Blocking as a friend of induction in verbal category learning

Linda J. Sorensen[1] · Dan J. Woltz[2]

Abstract We conducted an experiment to investigate the influences of interleaved versus blocked presentation of disparate verbal exemplars on the induction of category concepts. The practice schedules of four experimental groups were juxtaposed such that sets of exemplar–category associations were either solved in succession (i.e., blocked), systematically intermixed (i.e., interleaved), or presented with an incremental transition from blocked to interleaved practice. Counter to current trends in the literature, in which interleaving alone has been facilitative of induction in some tasks, we found that participants whose initial exposure to the category exemplars involved blocked presentation performed better in both implicit and explicit tests of concept learning. The results are discussed with respect to the impacts of task type, task difficulty, and exemplar relatedness on induction.

Keywords Blocking · Categorization · Induction · Interleaving

New category concepts are said to be learned inductively when they are learned by example. When a learner is exposed to exemplars that represent an unfamiliar category concept, acquisition of the concept depends on the induction of common features or relationships among the exemplars that serve to define the category. The mental operations of induction

differ from those demanded by methods of concept learning that rely on explicit verbal definitions. Whereas the cognitive processes demanded by the latter are similar to those used in other forms of declarative learning (e.g., fact and rule learning), inductive processes differ in that they are often implicit in nature (Reber, 1992). Furthermore, considerable evidence supports instructional principles that facilitate explicit forms of declarative learning (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007), including a robust literature demonstrating the benefits of practice schedules that incorporate a varied temporal distribution, or *spacing*, of the to-be-learned material (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, and Dempster, 1988, for reviews). In contrast, there is less evidence—and some disagreement—regarding the effect of temporally distributed practice on purely inductive concept learning. To address this disagreement, we investigated the effects on both inductive concept acquisition and transfer of four disparate presentation sequences of training stimuli.

Decades ago, Underwood (1952) declared that "for the perception of relationships among stimuli the needed assumption is that *appropriate responses to those stimuli be contiguous*" (p. 211, italics in original). Providing support for this hypothesis, Kurtz and Hovland (1956) found that the induction of nonsense-syllable category labels from simple geometric-pattern exemplars was hastened by the successive, or *blocked*, presentation of relevant stimuli, as opposed to an intermixed, or *interleaved*, format in which two instances of a given concept were never presented consecutively. Likewise, benefits to induction of blocked rather than interleaved stimulus sequencing were found, to varying degrees, in examinations decades ago of positive and negative concept instances (Hovland & Weiss, 1953), miniature linguistic systems (Foss, 1968), shape triads (Detambel & Stolurow, 1956), and low-relevance cues (Peterson, 1962). More recently, Carpenter and

✉ Linda J. Sorensen
  lsorensen59@gmail.com

[1] Urban Institute for Teacher Education, University of Utah, 1721 Campus Center Drive, SAEC 2248, Salt Lake City, Utah 84112, USA

[2] Department of Educational Psychology, University of Utah, Salt Lake City, Utah, USA

Mueller (2013) found that blocked presentation of stimuli surpassed interleaved presentation in the learning of rules for pronouncing French words.

Other researchers used computer-generated images (Zulkiply & Burt, 2013) and blob figures (Carvalho & Goldstone, 2014b, Exp. 2) to explore how varied presentation sequences might interact with the difficulty of category discrimination. They found interleaved exposure to be most beneficial when the task required the learner to determine the differences between highly similar, and therefore confusable, categories (a notion that Kurtz and Hovland had speculated about in 1956). Hence, empirical evidence simultaneously provides support for, but points to a boundary condition of, the *massing-aids-induction* hypothesis (Kornell, Castel, Eich, & Bjork, 2010). According to this hypothesis, presenting exemplars from a given category consecutively facilitates induction by allowing for the comparison, and subsequent encoding, of shared characteristics.

Temporal spacing is a feature of practice schedules that typically overlaps with the distinction between blocked and interleaved stimulus presentation. Better acquisition and transfer from temporally spaced repetitions of to-be-learned information has been demonstrated in many learning paradigms (e.g., paired associates, list learning, rule learning, and vocabulary acquisition). Ebbinghaus (1895/1913) noted that "with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time" (p. 89). He was describing the memorial benefits of temporal spacing for his own recall of nonsense syllables. Since then, the *spacing effect* has been demonstrated across myriad domains, settings, tasks, and age groups (see Cepeda et al., 2006, Dempster, 1996, Hintzman, 1974, and Melton, 1970, for reviews of verbal recall tasks; Shea & Morgan, 1979, provide an example involving motor learning). Of importance here, spacing is a feature inherent to some degree in interleaved presentation, and if blocked presentations are repeated, it can be instrumental there, too.

The observation of an almost ubiquitous superiority of spacing in memory experiments, in which the retention of repetitively trained items is later tested, prompted distinguished educational psychologist Ernest Rothkopf to reemphasize in 1977 that "spacing is the friend of recall, but the enemy of induction" (Kornell & Bjork, 2008, p. 585). Rothkopf's assertion of a boundary condition to the benefits of temporal spacing was directly tested by Kornell and Bjork in an experiment in which participants induced the painting styles of 12 artists. Using a within-subjects design, paintings by an artist were either presented consecutively (their *massed* condition) or distributed across time by being intermixed with other artists' paintings (their *spaced* condition). Indeed,

participants' ease-of-learning metacognitive judgments aligned with Rothkopf's stance: Fully 78 % perceived their learning as having been better when six of a given artist's paintings were shown consecutively. However, their accuracy on a transfer test revealed that the opposite was true: 78 % of the participants correctly classified new paintings by these artists (i.e., successfully induced the artists' styles) when the artists were learned under the spaced rather than the massed condition, despite their intuition to the contrary. These results were not expected by the authors, presumably because the findings deviate from historical evidence of inferior inductive learning when presentations of exemplars are spaced.

Kornell and Bjork's (2008) surprising findings sparked related research, including their own replications of the results with older adults (Kornell et al., 2010) and younger children (Vlach, Sandhofer, & Kornell, 2008). With a young adult sample, Zulkiply, McLean, Burt, and Bath (2012) extended the findings to verbal stimuli presented both auditorily and visually. Importantly for our purposes, a few studies have examined issues related to interleaving, in addition to temporally spacing the stimuli. For example, Kang and Pashler (2012) included combinations of both interleaved and spaced conditions in their presentation of works of art and found that increased temporal spacing alone was insufficient to produce improvements with induction. Rather, the interleaving of exemplars— and the enhanced discriminative contrast afforded by that presentation—was key to better learning. Additional support was found for this *discriminative-contrast hypothesis* in experiments conducted by Birnbaum, Kornell, Bjork, and Bjork (2013) using pictorial representations of birds and butterflies. On the whole, these findings represent a serious challenge to the assumption that interleaving impedes the induction of novel concepts from exemplars.

Building upon the intriguing metacognitive aspects of Kornell and Bjork's (2008) study, Tauber, Dunlosky, Rawson, Wahlheim, and Jacoby (2013) investigated the influence of self-regulatory processes during the formation of complex concepts. While learning to associate exemplar birds with their respective bird families, participants were allowed to decide for themselves the sequencing of to-be-studied exemplars after receiving prompts such as, "You just studied a _____. What would you like to study next?" (p. 360). These researchers found an overwhelming preference for selecting additional birds from a given family (blocked study) over mixing birds from different families (interleaved study), although there was ultimately no error difference by formats. However, as others have found in studies involving both verbal- and motor-skill learning (e.g., Birnbaum et al., 2013; D. A. Simon & Bjork, 2001; Zechmeister & Shaughnessy, 1980), interleaved study conditions typically result in better learning, even as participants' subjective judgments suggest the opposite.

Overview of the experiment

Although much recent evidence is compelling regarding the benefit of interleaved exposure of exemplars during inductive learning, we wondered whether this conclusion holds under all inductive-learning demands. Furthermore, we wondered whether the contrast between all-interleaved and all-blocked exposure obscures the potential benefits of each at difference stages of induction. The primary question addressed in the present study pertained to the optimal distribution of to-be-induced information; hence, in our design we adopted Kurtz and Hovland's (1956) recommendation to compare degrees of interleaved presentation. We juxtaposed practice schedules in which sets of exemplar–category associations were either solved in succession (i.e., blocked), fully interleaved, or presented with a gradual transition from blocked to interleaved practice.

Through a verbal inductive-learning task paradigm, the participants in this study learned categorization rules associated with several nonword category names (NCNs). We chose disparate categories to reduce confusability, while strategically selecting exemplars with somewhat obscured relatedness, both to each other and to their respective categories, to increase the difficulty of inducing the meaning of each category. We assumed that the potential benefit to induction of blocked exposure partly depends on the difficulty of recognizing relevant similarities among the category exemplars. Moreover, we reasoned that, in real-world concept acquisition, recognizing relevant similarities is sometimes cognitively demanding, because of interference due to competing, irrelevant features. For example, learning that whales belong to the mammal category is made difficult initially because of whales' salient but irrelevant similarities to fish. Under these task conditions, we hypothesized that the nontrivial induction of category concepts would depend on experiencing at least a modicum of initial contiguous presentation of the exemplars. However, to the extent that concepts were successfully induced during initial blocked practice, continued reliance on consecutive presentations of the exemplars should hinder transfer of the concepts to new instances, due to the absence of the variability afforded by interleaving.

Participants were informed that they would be learning new names for categorizing familiar objects. Over 19 blocks, exemplars were repeatedly presented and labeled with NCNs. The exemplars were familiar entities, yet discovering the commonalities among them was not trivial, due to the relatively novel feature overlap that defined category membership. For example, the exemplar *potato*, which would typically be categorized as a "type of food" or "type of vegetable," was categorized as something found "underground." The other three exemplars representing *BRASK* (the nonword for "underground") were diverse (*cave*, *roots*, and *tunnel*), rendering induction of the category challenging due to competing

associations (Chen, Ross, & Murphy, 2014). Moreover, the participants were never explicitly given the category definition, but instead were required to induce the common, category-defining feature of the exemplars associated with the new category name.

Intermittent tests of category learning were presented within the training blocks, and a final test of induction required the categorization of new exemplars presented during a subsequent transfer phase. To successfully categorize new "underground" transfer items from the illustration above, participants had to have (1) induced that *cave*, *roots*, *potato*, and *tunnel* share the common feature of existing beneath the ground; (2) associated the "underground" concept with the nonword *BRASK*; and (3) recognized that new exemplars (*worm*, *gopher*, *well*, and *aquifer*), similarly belonged to the "underground" category labeled *BRASK*.

Due to the complexity of inducing the exemplar–category matches and the presumed cognitive load, especially during interleaved blocks, we hypothesized that participants who were beneficiaries of at least a few blocks of blocked practice at the beginning of training, followed by many blocks of interleaved practice, would outperform the participants who only received interleaved practice. We evaluated this prediction with periodic testing during the acquisition phase and with performance on subsequent transfer trials. We further anticipated that this learning advantage would manifest itself when, at the conclusion of the experiment, participants were asked to explicitly describe the inductively learned categories associated with each new category name.

Method

Participants

The participants were drawn from the educational psychology subject pool at the University of Utah. Of the original 175 participants, the data from 15 (9 %) were eliminated according to one or both of the following criteria. First, if a participant committed more than 20 % errors during the last set of training blocks, when correct responses were visible on the screen and the participant had already experienced eight blocks of practice, this was taken to indicate a lack of effort (retained participants, $M = 4.8$ % errors, $SD = 3.3$; excluded participants, $M = 41.4$ % errors, $SD = 9.8$). Second, unrealistically fast response times (RTs; i.e., <1 s) during the transfer phase, which constituted the final test of learning, were taken as evidence of a lack of engagement in the task (retained participants, $M = 1$, 918 ms, $SD = 542$; excluded participants, $M = 580$ ms, $SD = 199$). The exclusions were distributed equally across the four experimental conditions, $\chi^2(3, N = 175) = 1.66$, $p = .646$, and the conclusions and statistical test outcomes were unaffected by replacing the outliers with engaged participants. The final

sample ($N = 160$) included 115 females and 45 males ranging in age from 18 to 51 years.

Apparatus and setting

While seated in computer carrels separated by sound-deadening panels, participants performed the experimental task on desktop computers with SVGA monitors and standard keyboards. The programming of all tasks was completed with E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). Up to five individuals participated during a single, 1-h session.

Materials

Each of the six conceptual categories was assigned a five-letter, single-syllable, pronounceable NCN. Care was taken to avoid any logical association between a given NCN and the actual category it represented. Eight diverse exemplars were selected for each category, then divided into two equivalent sets, one utilized during training and the other during transfer for a given participant. The four exemplars comprising each set remained constant, but which sets were presented in training and transfer was counterbalanced across participants. (A complete table of the categories and corresponding exemplar sets can be found in the Appendix.)

Experimental task

The purpose of each training trial was to promote an association between a given NCN and one of its four exemplars. Figure 1 shows a sample trial slide of the type that included an answer key across the top of each screen (a *visible* trial). The six NCNs were displayed in red uppercase letters, each above one of its four exemplars, shown in blue lowercase letters. At the center bottom of the screen was the item to be solved, consisting of one of the six NCNs and two exemplar response choices below, one correct and one incorrect. The two response choices displayed on any given trial always appeared in the answer key at the top for that trial, but only one was associated with the NCN being probed. In the early stages of practice—before the associations were learned—participants needed to search the answer key, find the pairing at the top that matched one of the two possible pairings at the bottom, and then respond with the correct answer ("z" for the response on the left side and "/" for the response on the right side).

Using the trial shown in Fig. 1 as an example, upon seeing *CHILT* with *mirage* and *frisbee*, the participant would look up, see *CHILT* with *frisbee* in the key, and then press "/." The incorrect exemplar appearing beneath the probed NCN on a given trial was randomly selected from the participant's training set, and these exemplars were used as incorrect alternatives with equal frequencies across the training trials. The ordering of categories in the answer key display was
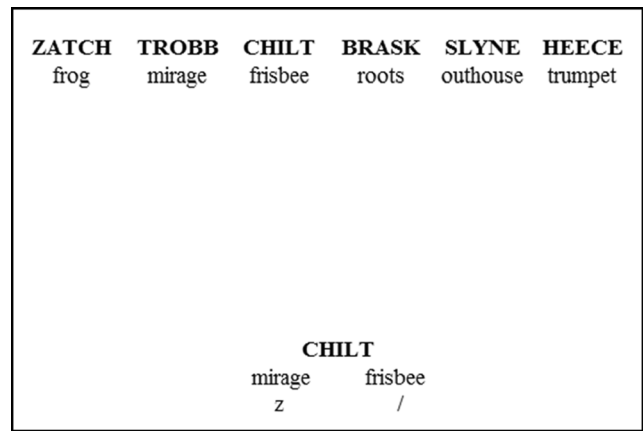

ZATCH	TROBB	CHILT	BRASK	SLYNE	HEECE
frog	mirage	frisbee	roots	outhouse	trumpet

CHILT

mirage   frisbee

z   /

Fig. 1 Sample of a visible trial

randomized for each trial, to increase the effort required to locate the correct response, thereby encouraging participants to learn the category concepts rather than conduct a predictable visual search.

Our intent in providing the answer key was to foster a relatively error-free learning opportunity that would resemble the implicit acquisition of natural categories in the real world. As we noted, no category definitions were presented. In addition, to discourage explicit learning strategies (e.g., memorizing exemplar–NCN associations or extensively analyzing the category features), a response deadline was imposed on each training trial. During the first training block, when participants presumably relied heavily on visual search, the deadline was 6 s. If a response was not made during this time period, the trial disappeared and the words *Too slow* appeared. This deadline was decreased gradually—by 1 s per block—to 3 s in the fourth block (constituting what we refer to as *Set 1*), and the pattern was repeated across Blocks 6–9 (*Set 2*) and 11–14 (*Set 3*) under the assumption that responses would be based increasingly on memory rather than visual search (cutoff times are included in Table 3 below).

There were two types of training blocks, reflecting the manipulation of blocked and interleaved practice schedules, respectively. Both types included 24 trials, with each of four training exemplars per category being presented once as the correct response. During a blocked practice block, the four exemplars from a category were clustered together, such as $A_1A_2A_3A_4 \; B_1B_2B_3B_4 \; C_1C_2C_3C_4 \; D_1D_2D_3D_4 \; E_1E_2E_3E_4 \; F_1F_2F_3F_4$ for categories A through F, each with Exemplars 1–4. In the subsequent blocked blocks, the order of the categories and exemplars was randomized (e.g., $C_2C_4C_3C_1F_2F_4F_1F_3A_3A_1A_2A_4D_1D_4D_3D_2B_1B_2B_4B_3 E_4E_2E_3E_1$). During an interleaved block, the four exemplars from any given category were interleaved with exemplars from the other five categories, and thus appeared on every sixth trial (e.g., $B_1D_2F_3A_4C_3E_2 \; B_2D_3F_4A_1C_2E_1 \; B_3D_4F_2A_3C_3E_4B_4D_1F_1A_2C_4E_3$). Regardless of the sequencing scheme, all participants experienced some level of temporal

spacing of the exemplars, either between blocks only (as in the blocked practice blocks) or both between blocks and within a block (as in the interleaved practice blocks). The order of category and exemplar presentation was randomized across blocks for each participant. Regardless of the type of practice schedule, feedback of the average RT and percentage errors was given at the end of each block, as was a summary feedback table presenting the data from all blocks to that point (see Table 1).

Periodic assessments of both memory for specific pairings and induction of the general categories were inserted during the learning phase, in the form of *blind* trial blocks, which presented trials identical to the visible trials, except that the answer key at the top of the display was removed. Blocks 5, 10, and 15 entirely comprised blind trials, and the response deadline was increased to 10 s in each of these blocks. Participants had been advised during the initial instructions and practice items to "make an effort to learn the new category names so that you no longer need to look up at the lists." During the 15-block learning phase, we further prepared participants for this occasional change in task presentation by reminding them at the end of each block: "During Blocks 5, 10, and 15, the answer key will cease to appear." The summary feedback table visually highlighted the occurrence and timing of the blind testing blocks that had previously been described in the verbal instructions.

Table 1 Sample summary feedback table

	Response Time	Accuracy
Block 1	2.25 s	87.50 % correct
Block 2	2.03 s	91.67 % correct
Block 3	1.72 s	87.50 % correct
Block 4	—— s	—— % correct
Block 5	—— s	—— % correct
Block 6	—— s	—— % correct
Block 7	—— s	—— % correct
Block 8	—— s	—— % correct
Block 9	—— s	—— % correct
Block 10	—— s	—— % correct
Block 11	—— s	—— % correct
Block 12	—— s	—— % correct
Block 13	—— s	—— % correct
Block 14	—— s	—— % correct
Block 15	—— s	—— % correct

During the training phase, participants received feedback at the end of each block that included not only their most recent RT and accuracy, but also a record of previous performance. (A similar screen appeared at the end of each block during the postbreak final testing and transfer phases, but only 12 blocks were listed)

Experimental design

Random assignment was used to place participants into one of four conditions, which varied only as to the ratios of blocked-to-interleaved practice that participants received. In all conditions, participants performed 15 training blocks with periodic blind testing blocks. Following these initial training blocks, but before the final testing and transfer blocks, participants performed an adaptation of the Thurstone Letter Series Completion Test (Schrepp, 1999; H. A. Simon & Kotovsky, 1963). To successfully solve the items in this task, participants were required to extrapolate the serial pattern from a limited but sufficient sequence of letters that embodied it. As a simple example, upon seeing *xyxyxy_ _*, a participant would have to determine the two letters that come next (*x* and *y*) and type them in. A more complex example is the series *pononmnmlmlk_ _*, which required detection of a pattern of longer length and involving letters in backward alphabetic order. (*l k* is the appropriate response.) Two blocks of 12 letter series items each were presented, ranging in difficulty from simple to complex within both blocks. Participants were given up to 4 min to complete each item but completed the entire task, on average, in less than 12 min.[1] This task was included in the experiment to ensure that performance on the subsequent test blocks reflected a persistent understanding of the category concepts.

Table 2 depicts the four between-group conditions, with the letter series task occurring immediately before all groups engaged in the interleaved testing and transfer blocks of Part 2. The participants in the all-interleaved, or *AI*, condition received no blocked practice but experienced 15 interleaved blocks in succession during the training phase; thus, their ratio of blocked-to-interleaved blocks was 0:15. The other three groups experienced incrementally greater amounts of blocked practice prior to shifting to exclusively interleaved practice, as follows: low-blocked (*LB*) = 5:10, medium-blocked (*MB*) = 10:5, high-blocked (*HB*) = 15:0.[2]

---

[1] The means and standard deviations [M (SD)] for the duration (in minutes) of the letter series interpolated task, overall and by condition, were as follows: overall, 11.90 (4.99); LB, 12.09 (4.27); MB, 12.82 (6.02); HB, 11.52 (4.76); AI, 11.14 (4.75). We observed no group differences, F(3, 156) < 1.

[2] We chose not to include in our design a fifth group that would experience massed exposures to exemplars throughout the entire experiment, on the basis of the following reasoning. The ultimate test of whether or not participants had induced the category concepts occurred during the transfer blocks, when simple memory for individual learning-trial exemplars would not likely support relatively fast and accurate categorization of novel exemplars. For the transfer blocks to be comparable across the experimental groups, the format needed to be either spaced or massed for everyone; spaced exposures seemed to us to provide the more rigorous test of successful induction. Thus, to avoid a potential confound, it was important that the blocks completed just prior to the transfer phase be spaced, as well, so that every group had at least some experience with that more-varied format before transfer.

Table 2 Design of the experiment

| | Part 1 | | | | | | | Part 2: | |
| | Training | | | | | | | Testing | Transfer |
	Set 1: Blocks 1–4	Testing Block 5	Set 2: Blocks 6–9	Testing Block 10	Set 3: Blocks 11–14	Testing Block 15		Blocks 16–19	Blocks 20–27
High Blocked (HB)	4 visible, blocked	1 blind, blocked	4 visible, blocked	1 blind, blocked	4 visible, blocked	1 blind, blocked	Interpolated task: Letter Series	4 blind, interleaved	4 blind, new exemplars, interleaved
Medium Blocked (MB)	4 visible, blocked	1 blind, blocked	4 visible, blocked	1 blind, blocked	4 visible, inter-leaved	1 blind, inter-leaved		4 blind, interleaved	4 blind, new exemplars, interleaved
Low Blocked (LB)	4 visible, blocked	1 blind, blocked	4 visible, inter-leaved	1 blind, inter-leaved	4 visible, inter-leaved	1 blind, inter-leaved		4 blind, interleaved	4 blind, new exemplars, interleaved
All Inter-leaved (AI)	4 visible, inter-leaved	1 blind, inter-leaved	4 visible, inter-leaved	1 blind, inter-leaved	4 visible, inter-leaved	1 blind, inter-leaved		4 blind, interleaved	4 blind, new exemplars, interleaved

During training, the answer key appeared at the top of every trial slide during the *visible* blocks, but was absent during the *blind* blocks.

Procedure

The session began with computer-administered instructions and one block of 12 practice trials with corrective feedback. Participants were instructed to utilize the displayed answer key as much as was needed to make the correct response selection, but nonetheless to respond quickly to each probe. Practice trials were followed by 15 training blocks, the interpolated task, and then four final blind testing blocks. Following Block 19, with no cue or foreknowledge, participants proceeded to the eight-block transfer phase (which had a 20-s timeout for each trial). During blind transfer Blocks 20–27, new exemplars for each category replaced the previous exemplars, providing the most demanding assessment of induction of the category concepts represented by the NCNs.

Explicit knowledge of the category definitions was measured by a six-item postexperiment questionnaire. After completing the computerized portion of the experiment, participants were asked to record on paper "the meaning you assigned to each nonword category name." A team of five raters independently rated the accuracy of each response on a 3-point scale (2 = *correct category definition*, 1 = *partially correct category definition*, and 0 = *no/incorrect category definition provided*). For the category "underground," for example, a response such as "beneath the ground" or "in the earth" scored a 2, "dark hole stuff" scored a 1, and simply listing an exemplar ("tunnel") scored a 0.

Results

The data are presented in three sections. First, we present descriptive data for performance during the learning phase in which the answer key was visible on all trials. Second, we present analyses of the three blind test blocks that were presented periodically during the learning phase to test the acquisition of concepts. Finally, and of primary importance, we present analyses of performance during the final testing and transfer phases. The p value was set at .05 in all statistical analyses.

Learning phase practice trials

Figure 2 presents the mean percentage errors and mean RTs for the three four-block sets of training blocks in the learning phase. We present these data descriptively to convey the general pattern of performance, since groups differed in their exposure to blocked and interleaved exemplar presentation. As is shown in the left panel of Fig. 2, the mean error rates were relatively low and consistent across groups. This generally low error rate reflected the availability of the correct answer to each trial with a visual search of the answer key, even before the category concepts were acquired. Although the mean error
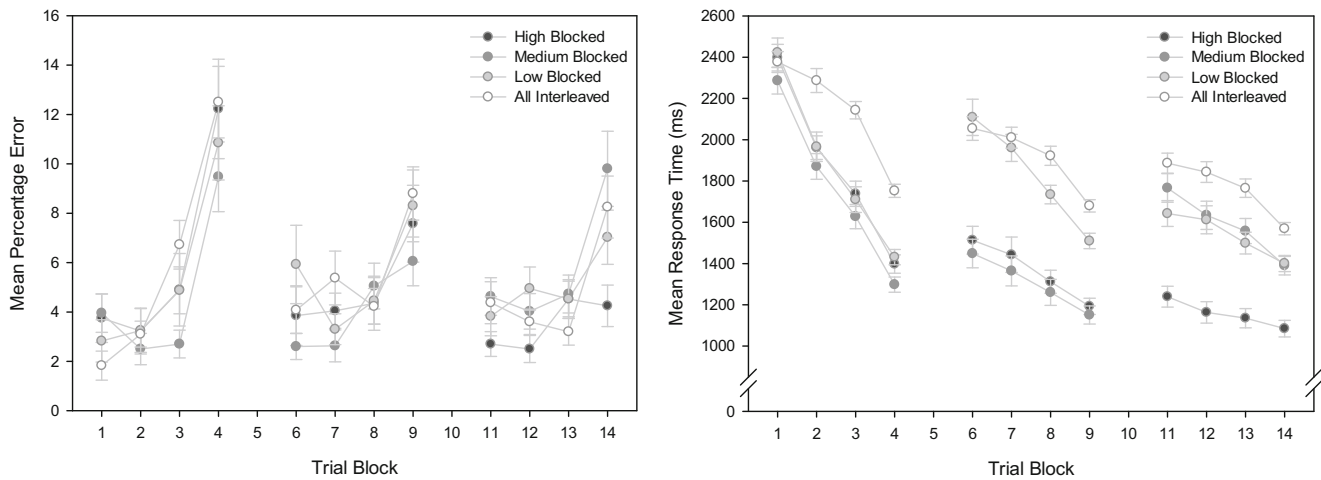
Fig. 2 Mean percentage errors (left panel) and mean response times (right panel) for the learning blocks, by presentation condition. Error bars represent standard errors of the means

rates were at or below 5 % for most practice blocks, there was a notable increase in errors during the final block of each set (Blocks 4, 9, and 14). Here the mean error rates approached and, in the case of Set 1, slightly exceeded 10 %. We suggest two explanations for these spikes. First, failure to respond before the deadline was counted as an error, and by the fourth block in each set, the deadline was reduced to 3 s. Table 3 presents the percentages of timed-out trials per block within sets. To the extent that the time limit influenced the error rate, this dependent measure reflected aspects not only of accuracy but also of speed of responding. Second, we speculate that the increase in errors immediately prior to the blind blocks was due to spontaneous self-testing by participants. Inasmuch as the summary feedback table appearing at the end of each block clearly indicated that Blocks 5, 10, and 15 would be "blind" as to the answer key, participants may have challenged themselves by reducing their use of the information at the top of the screen and, consequently, committed more response errors or timed out before responding.

The right panel of Fig. 2 presents the mean RT data for the three sets of training blocks. Here the pattern of performance reflects two phenomena. First, the mean RTs for all groups

Table 3 Cutoff times during training blocks and incidence of timed-out trials (TO)

Block Within Set	Cutoff (s)	Set 1:% TO	Set 2:% TO	Set 3:% TO
1st	6	0.4	0.1	0.0
2nd	5	0.7	0.3	0.2
3rd	4	1.9	1.3	0.8
4th	3	7.0	4.1	3.4
5th (blind)	10	0.1	0.0	0.0

declined with practice. This likely indicates that participants needed to search the answer key less frequently as category concepts were acquired. Second, group differences across the three training-block sets clearly show the impact of blocked versus interleaved exemplar presentation. Participants responded faster under blocked presentation as category labels in the probes remained constant over sets of four trials. Although this consistency of category labels across trials represented reduced difficulty of the learning trials relative to interleaving, the correct response for each item was not obvious until category concepts had been acquired. In each trial, participants had to decide which of two exemplars belonged to the category name, and without understanding the category concept, a visual search for the answer would be necessary even in the blocked condition.

The mean data for the blind test blocks for the four groups are presented in Fig. 3. The error data shown in the left panel are of primary importance, as lower error rates could only be achieved through the acquisition of category concepts or the memorization of individual exemplar–NCN associations. Several patterns across the three blocks reveal the impact of the different practice schedules on this acquisition. First, error rates were generally reduced with practice in all groups, as reflected by a main effect (linear) of blind block number, $F(1, 156) = 50.55$, $p < .001$, $\eta_p^2 = .25$. Second, the AI condition had more errors overall than the other conditions, which contained some blocked presentation of exemplars. As is evident in the left panel of Fig. 3, there was a substantial difference after just four blocks of practice between the three groups receiving blocked exemplar presentation and the group receiving interleaved exemplars, $F(1, 156) = 29.50$, $p < .001$, $\eta_p^2 = .16$. This contrast was still significant, but was reduced in magnitude, as blocked presentation groups made the
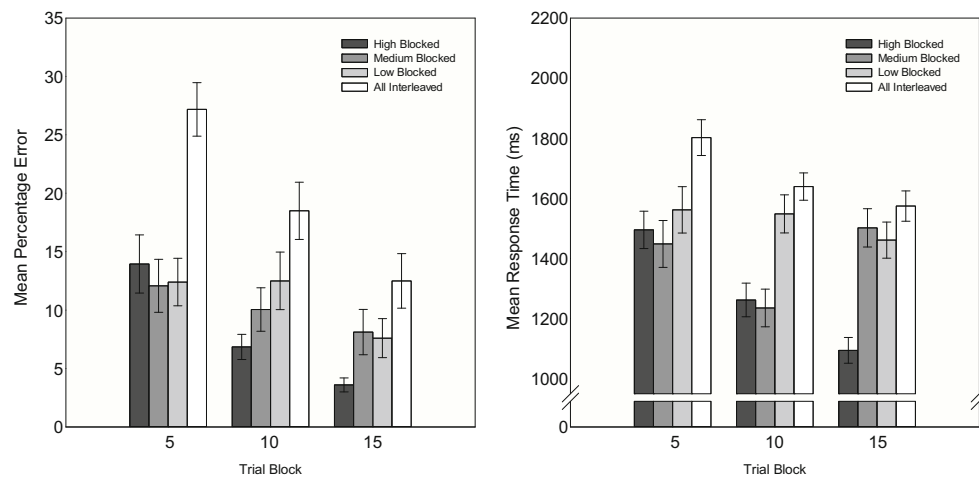
Fig. 3 Mean percentage errors (left panel) and mean response times (right panel) for the blind test blocks during learning trials, by presentation condition. Error bars represent standard errors of the means

transition to interleaved presentation, $F(1, 156) = 13.35$, $p < .001$, $\eta_p^2 = .08$, for Block 10, and $F(1, 156) = 8.44$, $p < .005$, $\eta_p^2 = .05$ for Block 15. Overall, the blind-block error data indicate that the presentation of even a few blocks of blocked exemplar practice resulted in more accurate performance when participants relied on acquired concept understanding in the three blind test blocks.

Finally, we examined the impact of switching from blocked to interleaved practice on blind-block errors. In Block 10, the LB group had switched from blocked to interleaved presentation in the preceding training blocks. As can be seen in the left panel of Fig. 3, their error rate in Block 10 did not increase substantially from Block 5, did not differ significantly from those in the two groups who continued with blocked practice, $F(1, 156) = 2.57$, $p = .11$, and was still somewhat lower than the AI group error rate, $F(1, 156) = 4.24$, $p = .039$, $\eta_p^2 = .03$. In Block 15, the pattern was similar. The two groups that had been switched from blocked to interleaved practice had fewer errors on average than the AI group, $F(1, 156) = 4.77$, $p = .032$, $\eta_p^2 = .03$. However, these groups made marginally more errors than the group that had had only blocked practice to this point, $F(1, 156) = 4.04$, $p = .049$, $\eta_p^2 = .03$. In total, the results of the blind block tests indicate that blocked exposure to the category exemplars in this task benefited performance, and that the benefit was not dependent on remaining in the blocked condition.

The RT means shown in the right panel of Fig. 3 primarily reflect the difference between blocked and interleaved presentation within the blind blocks that differed by group. In the four training blocks preceding the blind test blocks, the RT differences partly reflected time savings as participants learned category concepts and no longer needed to perform a visual search of the answer key (see Fig. 2); in the blind blocks, RTs only reflected decisions based on current understanding of the concepts or memory for specific

exemplar–label associations. The consistent pattern in Fig. 3 of longer RT means with interleaved presentation during the blind test blocks likely represents in part the additional time required to recall the associations from categories that differed on each subsequent trial. We also found evidence of learning across the three blind blocks for the two groups that had consistent presentation formats in the 15 blocks. Both the HB and AI groups had clear trends of reduced RTs over the three blind blocks.

Final testing and transfer trials

Figure 4 presents the performance data for the final test and transfer blocks that followed the break. Our hypothesis about the benefit of initial blocked practice was tested with a set of orthogonal contrasts. We first compared the AI group with the three groups receiving varying levels of blocked practice (LB, MB, and HB). Next, to investigate whether the amount of blocked practice mattered, we compared the HB group with the LB and MB groups combined, and finally we contrasted the LB and MB groups with each other. These contrasts for the error data represented the primary test of the previous learning-trial manipulations. All groups now performed under the same interleaved presentation format with no answer key available. Accurate performance, especially on the final eight blocks that contained new exemplars, could only result from accurate and generalized understanding of the categories.

Blocks 16–19 represented a final test of learning the original exemplar–category pairings. As is shown in the left panel of Fig. 4, the group that had had no blocked practice (AI) made more errors than the other groups, although the effect size was relatively small, $F(1, 156) = 7.94$, $p = .006$, $\eta_p^2 = .05$. The differences due to the amount of blocked presentation
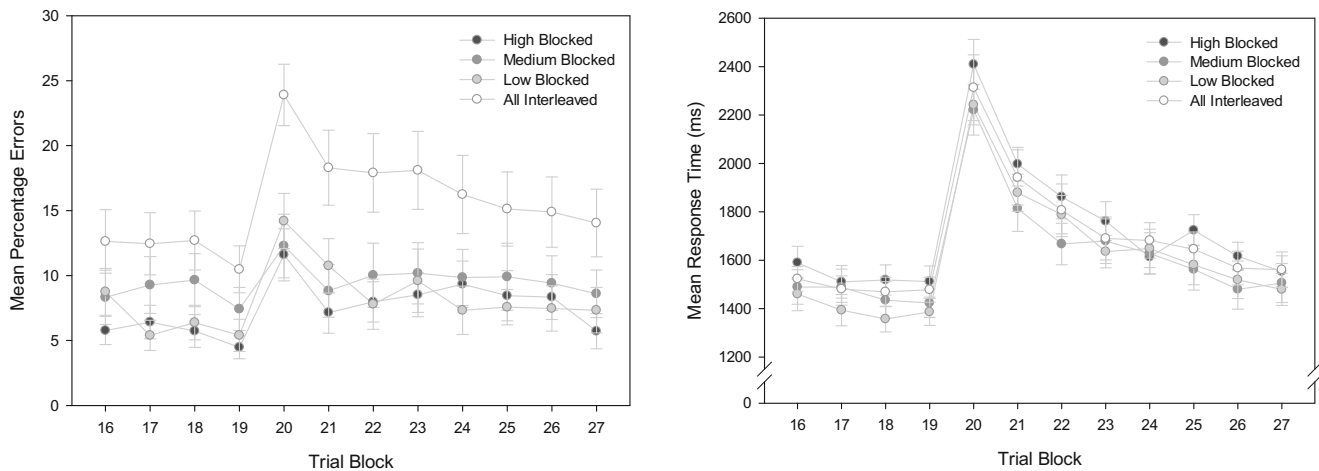
Fig. 4 Mean percentage errors (left panel) and mean response times (right panel) for the final test and transfer blocks, by presentation condition. Error bars represent standard errors of the means

were not statistically significant. However, the greater impact of no blocked presentation during learning occurred when new category exemplars were introduced in transfer Blocks 20–27.[3] As is shown in Fig. 4, the three groups that had had some amount of blocked presentation showed a small increase in errors when the new exemplars were first introduced (Block 20), but quickly regained the same level of accuracy they had demonstrated for the practiced exemplars. In contrast, the AI group showed a more dramatic increase in errors, never regaining their earlier level of accuracy. Apparently, the AI group was able to learn the exemplar–category associations reasonably well by the conclusion of the Part 1 training blocks, with only slightly more errors than those exposed to some blocked presentation. However, this learning was specific to the learning-task exemplars, and their general understanding of the concepts appeared to be weak relative to that acquired by participants in the other groups. In the statistical analysis of errors in the transfer blocks, the participants in the AI group committed significantly more errors than did those in the other three groups, $F(1, 156) = 12.9$, $p < .001$, $\eta_p^2 = .08$. However, the amount of blocked practice failed to differentiate between the other three groups, $Fs(1, 156) < 1$ for both contrasts.

Unlike the error data in Fig. 4, the RT data (right panel) show similar patterns of performance for all groups. The RT means are comparable across conditions in the final test

---

[3] One exemplar learned by all participants during the training blocks was inadvertently included in the transfer blocks, when participants should have encountered only previously unseen exemplars. Specifically, the exemplar *sparrow*—representing the category "can fly"—was old, not new, during transfer Blocks 20–27. It comprised 1/24 of the items in each of the eight transfer blocks, or 4 %. This was the situation for all participants, irrespective of which set of exemplars (A or B) had served as the training set and which served as the transfer set. The transfer data were reanalyzed after detection of the error, and the corrected data—with the *s* RT and error scores removed—are reported. No conclusions changed as a result of the adjustment.

blocks, and all participants took considerably more time to respond to the new exemplars when the initial transfer blocks were introduced. With practice, all groups responded more quickly, and no differences between learning conditions were evident.

Explicit memory test

Of the 160 participants, 152 completed the postexperiment explicit memory questionnaire. The eight participants who inadvertently did not receive the questionnaire were distributed as follows: three from the HB group, two each from the AI and LB groups, and one from the MB group. Table 4 displays the accuracy score means and standard deviations for the six learned categories by conditions. The interrater reliability (intraclass correlation) for the five raters scoring the responses was .96 or greater for all categories.

As with the previous performance tests of learning, we predicted that participants' explicit learning of the category concepts would be facilitated by experiencing at least some amount of blocked practice before receiving spaced practice. To test this hypothesis, we again first compared the AI group with the three partially blocked groups. We then compared the HM group with the combined LM and MM groups, and the LM and MM groups with each other, to investigate how much blocked practice was optimal.

The participants who had experienced no blocked practice (i.e., the AI group) were able to declare fewer correct category definitions at the conclusion of the experiment than were any of the other groups, $F(1, 148) = 15.48$, $p < .001$, $\eta_p^2 = .09$. This finding is consistent with the results observed in the blind-block error data during both the testing and transfer phases. Neither contrast involving the partially blocked groups was statistically significant, $Fs(1, 148) < 1$ in both cases. These results, combined with the previous findings, suggest that the inclusion of even a small amount of blocked

Table 4 Explicit learning of nonword category name meaning

| | | Category | | | | | |
		Green	Loud	Fly	Smell	Desert	Underground
Condition	*n*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
High Blocked (HB)	37	1.62 (0.79)	1.69 (0.45)	1.65 (0.74)	1.76 (0.63)	1.28 (0.91)	1.23 (0.95)
Medium Blocked (MB)	39	1.59 (0.82)	1.57 (0.59)	1.71 (0.68)	1.71 (0.68)	1.54 (0.76)	1.35 (0.90)
Low Blocked (LB)	38	1.42 (0.92)	1.62 (0.44)	1.69 (0.69)	1.53 (0.84)	1.36 (0.84)	1.19 (0.93)
All Interleaved (AI)	38	0.74 (0.98)	1.58 (0.61)	1.44 (0.89)	1.44 (0.86)	0.89 (0.90)	0.70 (0.92)

practice during inductive learning of this type has beneficial effects on the acquisition of category concepts.

Discussion

In this study, we found that a blocked format for presenting exemplars was indeed the friend of induction, just as Rothkopf would have foreseen (cited in Kornell & Bjork, 2008). This finding, although it might seem predictable, given the nature of our learning task, is noteworthy because it represents an exception to the current trend favoring interleaved practice for induction. We designed an experiment in which blocking would be expected to facilitate induction so as to test the boundary conditions of facilitation from interleaving, if it is indeed superior to blocking in most learning conditions. Furthermore, the key manipulation—gradually transitioning some participants from blocked to interleaved study—has been proposed recently by researchers in the field (Carvalho & Goldstone, 2015; Dunlosky et al., 2013; Rohrer, 2012), having not been investigated with category induction tasks before, to our knowledge. As such, these findings provide a narrow but potentially important theoretical contribution to the literature.

Our design involved transitioning three groups of participants gradually from blocked to interleaved practice of a moderately challenging induction task, while depriving a fourth group of any blocked practice at all. The most consistent finding across all phases of learning, testing, and transfer was that the participants receiving initial blocked, rather than entirely interleaved, exposure to the category exemplars performed better in both implicit and explicit measures of learning. Although the always-interleaved group had achieved RT parity with the other groups by the final testing and transfer phases, they committed more errors throughout the experiment. Notably, even after the first set of practice blocks, all three noninterleaved groups committed fewer errors than did the AI group on the first blind test (at Block 5). Indeed, it took 12 blocks of practice with the answer key visible for the AI group to achieve the level of blind-test accuracy that the other groups achieved after only four such blocks. The AI group's

tendency to commit more errors than the other groups during learning continued into the final testing phase, which occurred after an interpolated task, and in transfer, when new category exemplars were introduced. Finally, when asked to verbalize the category definitions at the end of the experiment, the AI participants demonstrated poorer explicit understanding of the concepts.

A second finding of interest pertained to the apparent ease with which participants made the transition from the blocked to the interleaved practice format. As expected, participants performed the category-learning trials faster when the exemplars were presented in blocked fashion than when the exemplars were interleaved. And, as participants switched from blocked to interleaved practice, their RTs increased relative to the still-blocked group(s). This result reflected the increased variability of trial content with the interleaving of categories, and one might assume that the prior consistency of blocked practice placed participants at a disadvantage in this switch. However, the average RTs of participants who had just switched to interleaved practice did not differ from those of participants who had always received interleaved practice. Even more telling, switching from blocked to interleaved practice did not increase errors relative to continued blocked practice. Evidently, even 15 blocks of blocked practice in the HB group did not reduce their ability to adapt to the greater demands of interleaved exemplar exposure.

A third finding of interest pertained to the impacts of differing amounts of blocked practice prior to interleaved practice. We assumed that initial blocked practice would facilitate induction, but we also expected that, following some degree of induction, interleaved practice would yield better learning and transfer than continued blocked practice. Several pieces of evidence suggested this was not the case. Blind testing interspersed in the learning trials did not differentiate the three partially blocked groups, regardless of when they had switched to interleaved practice. In other words, greater amounts of interleaved practice after an initial blocked exposure provided no advantage. In addition, no difference was apparent among the three groups with different amounts of blocked practice, in terms of either RTs or errors, during final testing. When the most-blocked group (HB) switched to

interleaved practice (at Block 16, during the final testing phase), they were no slower and made no more errors than the partially blocked groups that had five or ten previous blocks of interleaved practice. Finally, the three partially blocked groups did not differ in transfer performance, despite the differing amounts of interleaved practice that ostensibly should have promoted transfer.

Two related questions emerge from these results. First, why was the all-interleaved practice configuration relatively ineffective for learning and transfer, especially given the recent work by Kornell and Bjork (2008) and others? Second, why did the three partially blocked groups fail to differ from each other, despite the disparities in their practice schedules? More specifically, following the effectiveness of even five blocks of blocked practice, why did subsequent interleaved practice fail to produce better learning, and especially better transfer performance, than continued blocked practice?

To answer the first question, we compare the present task with others according to factors that have been shown to influence the efficiency and potency of inductive category learning. The first factor is the type of category learning in which participants engaged. Our task required participants to induce rules, a type of category learning thought to demand explicit reasoning, tax working memory, and result in categorization decisions that can easily be described verbally; an example would be the rule for distinguishing acute from obtuse triangles (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). By contrast, information-integration categorization is assumed to rely on implicit, procedural-learning-based processes and to result in rules that are difficult, if not impossible, to verbalize. Examples of this type of categorization include the decision rules employed by wine tasters, or "those used by artists to categorize unfamiliar paintings according to the Renaissance master who created them" (Ashby et al., 1998, p. 442). This last example describes a categorization task similar to that used in the research by Kornell and Bjork (2008) and others (Birnbaum et al., 2013; Kang & Pashler, 2012; Kornell et al., 2010; Wahlheim, Dunlosky, & Jacoby, 2011; Zulkiply & Burt, 2013). In these studies, an interleaved format was found to be a more effective than a blocked format, suggesting a possible interaction between categorization type and practice schedule.

The category-type argument is bolstered by the results from other studies that utilized rule-learning categorization tasks and similarly found blocking to be preferable to interleaving (e.g., Carpenter & Mueller, 2013; Goldstone, 1996; Kurtz & Hovland, 1956). The reason for the advantage of blocked presentation with rule learning may be that, when learning explicit rules, participants actively test hypotheses about incoming stimuli by comparing their responses to the feedback received (Hélie, Waldschmidt, & Ashby, 2010; Maddox, Ashby, Ing, & Pickering, 2004). This is a time-consuming, attention-laden process that may be facilitated by blocked presentation of exemplars, because the likely presence of at least

one previous exemplar in working memory allows for such comparison to occur. When the learning task involves information-integration categorization, on the other hand, an implicit, procedurally based learning system is triggered that unconsciously and gradually identifies slight covariations in at least two stimulus dimensions. Feedback processing is assumed to occur almost automatically, perhaps reducing the importance of having previous stimuli in working memory on the next trial. This distinction between learning types is important here, because the varying avenues for categorization may correspond, respectively, to learning processes that emphasize inferences about related features—thought to benefit from blocking—and those that emphasize discrimination processes—thought to benefit from interleaving (see Ashby et al., 1998; Markman & Ross, 2003).

The second factor thought to influence category induction is the level of within- and between-category similarity. Successful induction demands that learners find similarities between exemplars from the same category while discriminating between exemplars from different categories (Zulkiply et al., 2012). A combination of high between-category similarity and low within-category similarity within a stimulus set presents a considerable categorization challenge for participants (Carvalho & Goldstone, 2015), and our learning task incorporated a moderate amount of each. As an illustration of the former characteristic, the exemplar *well* represented the category "underground," whereas the exemplar *oasis* represented the category "desert." It is not difficult to imagine these concepts fitting into the same category, however, and, rather than discovering differences between categories during interleaved presentation, participants may have mistaken differences for similarities and associated *well* with "desert" (though probably not *oasis* with "underground"). As an illustration of the latter characteristic (and as can be confirmed through looking at the Appendix), it is not trivial to induce that the exemplars *lilac*, *locker room*, *rose*, and *outhouse*, even when appearing consecutively, share the feature of having a strong scent. In sum, the degree of combined within-category and between-category diversity of the exemplars in this experiment may have made it difficult to zero in on the category concepts in the always-interleaved format.

A third factor is task difficulty. Interleaved practice is typically held to be more challenging than blocked practice (Schmidt & Bjork, 1992), notwithstanding the notion that the difficulty may be "desirable" (Bjork & Bjork, 2011). When exemplars are blocked, commonalities between them are easier to detect because an immediately preceding exemplar may remain in working memory upon display of a subsequent exemplar. During the interleaved blocks, however, when same-category exemplars were separated by five disparate exemplars, the task became significantly more challenging. Because the participants in the LB, MB, and HB groups were at least repeatedly exposed to sets of exemplars like *alligator*,

frog, *spinach*, and *broccoli* in succession (though not always in this order) at the outset, they had an advantage for determining that ZATCH, the NCN with which these were always paired, represented "green." The participants in the AI group, repeatedly denied the benefit of such an explicit association, failed to induce the category concepts as fully as their peers.

Finally, exposure to a category's exemplars via interleaved presentation may be ineffective when participants' dominant preexisting associations interfere with the formation of the associations needed for concept acquisition. If induction requires the recognition of subordinate features and associations that are shared by multiple, otherwise-diverse exemplars, contiguous exposure may be essential. This explanation reflects Underwood's (1952) premise that the perception of relationships requires contiguity of responses. It also reflects the assumption that blocked but not interleaved exposure allows the relevant stimulus information to remain available in working memory.

With regard to the second question, why more interleaved practice did not benefit learning and transfer if it followed initial blocked practice, we believe the answer also is likely to be linked to the task characteristics. Once the defining feature had been induced for each category in our task (e.g., things that are green, objects underground, etc.), there was little else to be learned from continued exemplar categorization. One might expect additional retrieval practice to improve RTs, but the accuracy of categorization should not, and did not, change appreciably. Even for transfer of the category concepts to new exemplars, the supposed additional difficulty of interleaved practice may have done little to benefit understanding of the relatively simple category definitions. In addition, blocked practice had its own form of interleaving and temporal spacing that may have enhanced its effectiveness. Because the exemplars of each category consisted of diverse objects that had primary associations with other categories, the blocked condition required a degree of variability that might otherwise be absent in most blocked-practice tasks. Furthermore, repeated blocks representing each category constituted a form of spaced practice that may have been beneficial.

One might also ask, if blocked practice was so effective for this task, why was there little or no benefit to experiencing additional blocks of it? This finding may be informed by related research by Rohrer (2009) and colleagues (Rohrer & Taylor, 2006), investigating the spacing effect and overlearning. These authors defined *overlearning* as when a learner continues to study material after it has been learned to some arbitrary criterion—usually one perfect trial. In a verbal-learning study involving the recall of title–author pairings, they found that "overlearning is an inefficient use of study time, and the efficacy of spacing depends at least partly on the degree to which it reduces the occurrence of overlearning" (Rohrer, 2009, p. 1009). Although the participants in our three

partially blocked groups were never "perfect" as a whole, it is possible that, especially for the individuals in the HB group, the category concepts had been successfully induced prior to their switch to interleaved exemplars at final-testing Block 16 due to the combined benefits of blocked presentation and temporal spacing. To the extent to which this was the case for the HB group members, subsequent blocked study constituted overlearning and may have been superfluous, thus explaining the lack of a difference between the groups in the final analyses. This argument is bolstered by the data from the HB group's performance in Set 3 and blind test Block 15. This group alone remained virtually unaffected across this interval that concluded their blocked practice, committing 3.5 % errors and 3.6 % errors, respectively.

Regarding the relevance of the present findings to real-world educational practice, we believe there may be both direct and indirect potentials for applicability. The direct potential for this task paradigm, though it appears to be rather idiosyncratic, is that it could be utilized to teach students rule-laden concepts in a school computer-lab setting. We did, in fact, carry out one such learning activity at a private middle school in a neighboring state. Students were taught the rules for exponents implicitly through multiple blocks of rule–exemplar pairings in this error-free format, complete with the answer key at the top of the screen and periodic blind test blocks. In one 45-min session, sixth- through eighth-grade students quickly attained high levels of accuracy in pairing x^5/x^2 with x^3 (with a feasible foil of x^{10}), for example, and we can imagine the paradigm generalizing to a variety of content areas.

The indirect potential lies in the general finding that a proper matching of information type and presentation sequence should be among teachers' key considerations when they undertake to introduce students to any organizable content. Though teachers may not frame their content as "category learning," per se, component tasks such as analyzing, classifying, sorting, and organizing are integral to many learning activities, and these terms are found throughout the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). When teachers explicitly realize that the content they are about to teach (say, exponents) is rule-laden, placing heavy demands on working memory, they can provide students with a few consistent examples of each rule successively before switching to another rule, to maximize induction. Indeed, introducing a new concept or problem type in this fashion seems commonplace. Among the questions needing further study, then, are how much initial practice of the blocked type is enough, and how does the skill level of students impact the optimal number of repetitions (Dunlosky et al., 2013).

Despite the potential implications of these findings, this study was not without drawbacks. A possibly important way in which our task differed from those of the related studies we

examined was in the two-alternative forced choice recognition response format. Participants had to search for the correct response from among six exhibited in the answer key, but the ultimate decision was between two displayed answers. Kang and Pashler's (2012) participants instead studied paintings that were already correctly paired with the artist's name, and then they were tested by seeing paintings with three artists' names to choose from beneath (3AFC); similarly, Carpenter and Mueller's (2013, Exps. 1, 2, and 4) participants heard the correct pronunciation for French words initially, then were tested using a 3AFC auditory test. Other research had required participants to guess at category memberships during initial learning and to generate one of three categories unaided during testing and transfer (Carvalho & Goldstone, 2014a, 2014b). Still others had directly paired stimuli and category memberships during learning, then had used as many as 13 forced choice alternatives during testing (e.g., Birnbaum et al., 2013; Kornell & Bjork, 2008; Zulkiply & Burt, 2013; Zulkiply et al., 2012). The reduced complexity of recognition demanded by our task may have led to relatively superficial learning and performance.

Another drawback to our design was that the length of the test delay (during the interpolated letter series task) was relatively brief—less than 12 min. This is worth noting because the benefits to memory of temporal spacing are typically attenuated with shorter retention intervals (Cepeda et al., 2006); thus, our findings might have been different—likely offering an advantage to those in the AI condition—had the test delay been longer. Moreover, a longer test delay would better simulate a real-world learning situation. Due to these drawbacks, and the specificity of our manipulation and key variables, including the levels of within- and between-category discriminability, we acknowledge that our findings may best generalize to other tasks of active, rule-based induction in which the number of response choices is limited.

In summary, the results of this experiment suggest that interleaving may indeed be the enemy of induction, as Rothkopf purportedly claimed (Kornell & Bjork, 2008), or, at least that interleaved practice alone may be less than optimal for some forms of inductive category learning. The present study provided evidence that, for rule-based categorization involving disparate exemplars, even a relatively small amount of initial blocked practice may be sufficient to expedite induction of category memberships, and that interleaved practice, in the absence of blocked exposure, impedes learning. These data, although discrepant with some previous evidence, do not suggest that previous evidence for the advantage of interleaving is incorrect. Instead, these data suggest that the relative benefits of blocked and interleaved practice will vary with the nature of the inductive-learning task. Furthermore, they suggest that future research should contrast degrees of blocked and interleaved practice rather than focusing on all-interleaved versus all-blocked comparisons.

Appendix

Categories and exemplars – Set A					
IsLoud	CanFly	Smells	UnderGround	Desert	IsGreen
HEECE	CHILT	SLYNE	BRASK	TROBB	ZATCH
trumpet	biplane	lilac	cave	cactus	alligator
thunder	eagle	locker room	roots	sand	frog
cymbals	sparrow	rose	potato	sun	spinach
siren	glider	outhouse	tunnel	camel	broccoli
Categories and exemplars – Set B					
IsLoud	CanFly	Smells	UnderGround	Desert	IsGreen
HEECE	CHILT	SLYNE	BRASK	TROBB	ZATCH
jackhammer	kite	skunk	worm	mirage	grass
shout	mosquito	rotten egg	well	oasis	lime
scream	sparrow	cinnamon	gopher	thirst	pine tree
dynamite	frisbee	cloves	aquifer	heat	pickle

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481. doi:10.1037/0033-295X.105.3.442

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41,* 392–402. doi:10.3758/s13421-012-0272-7

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41,* 671–682. doi:10.3758/s13421-012-0291-4

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology, 5,* 1–11. doi:10.3389/fpsyg.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42,* 481–495. doi:10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22,* 281–288. doi:10.3758/s13423-014-0676-4

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi:10.1037/0033-2909.132.3.354

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General, 143,* 227–246.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43,* 627–634. doi:10.1037/0003-066X.43.8.627

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego: Academic Press.

Detambel, M. H., & Stolurow, L. M. (1956). Stimulus sequence and concept learning. *Journal of Experimental Psychology, 51,* 34–40.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. doi:10.1177/1529100612453266

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.) New York, NY: Columbia University, Teachers College. (Original work published 1895)

Foss, D. J. (1968). Learning and discovery in the acquisition of structured material: Effects of number of items and their sequence. *Journal of Experimental Psychology, 77,* 341–344.

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

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics, 72,* 1013–1031. doi:10.3758/APP.72.4.1013

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Hillsdale, NJ: Erlbaum.

Hovland, C. I., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology, 45,* 175–182.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26,* 97–103.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19,* 585–592. doi:10.1111/j.1467-9280.2008.02127.x

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25,* 498–503. doi:10.1037/a0017807

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology, 51,* 239–243. doi:10.1037/h0040295

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition, 32,* 582–591. doi:10.3758/BF03195849

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129,* 592–613. doi:10.1037/0033-2909.129.4.592

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 596–606. doi:10.1016/S0022-5371(70)80107-4

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards.* Washington, DC: Author.

Pashler, H., Bain, P.M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (Report No. NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, & U.S. Department of Education. Retrieved from http://ncer.ed.gov

Peterson, M. J. (1962). Some effects of the percentage of relevant cues and presentation methods on concept identification. *Journal of Experimental Psychology, 64,* 623–627.

Reber, A. S. (1992). An evolutionary context for the cognitive unconscious. *Philosophical Psychology, 5,* 33–51.

Rohrer, D. (2009). Avoidance of overlearning characterizes the spacing effect. *European Journal of Cognitive Psychology, 21,* 1001–1012.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24,* 355–367.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20,* 1209–1224.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217. doi:10.1111/j.1467-9280.1992.tb00029.x

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide* (Version 1.1). Pittsburgh, PA: Psychology Software Tools.

Schrepp, M. (1999). An empirical test of a process model for Letter Series completion problems. In D. Albert & J. Lukas (Eds.), *Knowledge spaces: Theories, empirical research application* (pp. 133–154). Mahwah, NJ: Erlbaum.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 179–187. doi:10.1037/0278-7393.5.2.179

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 907–912. doi:10.1037/0278-7393.27.4.907

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70,* 534–546. doi:10.1037/h0043901

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review, 20,* 356–363. doi:10.3758/s13423-012-0319-6

Underwood, B. J. (1952). An orientation for research on thinking. *Psychological Review, 59,* 209–220. doi:10.1037/h0059415

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109,* 163–167.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39,* 750–763. doi:10.3758/s13421-010-0063-y

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15,* 41–44.

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41,* 16–27. doi:10.3758/s13421-012-0238-9

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22,* 215–221. doi:10.1016/j.learninstruc.2011.11.002