Linking Cognitive Science to Education: Generation and Interleaving Effects

Lindsey E. Richland (Lengle@psych.ucla.edu)
Department of Psychology, UCLA
Los Angeles, CA 90095

Robert A. Bjork (RABjork@Psych.ucla.edu)
Department of Psychology, UCLA
Los Angeles, CA 90095

Jason R. Finley (JFinley@ucla.edu)
Department of Psychology, UCLA
Los Angeles, CA 90095

Marcia C. Linn (MCLinn@ucberkeley.edu)
Department of Education in Mathematics, Science, and Technology, UC Berkeley
4611 Tolman Hall
Berkeley, CA, 94720-1670

Abstract

Performance during a learning event is frequently used as a measure of learning; however, basic cognitive research suggests that this may be an unreliable predictor of long-term learning and transfer. Rather, in some training paradigms, higher error rates during training may predict greater retention and generalization of learning. One such paradigm is training in which learners generate, rather than read, during study opportunities. A second is training in which study for two sets of information is interleaved rather than grouped into separate blocks. Educationally relevant learning requires retention and the capacity to generalize information across contexts, leading to the hypothesis that these paradigms may have important implications for educational practice. An experiment is described in which the effects of generation and interleaving are examined using complex, educationally-relevant materials. Findings indicate that these effects are relevant for instructional design, but that there is not a straightforward relationship between laboratory research with simple materials and educational practice. Rather, the educational goal must be considered when determining the utility of generation and interleaving principles in designing instructional technology.

A large body of research on cognition exists in which simple materials have been used to understand the mechanisms underlying learning, retention, and transfer. Such research has led to rich understanding of principles that guide learning in laboratory contexts, but less is known about how these principles generalize to learning of the more complex content typically acquired in educational contexts.

This paper describes a study that links laboratory research on learning with educational practice. In particular, this study explores "desirable difficulty" (Bjork, 1994, 1999) findings that have been well studied in laboratory contexts and that may have important, albeit unintuitive, implications for educationally-relevant learning.

A common strategy for optimizing student learning in classroom contexts is the manipulation of study conditions to increase students' ease of comprehension and to improve students' performance on assessments embedded into the learning task. This is a strategy based on the assumption that performance during learning is a reliable index of knowledge and will predict learning and memory for the material over time. This assumption is made regularly by instructors as well as by learners themselves. Children and adults examine their own performance as a metacognitive strategy when making predictions about their own learning and retention over time (Benjamin & Bjork, 1996; Koriat & Bjork, in press) and when determining how to structure study time or when making decisions about how to optimize learning events (Son, under review).

While the relationship between performance and learning is regularly invoked in both educational practice and research on learning, other cognitive research indicates that this assumption may be misleading. A growing body of laboratory research indicates that increasing the difficulty of a learning activity in systematic ways may impair learners' performance during initial knowledge acquisition but improve learning and retention for this learning over time. Several specific principles have been identified and defined as "desirable difficulties" (Bjork, 1994, 1999). Desirable difficulties are specific types of manipulations of the learning context that increase learners' errors during learning and slow initial acquisition of materials but lead to increased memory for the relevant information over time.

Generation

Generation is an instructional strategy that impairs initial knowledge acquisition and slows learning but leads to increased likelihood of retention for learned information. The principle of generation describes the reliable phenomenon that having learners generate components of a learned stimulus produces better learning and retention that having learners read or re-read study materials. For example, Hirshman & Bjork (1988) found that requiring a learner to generate letters within the second word in a pair such as

bread: bu_tt_ _ results in greater retention for the word pair: "*bread: butter*" than does study in which the learner reads: *bread: butter*. After a five minute delay, participants in the read condition recalled 13% of word pairs in a free recall paradigm, which participants in the generate condition recalled 41%. This striking difference is particularly interesting because the generation tasks are so minimal in this type of a paradigm.

The effects of generation on retention and recall have been well studied and are robust for simple materials such as word pairs. However, the effects of generation on transfer and generalization of learning are not well understood.

Interleaving

Interleaving is a second principle for designing learning contexts in which learning is made more difficult during instruction, but retention and transfer are higher after a delay. Interleaving describes the method of structuring acquisition of two or more learning sets such that instruction alternates between the sets. This is in contrast to blocked instruction, in which learning for each set would be conducted separately. For instance they might complete instruction for one set before moving to the next set for instruction. A learner who is interleaving his/her learning would switch randomly between two sets of information during initial acquisition. The most typical strategy for teaching two sets of similar information (A,B) would be to serially teach set A and then set B. By contrast, interleaving these materials would mean to randomly switch between teaching parts of A and parts of B. So for instance a learning sequence might be: A1B1B2A2A3B3A4B4B5A5. Interleaving has been shown to be more effective for producing learning than blocking or massing learning (A1A2A3A4A5B1B2B3B4B5) on many diverse sets of stimuli including word pairs (Battig, 1979), motor patterns (Shea & Morgan, 1979), and word translations (Richland, Bjork & Finley, 2004).

In a training study, for example, (reference) manipulated training practice schedules for children learning to toss a bean bag into a target location from varying distances. Some children received training in blocks such that they that practiced multiple times from one distance in each block. Other children randomly practiced throwing from varying distances within each block of practice trials. During learning, mean performance for children in the blocked training performed was higher than mean performance for children in the randomized training condition. However, the opposite was true on a final test. After a delay, children were tested on a distance that was practiced by the blocked training group but that was not practiced by the random training group. Interestingly, children in the random training group performed higher overall than the children in the blocked training group – in spite of the fact that those in the blocked training group had practiced on the distance that was tested.

These findings suggest important implications for organizing instruction in more educational settings, potentially indicating that randomly interleaving related instructional materials might impair performance during learning but improve performance at a delay.

Platform: Web-Integrated Science Enrichment (WISE)

The current study explores the generation and interleaving effects using educationally-relevant science materials and an educational software platform used frequently in US classroom contexts. Specifically, an educational module was developed to teach astronomy materials using the Web-Integrated Science Enrichment (WISE) platform developed and maintained by Marcia Linn and colleagues (http://wise.berkeley.edu). This unique software enables researchers to design and modify science educational curriculum in order to systematically manipulate the order and structure of the materials. Further, the platform is highly approachable and is used successfully in thousands of middle and high school classrooms across the United States. Thus studies using the WISE platform can be systematically controlled, yet they are educationally relevant since very similar modules are regularly in use within educational practice. Explorations of interleaving and generation effects using WISE software therefore provide data that is more generalizeable to realistic classroom learning contexts than experiments using simple materials and abstract computer tasks or paper and pencil activities.

As shown in Figure 1, WISE is designed as a set of content pages that are navigated by the learner. The learner scrolls through pages by following the arrows on the left side of the screen. WISE contains many alternative forms of pages, including information pages, as in Figure 1, drawing activities, pages with links to other online materials, and pages in which the learner has an opportunity to study the material they have read about (see Figure 2).
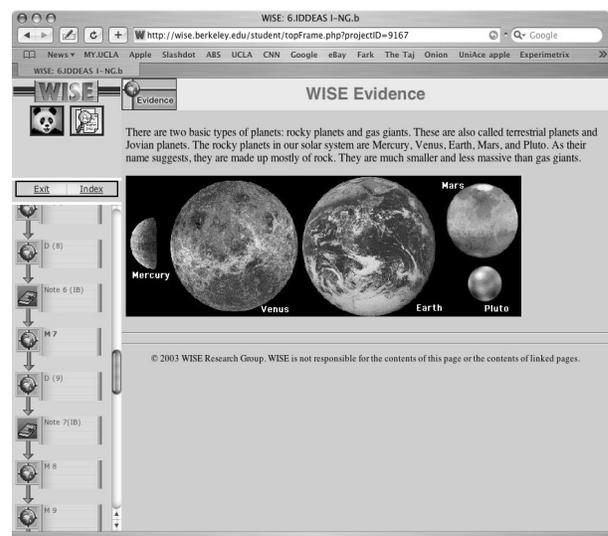

Figure 1. Sample WISE screen.

In the current study, the information pages and the study pages were manipulated to investigate whether generation and interleaving effects generalize from laboratory research

with simple materials to more complex learning with science educational content.
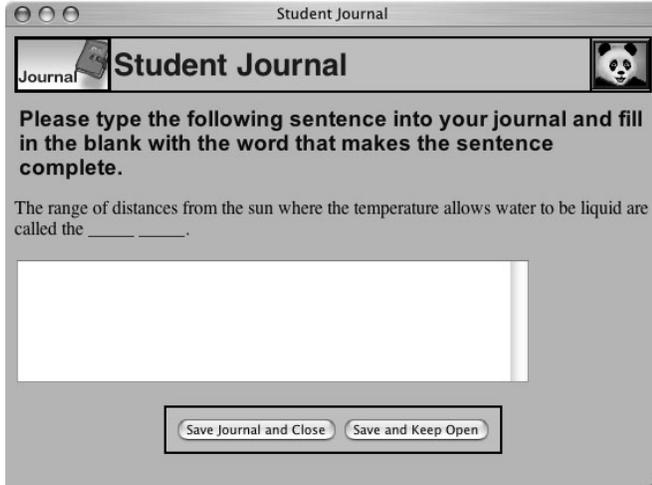

Figure 2. Sample generate condition study page.

Experiment

This experiment was designed to examine whether generation and interleaving effects identified in laboratory research as two principles of desirable difficulty would generalize to educationally-relevant materials. We hypothesized that generation during study would produce increased retention when compared with re-reading during study. Interleaving was hypothesized to produce more knowledge integration and flexible learning than blocking learning, resulting in increased transfer.

Methods

Participants

Undergraduates from the University of California, Los Angeles participated in this experiment for course credit. Data from 96 participants was used in the reported analyses. All participants were given a pretest that elicited information about relevant coursework. If a participant had taken a course on astronomy, their data were excluded from analyses. Participants who did not answer questions during training, or who did not return for the delayed posttest, were also excluded from analyses. 13 participants total were excluded for either relevant college coursework or for failure to complete the entire experiment.

Materials

Four versions of a science instructional module were created in a 2 x 2 between-subjects design. The first variable was *Interleaving*, in which the two learned sets of educational information were either presented in an interleaved order or a blocked order (see below for more details). The second variable was *Generation*, in which a subset of the learning materials was re-studied during instruction either as a generation/ retrieval test or as a re-read task.

As described more fully below in the description of the posttest, three dependent variables were measured after a 48-hour delay. The first was *Retention*, and this was measured through the repetition of sentence completion questions during learning and on the posttest. This measure was only used to assess the *Generation* manipulation. There were also two transfer measures. The first transfer measure was a fairly near transfer, requiring mostly retention: *Single Concept*, which measured transfer to new questions that required information about a single concept. The second measure of transfer was *Concept-Integration*, which measured learners' ability to flexibly apply their learning to new questions that required integrating concepts. The transfer questions included both sentence-completion and multiple-choice questions. Means were taken for each problem and then were averaged to create composite scores for each of the dependent variables.

The overarching theme of the science content used in the instruction was the search for life on other planets. Two distinct sets of information were presented. The sets were organized such that they were mutually informative, but they had distinct educational content. Specifically, the first set of information described the role of a planet's mass on the likelihood that a planet could be inhabited (Mass). The second set of information described the impact of a planet's distance from the sun on its habitability (Distance).

The WISE interface functions such that a learner navigates serially from one web page to a next page in a specified order. As shown in the Figure 1, the participant views a page within a viewer window, and on the left side of the screen s/he sees the order of all the following pages. When the participant finishes reading one page, s/he uses a mouse to click on the next page indicated by the arrows on the left side of the screen. In this experiment, each page had approximately one short paragraph of information and a relevant graphic. The graphics were visual representations of the information and generally did not add separate content.

The two manipulations of the generation variable were conducted using a feature of WISE that enables student participation. Study pages were inserted between 3 and 5 slides after initial instruction in which the participant can enter a response, as shown in Figure 2. These study pages either contained a sentence with one word underlined in which a subset of the information was re-presented, the read condition, or they contained a sentence completion activity, the generate condition. As in Figure 2, in the generate condition a sentence was presented with one word missing. This was the same word as was underlined in the read condition study pages. There were always either three or four pages intervening between the initial presentation of information and the study page.

Posttest

A posttest was developed to test participants' learning for the material presented in the WISE astronomy module. There were three types of questions on the posttest. The first type of question was the most complex and required the learner to

generate facts from both the mass and the distance categories of information. These were both open-ended short answer and fill in the blank questions. The second type of questions required participants to generate multiple facts from either the mass category or the distance category of information. These were both open-ended short answer and fill in the blank questions. The third type of questions were single fact fill in the blank questions. These were the same type as those used during the module as generation or study questions, in order to test retention. All of the questions asked during the module were included on the posttest as fill in the blank generation questions. There were also questions of this same type (single fact, fill in the blank) that were not studied during the learning module.

Procedure

All participants were tested individually on a Macintosh computer. They came for one hour for the learning module, and then they returned two days later for the posttest. There was always a 48-hour delay between learning and posttest. The learning modules were schedule for Monday, Tuesday and Wednesday, while the posttests were scheduled for Wednesday, Thursday, and Friday.

Participants were randomly assigned to one of eight counterbalanced versions of the WISE Astronomy module. There was an instructions page in which they were given directions for navigating the WISE interface by clicking on pages serially following the arrows on the left side of the screen (as shown in Figure 1). There was also a sample study task in which they were told that they would see some questions like this in which they would be asked to type a complete sentence into the space below. The sample note was consistent with the generation condition, so if a participant was assigned to a generation module, they would see a sample question in which one word was blank. If they were assigned to a read condition they would see a sentence with one word underlined.

When participants completed the learning module they were excused and reminded to return at the same time in 48 hours for the second part of the experiment. They were not explicitly told that they would be tested on their memory for the material in the module they had just completed.

One their return, participants were given the posttest in two separate parts. They were first given a paper packet with the complex open-ended and fill in the blank questions that required multiple facts either integrating mass and distance information, or just multiple mass or distance facts. Once they completed this part of the test this packet was removed and they were given the fill in the blank questions. The order of the two parts was held constant because the fill in the blank questions could potentially have served as prompts for facts required to answer the more open-ended questions.

Data

Generation

Participants' accuracy on single fact fill in the blank questions was analyzed in order to examine the generation effect and to assess whether generating facts during the learning module impacted learners' retention for the information over a 2 day delay.

The effects of Generation on Retention were examined through analysis of the sentence completion questions answered during learning as well as on a posttest after a 48-hour delay. During learning, the participants either read a sentence referencing material they had learned previously, or they generated a word to complete the same sentence. Participants were randomly assigned to either condition in a between-subjects design.

As shown in Figure 3, a two-level ANOVA was conducted to examine retention by comparing performance during instruction with performance after a delay, on the posttest. There was a main effect of extra study for the item during learning ($F (1, 58) = 45.2$, $p < .001$) as well as a significant interaction between retention and the generation variable ($F (1, 58) = 34.9$, $p < .001$) such that performance was highest during learning for participants in the read condition, but performance at a delay was highest for participants in the generation condition.

The effects of generation on retention were further examined using a two-level ANOVA to compare participants' performance on posttest items that were not seen during learning with performance on items answered correctly during learning. Generation was included as a between-subjects variable. There was a main effect of extra study on the item during learning ($F (1,58) = 16.0$, $p < .001$) as well as a main effect of generation over reading ($F (1, 58) = 7.2$, $p < .01$). There was also a significant interaction with the generation variable: $F (1, 58) = 23.7$, $p < .001$ such that items that were generated successfully were recalled more frequently after a delay than items that were re-read during study. As noted above, participants retained more of their learning in the generation condition than in the read condition.

For items not generated successfully in the generation condition, there was no difference between performance during learning and final test. In the current study, participants were not given feedback following incorrect generation during learning, so this finding might have been different if learners had been provided with corrective feedback. This suggests that for a classroom context, where retention of all items is optimal, feedback might be essential in designing generation opportunities.
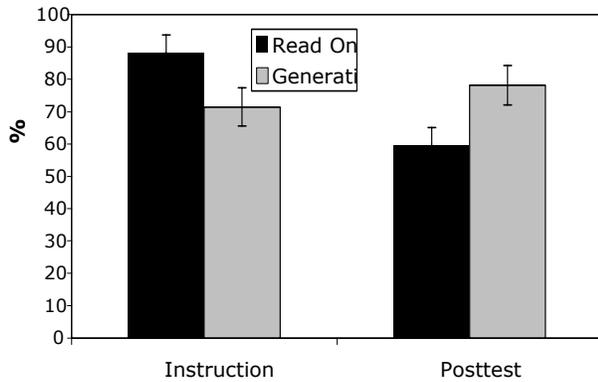
Figure 3. Interaction between Generating versus Reading during study and time of test.

Interleaving

As shown in Figure 3 below, participants' performance on two types of transfer questions from the posttest were compared. Data reveal presentation order (interleaving versus blocking) interacted with type of question, interacting that presentation order led to different results on the three types of questions ($F(1,61) = 5.54$, $p < .05$). On posttest questions that required recall of single facts, or integration of facts with a single domain of information, there was a slight advantage for participants in the blocked condition. On more difficult posttest questions that required more complex integration of information from the two domains, there was an advantage for participants in the interleaved condition.

This interaction suggests that although instruction may have been more difficult when structured in an interleaved format than when structured in a blocked format, participants were more able to integrate their learning in the interleaved condition than in the blocked condition.
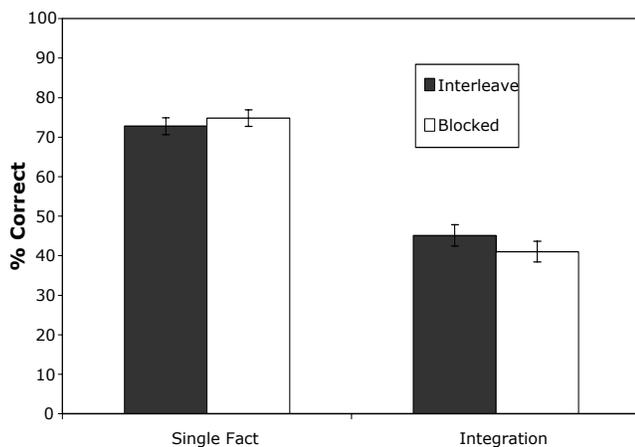

Figure 4. Effects of Interleaving on simple versus integration posttest questions.

Conclusions

These data reveal that principles of generation and interleaving generalize to educationally relevant materials. The manipulation of the Generate variable (generate versus read) study conditions revealed that participants made more errors during learning in the generate condition than in the read condition. However, after a two day delay participants showed the opposite pattern: participants in the generate condition recalled more correct answers than those in the read condition. This interaction within performance suggests that the generation effect impacted learners' retention of these complex science facts when assessments were embedded within the science instruction.

Findings also support the claim that interleaving two related sets of information during learning will produce improved ability to transfer learning to more complex, integrated problems after a delay. Specifically, the data revealed that the important effect of interleaving in these educational materials was the role of question type. Learners were not different on their performance on questions that required recall of single types of facts, but they were different on questions that required more complex reasoning and integration of the two sets of interleaved information.

The goal of most classroom science learning is to teach reasoning and learning that generalizes beyond the specific recall of scientific facts. Thus, this finding suggests that interleaving is a strategy for facilitating complex learning such that interleaved information can be integrated into a representation that allows for considering relationships between the information. However, interleaving may not be the most optimal way to enhance recall of specific fact types of information. The generation effect is more relevant to this type of learning and may be essential to promoting study that facilitates learners' recall of important facts.

Thus, this study demonstrates that the Interleaving effect and the Generation effect are useful for complex educationally-relevant materials, but it also indicates that consideration of the learning goals will dictate which type of principle would be most optimal for instructional design. Retention as a goal suggests that generation can be used to optimize learning. Even though learners might make more errors during learning, their retention for generated items will be higher than their retention for items re-read during learning. However, feedback or alternative supports might be necessary to assist learners with items that are not generated correctly during learning.

If integrated knowledge is a goal, instructors might optimize learners' acquisition by interleaving instructional materials. Although this has the potential to cause confusion for learners by requiring them to switch back and forth between materials, results indicate that this instructional design can lead to integrated conceptual knowledge that can be transferred to questions requiring knowledge integration.

Thus, this study finds that the desirable difficulty principles of generation and interleaving can facilitate instructional

design. Instruction with complex educational materials can build systematically on findings from simple materials in laboratory contexts, and in particular this study found the technology of WISE as an optimal platform for design of science astronomy lesson.

References

Benjamin, A. S. & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed*.), Implicit Memory and Metacognition: The 27th Carnegie Symposium on Cognition.* (pp. 309-338) Hillsdale, NJ: Erlbaum.

Battig, W. F. (1972). Intratask interference as a source of facilitation in transfer and retention. In R. F. Thompson and J. F. Voss (Eds.), *Topics in learning and performance (pp. 131- 159). New York:* Academic Press.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing*. (pp.185-205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher and A. Koriat (Eds*.), Attention and peformance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA: MIT Press.

Hirshman, E. L., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* 14, 484-494.

Koriat, A., & Bjork, R. A. (in press). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Experimental Psychology: Learning, Memory, and Cognition.*

Richland, L.E., Finley, J, R. Bjork, R. A. (2004). Differentiating the Contextual Interference Effect from the Spacing Effect. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*, pp. 1624. Mahwah, NJ: Lawrence 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.

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.

Son, L. (under review). Metacognitive control: Childrens' short-term versus long-term study strategies.