach together. Our discussions revealed that scale is a critical issue in phonics learning in the real world. Although ideas such as variability and the importance of interleaved learning (Carvalho & Goldstone, 2014; Kornell & Bjork, 2008; Rohrer et al., 2015) suggest the benefit of teaching many items simultaneously, this approach may not be as effective in the classroom, as there are too many GPC regularities, each of which can be instantiated by hundreds of words. As a result, most curricula group a small number of conceptually related regularities together (e.g., the short vowels, the long vowels). However, statistical learning (and connectionist thinking), with its emphasis on statistics over individual items (words), does not offer clear predictions.

We identified two competing predictions. First, the naïve associative account suggested that simultaneous instruction of GPC regularities that share a letter (e.g., EA, OA, AI) should be more difficult than non-overlapping ones (EE, OU, AI). This is because when teaching overlapping regularities, the same letter must be linked to two pronunciations. Moreover, given the enhanced need to isolate the vowel when GPCs overlap, we predicted a stronger effect of irrelevant consonant variability. In contrast, connectionist instantiations of statistical learning suggest that when learning a set of mappings, novel mappings that are closer to mappings that have been internalized already may be learned more easily, as they can be better integrated into a schema (McClelland, 2013). This predicts better learning for overlapping materials. It is not clear whether such networks will also exhibit the variability effect. As these competing predictions highlight, when we considered the more difficult problem of grouping at the level of regularities (not items), statistical learning as a broader theoretical framing was limited.

Our second field test thus simultaneously manipulated overlap and variability (Roembke et al., submitted) to address the broader issue of grouping GPCs. First graders (N = 277) were randomly assigned to four groups: high and low consonant variability, crossed with overlapping or non-overlapping vowels. They then underwent a similar procedure of two days of pretest, training and a posttest, with an additional retention test one–tp two weeks later.

Results were unexpected. First, although overlapping vowels led to much poorer performance at pretest (consistent with the naïve associative account), by the end of training, performance was equal between the two groups, and one week later, only the students learning overlapping vowels showed retention. Students learned more with the overlapping vowels. This latter finding is not consistent with the naïve associative account. Rather, it is more in line with the schema-based account in which learning is benefited when learners must embed the new GPC regularities in a network of similar ones (a schema) (McClelland, 2013). Second, we found a markedly reduced effect of variability that was only significant in some analyses, and not observed at retention. This was surprising given the close match between the training/testing tasks, and the materials used in our first field test (Apfelbaum et al., 2013). It suggests that the application of learning principles to reading curricula may depend on the micro-structure of training, such as the groupings of GPC regularities.
Interleaved and blocked presentation

While mulling over these puzzling findings, we had an insightful conversation with a teacher. She was discussing her thrice-annual administration of an oral reading fluency exam that gauges how quickly children can read short passages in order to monitor their progress. She noted that at the first assessment children seemed to do well, but then by the second assessment many of their scores drop. What was changing? We realized that the second round of assessment occurred when the teaching unit on short vowels (CAT, BET, HID) had ended and the one on long vowels (MATE, SIDE, BANE) had begun. The students' testing pattern thus may be explained by retroactive interference: New information “overwrites” previously known material. Of course, a long-known solution to retroactive interference is to interleave, rather than block, materials during training (McCloskey & Cohen, 1989; Mirman & Spivey, 2001). Thus, interleaving seemed like a useful potential principle for field testing.

At the same time, it occurred to us that the failure to demonstrate a robust variability benefit in our second field test (Roembke et al., submitted) may also be an effect of interleaving. Apfelbaum et al. (2013) trained GPC regularities in a partially blocked format: Children were trained on three short vowel GPCs (e.g., CAT) for 96 trials, then three digraphs (e.g., BEAK) for 96 trials, and then underwent a mixed block. This blocked design was chosen to match students’ school experience (where GPCs are typically blocked), and to ease the students into the training. In contrast, Roembke et al. (submitted) trained all six GPCs in an interleaved format. Could the interleaving be moderating the variability benefit?

To determine how interleaving affects learning, and to address this hypothesis about variability, the current study crossed a blocked/interleaved training format with consonant variability. This was a situation in which neither statistical learning nor connectionism offer clear predictions. However, it is nonetheless an important empirical question, as it speaks to two issues that could frame curriculum design: how to group GPCs for training, and which sets of items will be most effective for teaching them. Moreover, we hypothesize based on our prior work that we should observe a variability benefit with blocked training but no benefit for interleaved.

During study design, we encountered a second barrier in translating cognitive science to the complexity of real reading: A traditional blocked design would train children on each GPC regularity in isolation. This works fine in the context of a typical “study and test” design, or in the context of a simple motor skill. However, in this case, the tasks may become trivial: For instance, consider the Fill in the Blank task in which the participant sees a word with a missing vowel, hears the word and must select the vowel. In traditionally blocked training, all of the words would be EA words. Here, the same vowel string would always be the correct response, and students could unthinkingly select the same vowel over and over again. Or, consider Find the Word in which the student hears a word (beak) and selects the matching written form. If all response options include the same vowel (with different consonants; e.g., BEAK, LEAF, MEAT), a task strategy in Find the Word may be to find the word with the matching consonants, encouraging students to pay attention to consonants and not vowels.

Thus, GPC regularities would need to be grouped into smaller sets in order to implement a blocked design. But what sort of sets were optimal? Recent theoretical
work on interleaving indicates that the principle of contrast may underlie the benefits of interleaving. That is, students benefit from interleaving different materials (e.g., in this case, GPC regularities) on successive trials because it gives them the opportunity to determine which factors distinguish or contrast them (Carvalho & Goldstone, 2014, 2015; Rohrer et al., 2015). Whereas blocked presentation can help learners identify the defining characteristics for a given GPC, interleaved can promote learning of the features that distinguish categories or GPC regularities. In retrospect, this is consistent with our previous study (Roembke et al., submitted) in which overlap among the letters in a set of GPC regularities led to greater learning. Children need to learn that an A maps onto a different phoneme if it is paired with E (e.g., EA) than when paired with an I (AI). Presenting overlapping vowels within a block offers the opportunity to see the same letters in different configurations on adjacent trials.

If this principle is right, then there may be ways to harness contrast, even within blocked conditions, when blocks contain two GPC rules (to avoid the problems with training only one). The contrast principle predicts that pairing overlapping vowels such as OA and EA (which creates opportunities for contrast) would lead to better learning than blocked training with GPCs with no overlap (EA and OI). Thus, we compared interleaved training to two forms of blocking: contrast and no-contrast.

In the current study, students were trained on six GPC regularities (OI, EA, AI, OO, OU, OA). In the interleaved condition, all six were presented simultaneously in a block of trials. In the blocked/contrast condition, GPCs were presented in pairs within a block of trials, and pairs were constructed to maximize overlap among the vowels (e.g., OI with AI; EA with OA, etc.); in the blocked/no-contrast condition, GPCs were also trained in pairs within a block, but the pairings minimized overlap (OI with EA; AI with OO). These three blocking conditions were crossed with consonant variability/similarity. The design followed our previous studies. There were two days of pretest with a fixed number of trials on each day followed by three to seven days of training. Here, students worked for 20 minutes a day to complete 432 trials (over as many days as it took). Next, there was a posttest, and one to two weeks later, a retention test. All children learned the same GPC regularities, and the pre/postretention tests were identical for all six conditions.

Following Apfelbaum et al. (2013), we operationalize variability in terms of the distribution of letters across items that are irrelevant to the instructionally targeted GPCs (the vowels). Similarly, following Roembke et al. (submitted), we define overlapping items as those that share a letter in the relevant (instructionally targeted) portion of the word. In the real world, a given letter will contribute both to overlap and to variability as each letter participates in multiple mappings; however, in the context of this experiment—where GPCs for vowels are targeted—overlap and variability are orthogonal. It is important to note that unlike Roembke et al. (submitted), overlap was constant in all conditions as the same set of GPC regularities was learned in all conditions. Rather, blocking manipulated whether the contrast between overlapping strings is highlighted on nearby trials (within a block).

Our design was motivated around two empirical questions with both practical value and theoretical import: 1) Is there a role of different blocking/interleaving strategies in improving learning? and 2) is the benefit of variability observed by Apfelbaum et al. (2013) moderated by blocking/interleaving? Given prior empirical findings, we predicted that interleaving would lead to the best learning, but within the two blocked conditions,
contrastive blocking would be superior. This latter prediction is consistent with Carvahlo and Goldstone's (2014, 2015) view that interleaved training benefits learning (when the goal is to identify features that discriminate competing mappings), and it would extend this idea to a much more complex learning task. Better learning in the contrastive blocking scheme is also consistent with a schema-driven learning account: simultaneous training on two GPCs that do not overlap could make it difficult to integrate them into any schema (non-overlapping letter strings such as OI and EA have no inherent similarity, so they must be memorized). As a result, they may be more susceptible to interference when the next block of GPC regularities appears (McClelland, 2013).

With respect to variability, our prior empirical work suggests that there should be no variability effect for interleaved training (Roembke et al., submitted), but we predict a significant benefit for blocked training. If variability is beneficial in both blocked conditions, this would suggest that the general task demands of blocked training (e.g., fewer cognitive resources are required) are responsible for the effect. However, if a variability benefit is only observed in the blocked/contrast condition, this may suggest that variability plays a role in establishing contrast across trials or in integrating new items into schemas.

Beyond these theoretical implications, the empirical findings are meant to capture two critical issues in instructional design, a critical benefit of our field testing approach: what GPCs should be grouped together in a subunit within a curriculum? Moreover, within a subunit, what is the optimal structure of the items for promoting learning?

**Methods**

The study was timed to the spring semester of the first-grade year—a time when students would have some knowledge of these digraphs but would not have mastered them (given an examination of the curriculum in the districts).

**Participants**

Participants were first-grade students from 10 schools within the West Des Moines and Iowa City Community School Districts in Iowa. Participants were recruited over two years (2015 and 2016), but tested at the same point in the school year; four schools participated in both years. All first-grade students without learning disabilities (indicated by an Individualized Education Plan) were invited to participate. Parents of eligible children were first sent a letter detailing the study in combination with a consent form in both English and Spanish. This was returned to the school to obtain documentation of consent.

Five hundred and twelve students initially enrolled. Due to absences, snow days or noncompliance, 349 completed the entire experiment, an additional 42 completed everything but retention and were included in the analysis of posttest (N = 391), and 26 completed retention but not posttest (N = 375). An additional four were excluded as their pretest accuracy was below 20% on eight or more blocks (of 12). Thus, 417 students participated in the analysis in one phase or the other. Of these students, 49 students’ native language was not English. Data of nonnative speakers were analyzed with the rest of the sample, as our prior field tests did not show significant differences in the response to these learning principles in ELL students (nor did we observe any here).
**Design**

Students were taught GPCs for six digraph vowels (Table 1). Each of the GPC regularities was the dominant pronunciation for its letter string. This was determined by computing the proportion of monosyllabic words in Coltheart (1981) that contained the letter string and were pronounced with the intended phoneme.

**Blocks and trial structure**

The entire procedure consisted of two days of pretest, three to seven days of training, two days of posttest, and two days of retention testing one to two weeks after post-test. Each round of testing or trained subsumed multiple small blocks of trials across multiple tasks.

On each block of training or testing, children saw a selection screening containing all available tasks. They could select which task to complete. Each task block included 12 trials (evenly balanced on the GPCs available during that block). Subsequently, they returned to the selection screen to choose a new task from those remaining. Thus, participants had to complete each task before they could repeat a task.

During testing (pre, post, and retention), students were exposed to all six GPC regularities with two words/rule/task. There were 10 tasks, leading to 120 unique trials over two blocks. On day 1, students completed one block and were automatically logged out of the program afterward. They completed the second block on day 2.\(^1\) The three tests were identical to each other and across all groups.

During training, students underwent six blocks of six tasks (four were withheld for testing), with 12 blocks/task within a block. This led to 432 trials. Students used the program for a fixed time (instead of a fixed number of trials as in testing). This number ranged from 20 to 23 minutes across schools, depending on how much time the school was able to allot to this study. Children completed as many tasks during this period as possible, and advanced to subsequent blocks if there was time. After the fixed period, they were automatically logged out of the system and continued where they had left off the following day.

**Experimental conditions**

Children were randomly assigned to conditions after pretest. There were six experimental groups crossing blocking (blocked/contrast, blocked/no-contrast, interleaved) and consonant variability (similar, variable). Further, there was an additional grouping factor that controlled whether a given task appeared in training or was used only during testing.

---

**Table 1. GPC regularities used in this study.**

<table>
<thead>
<tr>
<th>Spelling</th>
<th>Pronunciation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>/i/</td>
<td>MEAN, HEAT</td>
</tr>
<tr>
<td>OA</td>
<td>/o/</td>
<td>COAL, BOAT</td>
</tr>
<tr>
<td>OI</td>
<td>/ɔɪ/</td>
<td>COIN, JOIN</td>
</tr>
<tr>
<td>AI</td>
<td>/ei/</td>
<td>RAIN, SAIL</td>
</tr>
<tr>
<td>OO</td>
<td>/u/</td>
<td>FOOD, HOOP</td>
</tr>
<tr>
<td>OU</td>
<td>/au/</td>
<td>LOUD, HOUR</td>
</tr>
</tbody>
</table>

\(^1\)To make up for absences, a small number of children performed both testing blocks on the same day.
(levels: A, B) for 12 groups. Random assignment was conducted after pretest and assignment balanced pretest score, gender, and ELL status.

**Variability** was manipulated by selecting items for training with similar consonant frames (the testing items were the same in both variability conditions). We strived to maximize (or minimize) the number of complete frames that were shared (e.g., POACH, POUCH, and POOCH), but we also maximized the number of items sharing the onset and/or coda consonant.

**Blocking** was manipulated by controlling which items were presented during a given training block. In the interleaved condition, GPCs were randomly assigned to blocks with items from all six GPCs appearing in each block. In the two blocked conditions, students performed a block of 72 trials in which all of the items were selected from two GPCs. Across blocks, each GPC was used twice (Table 2). In the blocked/contrast condition, each block trained students on pairs of GPCs that maximized overlap by using digraphs that shared a letter (Table 2). In the blocked/no-contrast condition GPCs were paired to avoid overlap. As there were only 12 items per vowel during training, the blocked conditions included some repetition of the same item within a block (but never within the same task).

**Tasks**

Training and testing used a variety of simple tasks (Table 3) selected from our prior studies (Apfelbaum et al., 2013; Roembke et al., submitted). Some tasks did not appear during training (to assess generalization to new tasks). The assignment of tasks to training/generalization was counterbalanced, so that a generalization task for one subject appeared as a trained task for others. Assignment of tasks to condition was equated on the word/nonword status of the task, and on the difficulty of the tasks in prior work.

**Feedback**

During training, students received feedback in two forms. First, if students responded correctly they received positive reinforcement (“good job!”). If they were incorrect, they received negative reinforcement (a buzz followed by “Try again!”), then the answer they had incorrectly chosen was removed and they had to correct their response. On correction trials for relevant tasks, the response screen was also enriched with auditory cues for the responses (children could click a button to hear the sound of each available letter). Subjects had two opportunities to correct their answer before moving on to the next trial. Second, we used a point system to maintain motivation during training. If participants responded correctly on the first attempt, they received 10 points; on the second attempt, they received 5 points, and on the third attempt no points. No feedback on performance was given during testing.

<table>
<thead>
<tr>
<th>Table 2. GPCs used in each block.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>
Table 3. Tasks used in this study. Roles refer to role at test: trained task were encountered during training, Gen (Generalization) appeared at test only.

<table>
<thead>
<tr>
<th>Task</th>
<th>Role</th>
<th>Set A</th>
<th>Set B</th>
<th>Items</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find the word</td>
<td>Set</td>
<td></td>
<td></td>
<td></td>
<td>Hear a word/nonword and select it from among 10 printed options.</td>
</tr>
<tr>
<td>Find the nonword</td>
<td></td>
<td>Train</td>
<td></td>
<td>NW</td>
<td></td>
</tr>
<tr>
<td>Fill in the blank (word)</td>
<td></td>
<td>Train</td>
<td></td>
<td>W</td>
<td>Hear a word/nonword and see a consonant frame. Choose vowel to complete the</td>
</tr>
<tr>
<td>Fill in the blank (nonword)</td>
<td></td>
<td>Train</td>
<td></td>
<td>NW</td>
<td>item from among 10 options.</td>
</tr>
<tr>
<td>Make the word</td>
<td></td>
<td>Gen</td>
<td></td>
<td>W</td>
<td>Hear a word/nonword and choose the letters to spell it from 10 options for</td>
</tr>
<tr>
<td>Make the nonword</td>
<td></td>
<td>Gen</td>
<td></td>
<td>NW</td>
<td>each position (onset, vowel, coda).</td>
</tr>
<tr>
<td>Word family</td>
<td></td>
<td>Train</td>
<td></td>
<td>W</td>
<td>Hear a vowel and coda consonant (e.g. EAT), and find a word/nonword that</td>
</tr>
<tr>
<td>Nonword family</td>
<td></td>
<td>Train</td>
<td></td>
<td>NW</td>
<td>contains those sounds from 10 options.</td>
</tr>
<tr>
<td>Change the word (vowel)</td>
<td></td>
<td>Gen</td>
<td></td>
<td>W</td>
<td>See a word, change it to another (“Change the vowel in cat to make coat”)</td>
</tr>
<tr>
<td>Change the nonword (initial)</td>
<td></td>
<td>Gen</td>
<td></td>
<td>NW</td>
<td>by selecting new vowel from 10 options.</td>
</tr>
</tbody>
</table>

**Items**

Items were individually examined by the research team and by the private sector partners to ensure they were appropriate for first-grade children and correctly instantiated the intended GPC.

**Testing items**

Items used for testing were the same in all conditions and for all testing sessions. For each GPC regularity, we selected six words and six nonwords according to three categories: two items that appeared in training (trained); three items that followed the target GPC, but were untrained (novel, two of these included consonant clusters [e.g., trout, paint] to enhance difficulty); and one item used a GPC that did not appear in training (baseline).

**Training items**

Two lists of words were used for training—one for the variable condition and one for the similar condition (the blocking manipulation concerned trial order not the selection of items). For each GPC regularity, six words and six nonwords were selected for each list. Although most training items followed the simple CVC schema, we included several with consonant digraphs (CH) to increase the number of items available for vowels like OI that generally do not have as many words that are known by first graders.

The two lists differed in the amount of variability among consonant frames. In the similar list, items shared onset or offset consonants or both: Each item shared the complete consonant frame with an average of 2.36 other items (e.g., poach, pooch), and shared its onset consonant with 8.03 other items and its coda with 14.59. In contrast, the variable word list minimized this similarity: Each item shared its full frame with only 0.1 other words. They shared onset consonants with 3.85 items and coda consonants with 7.4 items. As some words needed to be repeated from training to test, this meant that several words had to be used in both the variable and similar list. This (in addition to restrictions of the language) constrained our ability to completely control variability.

The words were also matched across variability conditions on frequency using the Children’s Printed Word Database (Masterson, Stuart, Dixon, & Lovejoy, 2010) and imageability (when available), using (Coltheart, 1981). Some words did not appear in this corporal;
in the first analysis they were treated as a frequency of 0. There was no difference in frequency ($M_{\text{Similar}} = 65.8; M_{\text{Variable}} = 105.5; p = .114$). We also compared groups by excluding those items and also found no difference ($M_{\text{Similar}} = 107.8; M_{\text{Variable}} = 140.7; p = .25$). There was also no difference in imageability ($M_{\text{Similar}} = 492.39; M_{\text{Variable}} = 519.76; p = .284$).

**Stimuli**

All items, training instructions, carrier sentences and phonetic cues were recorded by a phonetically trained female talker, speaking clearly and slowly. Recordings were conducted in a soundproof room using a Kay Elemetrics CSL 4300b (Kay Elemetrics Corp., Lincoln Park, NJ) at 44,100-Hz sampling rate. Items were recorded in a neutral carrier phrase (He said...) to control for prosody and rate. The words were then excised, 100 msec of silence was added, and edited to remove clicks, pops and extraneous articulations.

**Procedures**

Children participated during the school day in large groups in a computer lab or media center. Students were assigned a unique login username and password to track their progress. Items and task instructions were presented over headphones to minimize disruption from other students taking part in the study at the same time. Completed tasks were marked with a checkmark and showed how many points were awarded within that task. Each block of tasks was presented with a new color background.

Children received nonspecific encouragement (e.g., “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 students’ accuracy.

Each task began with auditory instructions. Each trial started with shorter spoken instructions which included the target stimulus. For instance, in the Find the Word task, children were told to “Find the word meat” at the beginning of a trial. This was accompanied by 10 written items on the screen, one of which was the target stimulus. Each trial was completed by choosing a response option with a mouse (see Table 3 for an overview of all tasks).

**Results**

Preliminary analyses showed no differences in the baseline words (which tested untrained GPCs) at pretest or retention; thus they were excluded from all analyses. In all three groups, a clear learning effect was observed between pre and posttest, with some drop off at retention testing (Figure 1). Large effects of blocking were not immediately apparent. However, there is a clear moderation of the effect of consonant variability across conditions, with a benefit for similar consonants in the blocked/no-contrast group, a benefit for variability in the blocked/contrast group and no difference in the interleaved condition. Slight differences in pretest were likely due to attrition (since the groups were balanced by pretest at random assignment); however, a blocking x variability ANOVA on pretest scores revealed no significant differences (all $F < 1$).
To document these results statistically, we used an ANCOVA with accuracy as the dependent variable and pretest as a covariate to examine the effect of blocking and variability (between-subject factors) (see Dugard & Todman, 1995; Van Breukelen, 2006). To maximize power, this ANCOVA included both the posttest and retention, with an additional within-subject factor, test-phase. As some participants were missing one phase or the other (posttest: 6.2%; retention: 10.1%), missing data points were replaced with cell means. Figure 2 shows the marginal means (after accounting for pretest differences) within each condition. We also conducted an item analysis (Clark, 1973) to document that effects generalize across items. As this used a substantially different design (all effects were fully within-item), it is described in Supplement S1, and we incorporated experimentally important (or discrepant results) into the discussion here.

As expected the covariate (pretest) had a strong effect ($F(1,410) = 472.5, p < .0001$). There was a significant main effect of test-phase ($F(1,410) = 23.89, p < .0001$) with an overall reduction in performance between immediate posttest and retention. There was no main effect of either blocking ($F < 1$) or variability ($F < 1$). However, there was a blocking x variability interaction ($F(2,419) = 3.21, p = .041$; see Figure 2a and b), which was also significant by items ($p = .013$). This did not further interact with phase ($F < 1$).

This was followed up in several ways. First, we conducted two 2 × 2 ANCOVAs to try to identify the nature of the interaction. These showed no blocking x variability interaction when we only considered the blocked/no-contrast and interleaved conditions ($F < 1$), but a marginal similarity benefit ($F(1,278) = 3.22, p = .074$), and no main effect of blocking ($F(1,278) = 1.42, p = .24$). In contrast, when we compared the two blocked conditions, we found an interaction ($F(1,265) = 5.03, p = .026$), and no main effects (Fs < 1). Thus, the interaction appears to derive from the distinctiveness of the blocked/contrast condition relative to the other two.

Next, we conducted separate ANCOVAs examining the effect of variability (accounting for pretest) within each blocking condition (Figure 2a). This showed no effect of variability in the blocked/no-contrast condition ($F(1,133) = 1.79, p = .183$); though there was a

---

Figure 1. Accuracy as a function of phase (pretest, posttest, retention) and variability condition for the a) blocked/no-contrast group; b) blocked/contrast group and c) interleaved group. Error bars reflect SEM. For reference, the grand mean across all conditions is plotted with a dotted line.

---

2We also conducted separate ANOVAs on the posttest and retention data using only subjects that had data for that phase and found the same general pattern of results.
significant similarity benefit by items ($p < .0001$). There was a marginal variability benefit in the blocked/contrast condition ($F(1,131) = 3.07$, $p = .082$) that was significant by items ($p = .02$). Finally, there was no effect of variability in the interleaved condition ($F(1,144) = 1.48$, $p = .226$; by items, $F < 1$). Thus, partial evidence exists for a variability benefit in the blocked/contrast condition, but overall the results are more consistent with a crossover pattern.

In addition to the variability x blocking interaction, the primary analysis also found a testing-phase x blocking interaction ($F(2,410) = 4.66$, $p = .010$; Figure 2b). As before, this was evaluated with separate ANCOVAs examining the pairwise blocking comparison within each phase. At posttest, this found a difference between the interleaved and blocked/no-contrast condition ($F(1,278) = 4.46$, $p = .036$), and no difference between the two blocked conditions ($F(1,265) = 1.43$, $p = .23$). However, at retention, there was no difference between either contrast (both $F < 1$).

Finally, the phase x variability ($F < 1$) and three-way interactions ($F < 1$) were not significant.

**Learning across conditions**

The preceding analysis could not determine if there was learning at all (e.g., a difference from pretest). We therefore conducted a series of paired t-tests asking whether each of the six subconditions differed from pretest at both posttest and retention. Here, a significant effect would indicate the presence of learning. These were an exploratory analysis, and we caution the reader not to draw strong claims from any individual comparison, but rather to look at statistical patterns as a whole.

Results are presented in Table 4. We saw significant evidence of learning in virtually all conditions. The exceptions (which accord with the prior analyses) indicated 1) only marginal evidence of learning for the blocked/no-contrast group with variable words; 2) no evidence of retention in the blocked/contrast group with similar words; and 3) no evidence of retention in the interleaved group with variable words. Although we do not
make claims about each condition, this pattern complements the crossover interaction shown in the prior analysis, as here we see that variability shows a benefit with blocked/contrast training, but a cost in blocked/no-contrast or interleaved training.

Low-performing readers

One concern is that many of the children performed fairly well at pretest—this is likely due to the large variability in decoding ability in first grade, coupled with the fact that testing was necessarily spread out throughout the semester. However, this raises the possibility that results could simply reflect variation in performance (e.g., testing) among readers who had already mastered the regularities. Moreover, in educational practice, variability and to a lesser extent interleaving and/or contrast are often avoided (particularly for low-performing readers) to avoid overwhelming children, yet here we see a potential benefit. Thus, it is important to ask if low-performing readers also show these effects which appear to conflict with many educators intuition.

To address both issues, Supplement S2 conducted a median split in which the data were separated by pretest ability. Low- and high-ability readers were analyzed separately using the same ANCOVA as before. In low-performing readers, the crucial variability x blocking interaction was significant \((p = .03)\), and numerically almost twice as large. A significant benefit of variability occurred in the blocked/contrast condition \((p = .037)\). There were no effects in the high-performing group. This shows a clear pattern for the variability benefit and suggests stronger effects in the group that started with the least mastery.

Generalization

Followup analyses (Supplement S3) also examined whether any of these effects were moderated by whether tasks and items were trained or held out for testing only. These analyses showed a generalization decrement in both cases, but no moderation of the variability x blocking interaction. This suggests that the learning benefits (or costs) from the variability and blocking manipulations are based on the abstract encoding of regularities among sounds and letters, not memorizing items, or optimizing task performance. In similar fashion, we examined the effect of item type (word/nonword). Here, there was no decrement in performance for nonwords, supporting the idea that children were learning an abstract rule, not memorizing words, and there was no moderation of the blocking x variability interaction.
Training data

Lastly, we conducted an analysis of the training data, examining accuracy in training as a function of block number. A limitation of this analysis is that GPCs vary in their inherent difficulty (Figure 3a), and different GPCs were assigned to different blocks (which could not be counterbalanced due to technical limitations). Thus, this analysis should be treated as exploratory.

Figure 3b shows accuracy as a function of training block. We see an upward trend in all three blocking groups, but rather large block-to-block fluctuations in the two blocking groups. Some of this may be due to differences between no-contrast and contrast versions of blocking, although some is likely due to the base difficulty of the GPCs that happened to be used on that block. Thus, for each participant, for each block, we computed a relative measure of performance as the difference between performance on that block and their own pretest performance on the GPC regularities used on that block. This is shown in Figure 3c. Again, the interleaved group shows a smooth increase, and block-by-block fluctuations are somewhat reduced. However, we also see little upward trajectory in the blocked/no-contrast group, whereas the blocked/contrast group shows increasing performance.

This was evaluated in an ANCOVA examining relative accuracy as a function of trial block (within-subject), blocking (between), and variability (between) with pretest accuracy as a covariate. We found an overall main effect of blocking ($F(2,410) = 5.05, p = .007$). This was due to relative accuracy being greater in the blocked/contrast group ($M = .14$) than either the blocked/no-contrast ($M = .12; p = .033$) or interleaved group ($M = .11; p = .002$). Performance was expected to be better in the blocked than interleaved conditions (since only two GPCs were relevant on a given trial); however, this cannot account for the difference between the two blocked conditions. There was no effect of variability or a blocking x variability interaction ($F < 1$). We also found an effect of trial block ($F(5, 2050) = 6.21, p < .001$), with better performance late in the experiment. The main effect of trial block did not interact with variability ($F < 1$). However, it did interact with blocking ($F(10,2050) = 15.27, p < .001$), and there was no three-way interaction ($F < 1$). Follow-up ANCOVAs examined performance within each blocking condition with a trend analysis to identify the effect of increasing block. We found significant linear trends in the interleaved ($F(1,144) = 57.79, p < .001$) and blocked/contrast ($F(1,131) = 21.10, p < .001$), but not in the blocked/no-contrast condition ($F(1,133) = 2.42, p = .12$).

Figure 3. a) Accuracy at pretest for each of the six GPC regularities. b) Accuracy over the course of training. Pre and posttest results are shown for reference (horizontal line shows pre-test accuracy). Each block is labeled with the pair of vowels trained on that block (top row: blocked/contrast, bottom: blocked/no-contrast). c) Accuracy on each block minus mean accuracy for the corresponding GPC correspondence at pretest.
As a whole, the better training performance in the blocked/contrast than the blocked/no-contrast condition suggests that overlap among consonants promotes better performance during learning (consistent with Roembke et al., submitted). Further, if we consider the change over blocks as a marker of learning, this also suggests greater learning in the blocked/contrast and interleaved conditions. It is not clear why effects of variability (or interactions) were not observed in this analysis, though we note that with the between-item design, and the issues regarding the assignment of GPCs to block, we may not have had the power to detect it.

General discussion

We start our discussion with three critical limitations of this study. Next, we discuss the results and their theoretical import. Finally, we conclude with broader thoughts on the field testing exercise and the role of cognitive science in education research.

Limits

It is important to acknowledge three limitations of the present study.

First, all the effects we observed were small, although they were statistically robust, particularly in low-performing children. This suggests that the benefits of blocking or variability may not be predictable for any given child. It may also be possible that these principles play out differently in children with different profiles of abilities (c.f., Perrachione, Lee, Ha, & Wong, 2011). As we were attempting a large sample size, we did not have the resources to conduct individual differences measures. A second possible explanation is that with only a short training regime, learning gains were numerically small. The average child improved from 69.2% to 75.7% over the training. This small range may have limited our ability to see large effects of the experimental factors.

Second, given the goals of this study, we focused only on dominant GPCs rules (e.g., that EA makes /i/ as in MEAT, not the EA→/ɛ/ rule as in DEAD). There may be other principles that govern learning these sub-regularities or true exceptions (Armstrong, Dumay, Kim, & Pitt, 2017; Kim, Pitt, & Myung, 2013).

Third, consistent with our goal of field testing, it is important to note that this was not a test of a curriculum and should not be interpreted as effectiveness research. Rather, we see these effects as potentially highlighting dimensions on which smaller parametric choices regarding how to structure activities or units within a curriculum to promote better learning.

Findings

Two factors were manipulated in this experiment: blocking and variability.

The blocking manipulation was broadly motivated by several goals. First, it captured the issue of how to best group GPC regularities in subunits of a curriculum, which is a form of blocked training. Second, blocked vs. interleaved training is an interesting principle of learning in its own right. Third, the two blocked conditions (contrast vs. no-contrast) were intended to test the idea that contrast on adjacent trials drives some of the benefits of interleaving (Carvalho & Goldstone, 2015). And fourth, our prior work led us to expect the benefits of variability to be impacted by blocked training.
We found mixed evidence for an overall effect of blocking. At immediate posttest, there was clear benefit for interleaved training, though this was lost by the retention phase, and may have been driven by the high-ability students (Supplement S2). Similarly, during training, the interleaved condition showed an upward trajectory, whereas the blocked/no-contrast condition did not, suggesting greater learning. Intriguingly, the blocked/contrast condition also showed an upward trajectory, suggesting it may not have been interleaving per se, but the opportunity to contrast overlapping items from trial to trial that benefitted learning. It is notable though, that at posttest (at least one day after training) some of these differences were not observed (the benefit for blocked/no-contrast training), and there was no effect of blocking at retention. Together, these results suggest a benefit for learning when contrasting items (e.g., GPCs that share a letter and thus need to be sorted out) are presented simultaneously; however, consolidation or decay may rapidly even out differences in performance due to overall effects of blocking.

In sum, our results suggest that blocking interleaving alone does not exert a large effect, particularly over the long haul, and when compared to blocked conditions featuring two regularities. Nevertheless, the relatively weak main effects of blocking may also be consistent with Carvalho and Goldstone’s (2014, 2015) claims that interleaving may not be uniformly superior, but may lead to better learning in only some conditions (e.g., moderated by variability).

The effect of variability produced no overall benefit. This, coupled with the relatively weak effects of variability observed by Roembke et al. (submitted), suggests that irrelevant variability may not be a uniformly beneficial principle in reading, but may only help under certain conditions. That was exactly what we observed here: There was no effect of variability in the interleaved or blocked/no-contrast conditions. These results help explain the lack of a robust variability benefit in Roembke et al. (submitted)—when six GPC regularities are taught at once (as was also done in that experiment), consonant variability matters less. However, in blocked/contrast training, we observed a marginal benefit of variability overall (which was significant by items) and a significant benefit when we considered only low-performing students (Supplement S1). These were also supported by individual comparisons showing a pattern of learning and retention indicating heightened learning when variability was coupled with blocked/contrast and poorer learning retention otherwise (Table 4). Further, these findings generalized to novel items and novel tasks, and appeared in both words and nonwords. There was no interaction of variability with test-phase. This suggests that the benefits of variability in the blocked/contrast condition reflect learning a durable regularity among letters and sounds, not memorization of items or task-specific learning.

This variability benefit with blocked/contrast training makes intuitive sense when we consider the task demands. When learning GPCs that share letters, children must link the same individual letter to two different sounds. In AI and OI, for example, the I can be linked to two pronunciations. Such one-to-one letter/sound associations are not a useful sort of mapping to acquire; rather the learner must learn that the conjunction of letters is what predicts pronunciation. Blocked/contrast training explicitly highlights this fact across trials as the pairs of GPCs in a block always shared a letter. This highlights the overlap that Roembke et al. (submitted) found to benefit learning.

Footnote: In fact it may do so even better than interleaved training as the likelihood of a contrast from trial to trial is much higher.
A similar problem is created by consonant similarity—when learning that BEAT, MEAT, and HEAT make the /i/ sound, the learner may incorrectly associate the T with the /i/ sound. Thus, blocked/contrast training in the context of similar consonants may compound the problem, making it difficult to sort out the array of potential associations that are needed to accurately link sound to meaning. This could potentially overwhelm the benefits of overlap. However, consonant variability can minimize this by preventing learners from associating any individual consonant with the pronunciation, helping them focus solely on the more problematic vowels. That is, in conditions where the child is already struggling to identify the digraph as a unit, further binding irrelevant consonants (in the similar condition) can cause problems, whereas variability heightens attention to the vowel.

In contrast, in the blocked/no-contrast condition, the overlap between vowels is minimized. Thus, learners may adopt a somewhat different strategy more akin to memorizing the vowels. Under such conditions, consonant similarity may play a less interfering role, and may even help by creating as similarity structure among the words enabling schema-driven learning processes (McClelland, 2013). Interleaved training among all six GPCs may also minimize contrast (since the chances of two overlapping vowels appearing side-to-side are smaller). At the same time, forcing the child to simultaneously figure out all six GPCs could similarly drive a more schema-driven mode of learning.

As a whole, these results point to a broader principle that learners need opportunities to contrast or relate material (Carvalho & Goldstone, 2014, 2015; Deng & Sloutsky, 2015; Kloos & Sloutsky, 2008). Indeed, Carvalho and Goldstone (2014, 2015) suggest there should be no uniform benefit of blocking. Instead, benefits of blocking/interleaving should be moderated by the nature of the mappings being learned: When the primary goal is to identify features that discriminate GPCs, interleaving will benefit learning; however, when the goal is to identify the constellation of defining features of each GPC (category), blocking may promote learning. Nevertheless, our results are not straightforwardly predicted by this account: One would imagine that it is the high similarity conditions—where learners need to discriminate among the vowels—that should benefit from contrastive blocking (or interleaving), whereas we find the converse.

Instead, our findings may be more consistent with connectionist approaches (Harm & Seidenberg, 1999; McClelland, 2013). Such models utilize similarity more broadly, not just among relevant units. Under this view, the need for contrast cannot be isolated to a relevant rule (e.g., the vowel), but rather appears to take the whole word into account, both in terms of related GPCs (that share vowels) or related items (that share consonants). Indeed, the effects of contrast and variability appear to complement each other. Consider just the two blocked conditions: We observed relatively good learning in the blocked/no-contrast + similarity and the blocked/contrast + variability subconditions (Figure 2; Table 4). In each, one factor creates some global similarity (in either the vowels or the consonants), and one does not. In contrast, where there is poorer learning, we see either no global similarity (blocked/no-overlap + variability) or too much (blocked/contrast + similarity). Thus, contrast and similarity seem to contribute to learning in similar ways. This is consistent with a schema-driven account which requires some degree of similarity to sort out the relations among the to-be-learned mappings. However, too much similarity leads to challenges and interference.
Finally, when we examined **performance over the course of training**, we found volatility from block to block in the blocked designs (Figure 3b). Some of this is most directly due to difficulty of the GPCs and items employed on a given block. However, we cannot rule out the possibility that retroactive or proactive interference may have played a role, consistent with the teacher’s intuitions about performance (discussed in the introduction). In contrast, the interleaved condition showed more continuous and even development. Here, training performance was reduced relative to the blocked conditions even as average accuracy after training was equal, and less sensitive to item factors (variability). This suggests that interleaving may be a viable route to long-term retention, and training under interleaved conditions may show a cleaner estimate of ongoing learning (e.g., for progress monitoring by teachers), even if it appears to lead to momentary difficulties during acquisition.

Although no main effect of trial-block was seen in the **blocked/no-contrast** condition during training (Figure 3c), this finding should not be interpreted as no learning in this condition: Pre/postcomparison showed gains in this condition at test. Thus, we suggest that learning in the **blocked/no-contrast** condition must have taken place largely in the first exposure to a given GPC, whereas it was more drawn out in the two contrastive conditions. This slower form of learning may be beneficial for integrating new GPCs into schemas reflecting similarity relationships across multiple GPCs (e.g., EA with EI, AI, and IE).

However, it is important to note that these differences at training did not necessarily carry over to test or to long-term retention. This could reflect offline consolidation, decay, or ongoing experience with reading outside of the training that helps integration. However, even if short lived, these effects are not irrelevant for education: In real reading education, learning does not happen in short bursts with long delays of no training. Rather, learning is continuous: New GPCs build on previous ones and there is ongoing re-exposure to learned GPCs (either as reviews, or in everyday text). Thus, the more ongoing measures of performance (during learning or at immediate posttest) may be more characteristic of performance in typical pedagogy than a delayed retention test, suggesting a potential benefit for interleaving or for blocks based on contrasting elements. We return shortly to the issue of what constitutes learning.

It is also important to note that training overestimates performance in multiple ways. We see this most clearly in Figure 3b where performance increases dramatically from pretest to block 1, and drops from block 6 to posttest. This likely reflects the added motivation of feedback during training (which was not present at test). In the blocked conditions, this is also likely due to the fact that response sets are implicitly limited to items from only two GPCs, making these tasks easier.

**Implications for education**

This study was not intended as a clinical trial, but it offers some useful ideas that should be carried forward into education research. As a field test, these ideas should not be treated as clear “guidelines” or advice. Rather they suggest ways in which the microstructure of a curriculum or a remediation might be designed to optimize learning. Such a curriculum or remediation, however, still requires randomized testing.
First, this study (and our prior work) suggests the need to reconsider what GPCs are being taught together. Roembke et al. (submitted) suggest that simultaneous training on overlapping GPCs may depress performance at first but lead to longer-term gains. In this way, standard groupings such as short vowels (e.g., A, I, O, in isolation) may not offer that kind of overlap. The present study adds to that by suggesting that achieving this contrast need not require throwing all GPCs together—small sets of targeted contrasts (particularly when item properties [e.g., variability] are controlled) can be as effective. Indeed, the design of the blocked/contrast condition was somewhat of a “round-robin” with each GPC contrasting with two other GPCs in targeted blocks (e.g., OA with EA in block 1, OA with OI in block 4). This may be a useful teaching strategy to consider down the road.

Second, it suggests that item characteristics (variability) can matter when blocks are restricted to a small number of regularities. In particular, variability seems to help when blocks highlight overlapping or contrastive GPCs, but not with non-contrastive blocks. Moreover, when multiple regularities are taught simultaneously, these item characteristics may wash out. There may be yet undiscovered principles in this regard that could be deployed. Moreover, it is not clear what counts as a “small” block—here we operationalized it as a short burst of about 60 trials of two GPCs. However, in a larger, more extended curriculum, even six GPCs may function that way.

Finally, these principles (contrast, variability) may be difficult to employ in a real, teacher-led curriculum. They require many items and fairly careful control over blocking of GPCs. In traditional educational approaches, these ideas could be directly implemented in spelling lists (which could adopt more or less variability), and in the order of teaching units. However, computer-based supplementary activities may offer an interesting learning opportunity. Whereas primary instruction can and should be left to the teacher, supplementary practice activities on the computer can bring to bear far more items that are carefully structured into contrastive blocks, along with other desirable properties such as immediate feedback and the ability to deploy multiple tasks. Indeed, this was part of our original motivation of working with the private sector; as we discussed this work with our partners, their instructional designers have been able to use these insights to select items, and structure curricula for optimal training. A crucial step down the road would be a randomized control trial of a supplement or remediation which directly tests these ideas. That is, one could compare a supplement in which items are selected to maximize contrast and/or variability (compounding what may be small effects across multiple studies) to a more standard one in which item choices are made by more typical practice.

**Bridging cognitive science and education**

The introduction offered a narrative of our systematic attempts to bridge from cognitive science to instruction. There were important conceptual challenges we faced that raise important issues for both domains.

First, the translation of what would appear to be a simple laboratory manipulation—blocking—was not straightforward. When considering the high dimensional mapping between letters and sounds, laboratory work was too limited to be of help. Much work on blocking only employs a small number of items to be learned, and often in paradigms in which subjects are motivated to learn. Consequently, they can “get away” with a single
to-be-learned mapping per block. That was not possible here—limiting the trials to a single GPC regularity made the task trivial. Intriguingly, this is worth consideration in the real world, where typical spelling curricula often focus on only a few letters at a time. Although the triviality may not be apparent in the classroom (where tasks are much more open ended than the forced choice task used here), at an implicit level, this property may still hold: In a unit on silent E, for example, children may rapidly figure out that all the words need to end with E. Confronting this issue opened up multiple degrees of freedom for ways to construct smaller blocks. Although our particular work was motivated by the idea of contrast as a structuring principle, there may be others. Future work in the cognitive sciences should consider higher dimensional learning problems (e.g., many categories, many mappings) to help isolate these principles.

Second, the design of our training was motivated by standard cognitive practice, but differed markedly from educational practice. Most spelling curricula, for example, might give children 20 words—all from the same “family.” This contrasts with our use of 72 training items and 72 test items. Of course, a large number of items is crucial in psycholinguistics, and here it likely has an added benefit—encouraging children to abstract a regularity, not to map an item. Similarly, our training featured a high number of nonwords; this was motivated by cognitive models such as (Seidenberg & McClelland, 1989) as an unambiguous way to target the direct mappings between sound and print. However, nonwords are controversial in education practice (because they are not items a child will “need” in real reading). These are straightforward design choices for cognitive science, but perhaps unexpected in educational practice. However, the effect of these choices may be to iron out differences, to make children expect to encounter new words (and therefore to learn GPCs and not memorize words), and to behave more flexibly. Indeed, this may be why we observed no moderation of our effects by new tasks or items, and not even a main effect between words and nonwords. Again, this suggests the possibility of using computer-based supplements to implement such designs.

Third, unlike either domain, however, we used multiple tasks. In a typical category learning experiment, for example, listeners might do a forced choice categorization task over and over again. Similarly, a typical spelling curriculum may include one task (Spell the Word), and students repeat it all week. In contrast, we used many tasks. In fact, there are dozens of tasks that can tap the mapping between sounds and letters. We used 10 here, but our other work has identified far more (Apfelbaum et al., 2013; Roembke et al., in press). Using multiple tasks is was motivated by our assumption of procedural learning as an underlying learning mechanism. Procedural learning requires many trials with active feedback, but getting a child through 400 trials of the same task would be nearly impossible. Varying the tasks with only a small number of trials of the same task overcame this issue. However, the inclusion of multiple tasks could have offered yet another form of contrast or variability to improve learning. This question is worthy of future laboratory and field work.

Fourth, what counts as knowledge? And what counts as learning? This design (many items, many tasks) affords a multidimensional test of children’s GPC knowledge. Take the simplest question: Did the children know the digraphs before we began? Chance was at 10–12.5% and only three children were below 20% at pretest. However, at the same time, in these tasks, children could perform above chance based on their knowledge of the foils, even if they had not encountered the digraph before. For example, in the Fill in the Vowel
task, if the target word was BEAK, they may be able to rule out OO (since they know that
would make BOOK) or AI (since that would not make a word). Learning a GPC regularity
is not a simple matter of memorizing the rule, but rather of being able to flexibly deploy a
new rule across many contexts, and developing a system which can implicitly select the
correct rule for a given word/context. Thus, even a child ostensibly performing at a high
level in this task has room for improvement and fine tuning. Indeed, knowledge or skill is
continuous, particularly when we consider that GPC knowledge is embedded and builds
on oral vocabulary, phonology and other aspects of reading. There is not likely a point
where children know nothing, and even a well-performing child may need continued
exposure and practice.

Learning may be an equally nebulous construct to measure. Here, the most rigorous
definition may be the difference between a well-matched pre and posttest, yet, when this
posttest is delayed (our retention tasks), differences can emerge (e.g., the main effect of
blocking disappeared). In fact, even our so-called immediate test was removed by at least a
day from training. This suggests that more immediate measures of learning (e.g., performance
during training) may be useful to isolate learning from consolidation and decay. However,
during training, the presence of feedback and task constraints (e.g., blocking) can cause such
measures to overestimate abilities. However, despite this, change over the course of training
may be a useful metric (even if the absolute magnitude of performance is inflated). In the real
world, this may get worse: Real education does not always have neat blocks, and a retention
task two weeks later could reflect interference (or benefit) from more learning in the inter-
vening period. This suggests a need to consider learning more broadly.

Finally, and perhaps most important, our work has encountered challenges in translating
simple cognitive theories like statistical learning to the real world. Statistical learning is framed
around items—what is the likelihood of a letter being mapped to a sound across many words?
However, pedagogy is not; it is structured around concepts like a GPC regularity.
Consequently, cognitive theory may not offer immediate guidance or clear predictions
about materials structured via other ways. These are not independent approaches: Decisions
at the level of GPCs have clear consequences for statistics at the item level. Computational
work may help translate these ideas; but there is also a need for laboratory work that is
motivated more directly by the kinds of framings used in the real world.

Lessons learned: Field testing and education research

As an example of a field-testing enterprise, this study offers important broader messages.
First, as this work and our prior studies show (Apfelbaum et al., 2013; Roembke et al.,
submitted, in press), the use of an Internet-delivered training paradigm offers several
advantages for this line of work. Computers offer the ability to more carefully control and
manipulate the items and training structure, and they can deliver immediate feedback
which promotes better learning (Ashby et al., 2002). They also offer high fidelity of
implementation, and they permit within-classroom manipulations, increasing statistical
power. However, even in highly controlled circumstances, the results were noisy and
effects were small. Large samples are necessary. However, we note that a large sample,
tested mostly automatically and over a relatively short time, may end up being more
economical than teacher intensive, longer-term work with smaller samples.
Part of the variability may come from child factors that need to be better measured and controlled in future work. Children varied substantially in their preexisting abilities with these GPCs. We attempted to mitigate this by using pretest data for stratified random sampling. However, differences in ability were still a factor: Much of the sensitivity to the blocking x variability manipulation derived from low-performing participants (Supplement S2). It perhaps would have been more effective (for the purpose of this study) to have simply dropped the high performers all together. Similarly, it is also important to time field tests to a common point in the curriculum, but of course, there are limits to this—to test sufficient children \ we had to test throughout an entire semester. The other major factor that needs to be considered is motivation. With computer-based training and testing, children can be easily tested in groups, but this comes at a cost. It can be hard to monitor individual children’s level of motivation throughout the experiment, and children may be better motivated by a human than a computer. While gamifying such field tests can help (Boyle et al., 2016; Hines, Jasny, & Mervis, 2009), this may come at a cost as children adopt strategies to “beat the game,” rather than learn the reading skills (Reed, Martin, Hazeltine, & McMurray, submitted).

However, ultimately, as we described, effects may be small in part because the learning paradigm was brief. Again, this speaks to the need to balance the needs of rapid field testing (which can be done in a week or two but requires large samples) and a longer-term learning paradigm (which may show larger effects with a smaller sample size). Education research is time consuming and expensive—and field testing represents one point in a complex cost-benefit array of sample size, individualization, and efficacy.

Finally, this work (as well as our prior studies) makes a strong case that curricular questions can be translated to learning scientific constructs, though as we described, the basic science is not always at a point where this can be done in a straightforward manner. Every curriculum or pedagogical activity must make significant choices about the specific items being used, the constellation of skills being taught, and the types of tasks being used. Although pedagogical goals must constrain these choices, our work suggests that parametric field testing can also help identify the particular parameter settings that may optimize learning. That is, even absent a strong theoretical motivation, field tests may play a useful role in pedagogical development. However, such field tests can lead to practices that may appear counterintuitive pedagogically—both variability and contrast may appear to make things challenging for children even as they help learning. In this way, it may be beneficial to engage broader discussions between educators and cognitive scientists to put these ideas into better perspective.

At the same time, this is not a one-way street. Results from field testing may also push basic learning theory. Roembke et al. (submitted) did not anticipate a strong benefit of simultaneous training on overlapping GPCs items, nor did we anticipate here that blocked/no-contrast training would show the learning performance that we observed. This suggests a need for cognitive science to consider perhaps more sophisticated learning problems, or to consider the way in which groupings at the level of regularities can play out in a statistical learning framework. Such work must consider the fact that learners may vary in their preexisting knowledge, that items may not always be perfectly balanced, and that they must be embedded in a complex set of knowledge about not just the GPC rules being taught, but also the other rules not being taught (e.g., the exceptions).
Acknowledgments

This project was supported by NSF BCS Grant #1330318 awarded to EH and by IES Grant # EDIES15C0023 awarded to Foundations in Learning, Inc. We are grateful to Carolyn Brown, Jerry Zimmerman, Jason Smith, and Eric Soride at Foundations in Learning for ongoing support of the project, and as well as theoretical insight into the problems of reading. We also thank Mike Freedberg and Jamie Klein-Packard for assistance in preparing stimuli and testing the children; and Alicia Emanuel, Shellie Kreps, and Tara Wirta for assistance with data collection.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Division of Behavioral and Cognitive Sciences [1330318]; Institute for Education Sciences [EDIES15C0023].

References


