WORK-LEARNING RESEARCH, Inc.

Measuring Learning Results:

Creating fair and valid assessments by considering findings from fundamental learning research

Will Thalheimer, PhD

A Work-Learning Research Publication.

Published April 2007

Questionmark

How to cite this report using APA style:

Thalheimer, W. (2007, April). *Measuring Learning Results: Creating fair and valid assessments by considering findings from fundamental learning research.* Retrieved November 31, 2007, from http://www.work-learning.com/catalog/

Obviously, you should substitute the date on which the document was downloaded for the fictitious November 31 date.

Published April 2007

The Author: Will Thalheimer is a research psychologist specializing in learning, cognition, memory, and performance. Dr. Thalheimer has worked in the corporate-training field, beginning in 1985, as an instructional designer, project manager, product leader, instructor, consultant, and researcher. He has a PhD from Columbia University and an MBA from Drexel University. He founded Work-Learning Research in 1998 with the purpose of helping instructional designers, e-learning developers, trainers, and performance consultants utilize research-based knowledge to build effective learning-and-performance solutions.

Dr. Thalheimer can be contacted at info@work-learning.com or at the Work-Learning Research phone number below for inquiries about learning audits, workshops, high-level instructional design, and consulting on e-learning, classroom training, evaluation, and learning strategy.

Work-Learning Research, Inc.
Somerville, Massachusetts USA
(617) 718-0067

www.work-learning.com

Introduction

Hi. I'm Dr. Will Thalheimer, a researcher and consultant in the field of learning and instructional design. I help people create more effective learning interventions by building bridges between learning research and learning practice. There is wisdom in both camps, but only by integrating research and practice can we maximize our learning outcomes.

In writing this report on using fundamental learning research to inform assessment[1] design, I am combining two of my passions—learning and the measurement of learning. As an experienced learner and learning designer, I have come to the belief that those of us responsible for designing, developing, and delivering learning interventions are often left in the dark about our own successes and failures. The measurement techniques we use simply do not provide us with valid feedback about our own performances.

The traditional model of assessment utilizes end-of-learning assessments provided to learners in the context in which they learned. This model is seriously flawed, especially in failing to give us an idea of how well our learning interventions are doing in preparing our learners to retrieve information in future situations—the ultimate goal of training and education. By failing to measure our performance in this regard, we are missing opportunities to provide ourselves with valid feedback. We are also likely failing our institutions and our learners because we are not able to create a practice of continuous improvement to maximize our learning outcomes.

This report is designed to help you improve your assessments in this regard. I certainly won't claim to have all the answers, nor do I think it is easy to create the perfect assessment, but I do believe very strongly that all of us can improve our assessments substantially, and by so doing improve the practice of education and training.

I would like to thank Questionmark for agreeing in advance of my research and writing to license this report for the benefit of their clients. Questionmark is available on the Web at www.questionmark.com and by phone at 800-863-3950 (North America), +44 (0)20 7263 7575 (United Kingdom) or +32 2 298 02 01 (Europe).

[1] I use the term "assessment" to refer to the systematic gathering of data regarding learning outcomes, usually by requiring learners to demonstrate their ability to answer questions, make decisions, engage in specific tasks, or perform specific skills.

Using Assessments to Gauge Future Performance

Most learning assessments look backward. They are designed to tell us how much someone has previously learned. In other words, they tell us how much someone has learned to-date. This backward-only approach makes it virtually impossible for us to collect valid feedback on the effectiveness of our learning designs. With our current assessment practices, we often fool ourselves about our own performance as designers and facilitators of instruction. This might not be so bad if our assessment outcomes were balanced, but the biases seem heavily weighted in favor of making us look good.

This paper will describe how to design learning assessments that better predict the future—that tell us how well our learners will be able to retrieve the information that they have already learned. The paper will also describe how to avoid the all-too-common biases in our current learning-assessment practices. To put it another way, the ideas in this paper will help you create more valid learning assessments, providing you with better feedback about how you're really doing as a learning professional.

The Benefits of Looking Backward

Before moving forward, let me put my criticism of current assessment practices in perspective. While I don't think the backward-looking approach to assessments is sufficient, it does have some benefits. First, backward-looking assessments can motivate learners to pay attention and to study. This approach is not perfect—as exemplified by the cramming behavior that leads to inadequate memory retention—but it can motivate learners beyond the lower thresholds of learning behavior. Second, backward-looking assessments can be somewhat predictive of future retrieval performance, though the metrics tend to fail to account for the vagaries of forgetting. In other words, backward-looking assessments may be good at evaluating the learning intervention's ability to enable immediate retrieval, but they are poor at evaluating the learning intervention's ability to minimize forgetting.

In summary, backward-looking assessments provide some value, but that value comes at significant cost, including the cost of promoting inadequate learning behaviors and ineffective learning designs.

What do Assessments Measure?

Human memory is not like computer memory. With computers, inputs become outputs—items that go into computer memory are retrieved in an identical form. The human memory system doesn't work that way. Information that can be retrieved today may not be retrieved tomorrow. It may not be fully retrieved. It may not be accurately retrieved. It may be retrieved in some situations but not others. It may be retrieved in the presence of some cues, but not others. Human memory—especially the process of retrieval—depends on many factors.

Let's look at an example. Suppose we are running a seminar on the history of buggy whips. The seminar takes place on April 1st, running for six hours in our organization's blue seminar room (from 9 o'clock in the morning until 3 o'clock in the afternoon). We develop an extensive assessment to measure the results of the learning. The assessment is delivered between 3 o'clock and 4 o'clock. It asks our learners to retrieve specific information about the history of buggy whips.

What, then, does our assessment measure? Does it measure the learners' knowledge of buggy whips? Does it measure their ability to retrieve information about buggy whips? It probably provides some measure of each of these constructs. However, when we fully consider the workings of human memory, we can utilize a more precise metric. So, to be precise, our one-hour assessment measures the ability of our learners to retrieve specific information about buggy whips as they sit with their fellow learners in the blue seminar room on April 1st between three and four in the afternoon.

While this wordy description may seem gratuitous, it is not. Each of the contextual elements described (for example, "in the blue seminar room") will affect our learners' ability to retrieve. Here's a short list:

- If they sat in a different room during the assessment, they would probably retrieve less of what they had learned.

- If they sat with different learners, they may retrieve less of what they had learned.

- If they took the assessment two hours later, a day later, or a month later, they would probably retrieve less of what they had learned.

- If they were asked questions that differed from the questions presented on the beginning-of-the-day pretest, they would probably retrieve less of what they had learned.

- If they were asked questions that used words that differed from the words used in the seminar—even if those words were synonymous with the words used—they would probably retrieve less of what they had learned.

- If they were asked questions that raised or lowered their state of anxiety in comparison to the level of anxiety during the seminar, they would probably retrieve less of what they had learned.

Even without any changes to our buggy-whip assessment, our learners' scores will be positively correlated with their ability to retrieve information about buggy whips in other situations and at other times. However—and this is a key point—when we take human learning and memory into account, we can build assessments that are significantly better in being predictive of the situations that our learning interventions are designed to support. By having more-predictive assessments, we can get better feedback on our own performance as creators of instruction. So, for example:

- If we develop a workshop designed to help people administer CPR over the coming year if faced with a stopped-heart emergency, we can build a more predictive assessment of their ability to remember what to do by delaying the assessment one week instead of providing the assessment immediately at the end of the workshop.

- If we develop a history course to help our learners be better citizens in a democracy, we can build a more predictive assessment by getting rid of questions that ask about past events and instead use questions that ask the learners to decide how to apply what they've learned to current policy debates.

- If we develop a course to teach Microsoft Excel, we can build a more predictive assessment if we provide the assessment on the computer instead of on paper.

- If we develop a course to teach statistics concepts to 10th graders to help them perform well in subsequent classes and real-world situations, we can build a more predictive assessment if we avoid hinting about the types of problems asked. To be specific, a more predictive assessment will not label the problem types as t-test, ANOVA, or regression problems, as such hints will not always be available in future courses or real-world uses.

What influences retrieval? Here's a short list:

- The passage of time makes it less likely that learners will be able to retrieve the information they learned.

- The more the retrieval context mirrors the learning context, the better the retrieval.

- The more learners receive retrieval practice during learning, the better the subsequent retrieval.

- The more learners focus on the relevant aspects of the learning material, the better the subsequent retrieval of critical information.

- The more the learners practice making decisions and taking actions in realistic situations, the better their retrieval in real-world situations.

- The longer the duration between learning events that reinforce the same learning points, the better the subsequent long-term retrieval.

For our assessments to capture these characteristics of the human learning system, they must be designed in appropriate ways. Unfortunately, a fair percentage of the assessments we currently use in education and training fail to account for these fundamental human learning factors. Here’s a short list of the potential issues:

- Assessments are provided to learners only at the end of learning, not after a delay.

- Assessments don’t capture our learning interventions’ ability to minimize forgetting and enable future retrieval.

- Assessments are provided to learners in the same context in which learning took place.

- Assessments include items that are of secondary relevance or importance.

- Assessment items measure memorization, not decision making or performance.

- Assessment items do not mirror the learners’ future retrieval contexts.

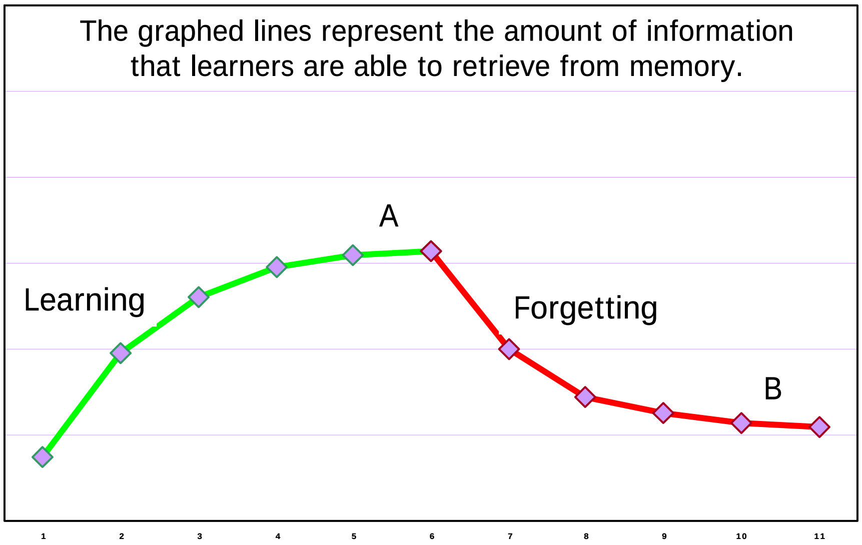
- Assessment items inadvertently hint at correct answers through the surface characteristics included in the items.

Our current assessment designs often only measure our learners’ ability to retrieve in the assessment situation—at that specific time and place, and in response to specific assessment-item cues. While such designs may be adequate as a way to motivate a modest level of learner intensity and to enable us to assign grades, they are not sufficient in giving us valid feedback about how well our learning designs are preparing our learners to retrieve information in important future situations. To get feedback on our ability to produce this future retrieval, we have to build assessments that can better measure future retrieval.

Measuring the Potential for Long-Term Remembering

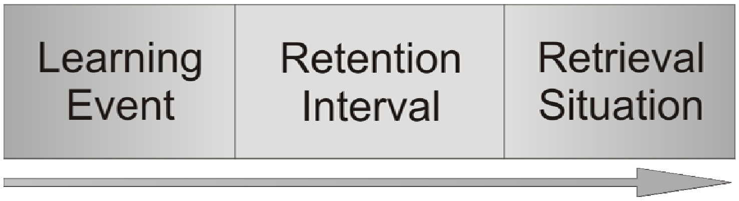
Why is it so critical that assessments measure the potential for future retrieval? Look at the graph below, representing the typical learning and forgetting curves. As our learners learn a body of knowledge, they gradually improve in their ability to retrieve information. However, if the information is not utilized after a learning event, learners gradually lose their ability to retrieve the information.

When the Learning is Not Utilized



Looking at the graph above, if we measure retrieval immediately at the end of learning (Point A), the results will not accurately reflect retrieval performance at the lower ends of the forgetting curve (Point B). In situations like these—in which the learners don't immediately utilize what they've learned—measuring retrieval at the end of learning (Point A) produces an inflated prediction of future retrieval. When we measure retrieval at the end of learning, we lie to ourselves about our performance. We also forgo the opportunity to get valid feedback so that we can improve our instructional designs.

Not all of our learning interventions will suffer the learning-forgetting curves illustrated above, but most will. The problem is this: almost all learning activities are followed by significant "retention intervals" in which forgetting occurs. Retention intervals begin at the end of a learning event and continue until retrieval is required. Look at the diagram on the following page.



Retention intervals can range from five minutes to fifty years or more. The longer the retention interval, the more forgetting will occur. Even relatively short retention intervals enable forgetting. So, for example, if our learners participate in a day-long workshop on Monday—even if they learn something that they'll be able to apply the next morning, their retention interval will still be about 16 hours or so (from 5 in the afternoon on Monday until 9 on Tuesday morning). A lot of forgetting can occur in that period.

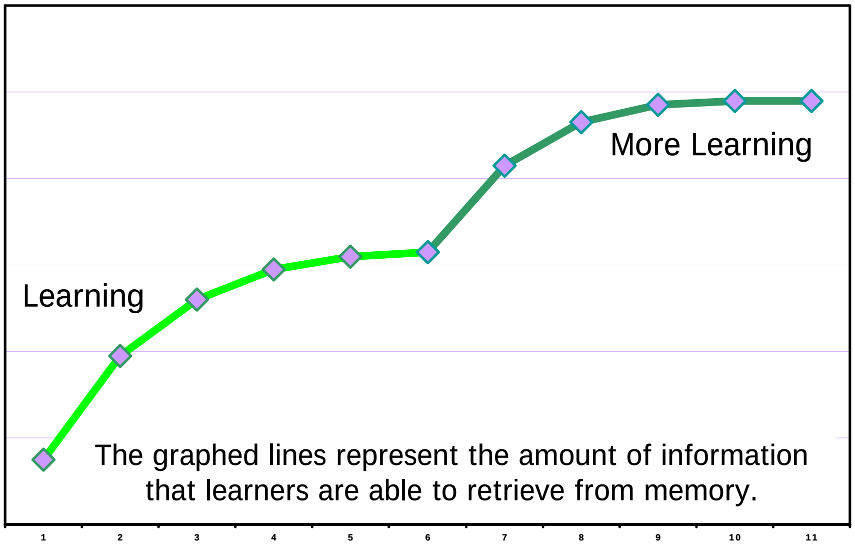
Typical training and education scenarios tax learners' memories even further. Our learners learn a lot of information, but only utilize some of that information soon after learning. For example, employees who learn how to handle emergencies will only face some of those emergencies in the first month after the training. Students who learn statistics may run 20 t-tests soon after their course ends, but may use few regressions, and zero factor-analyses. Programmers who take an online course may use some of what they learned later on the same day, but other information they learned may not be required for six months. Managers who practice interviewing techniques may not use what they have learned for two months, or may use it right away and then not for six months.

Even so-called just-in-time learning can suffer from the retention-interval problem. Suppose an employee can't figure out how to accomplish a task, so he accesses his company's knowledge management system to get help. It takes him 15 minutes to find the right information. Immediately after seeing a video on how to do the task, he implements the solution. So far so good: there are no retention-interval problems in this just-in-time scenario. However, eight months later the employee has to accomplish the same task again—and try as he might, he can't remember what to do. Those eight months represent a debilitating retention interval, especially if every eight months the employee has to spend 15 minutes searching for the correct information—even supposing he remembers that it was the system that helped him solve the problem previously[2].

[2] Knowledge-management systems can be set up to remember the information that learners previously utilized, but users have to remember to use the system. There are also negative learning effects of searching through inappropriate material, causing memory-interference issues. There are also negative productivity effects of not remembering, as the learner is likely to waste time before accessing the knowledge management system. Of course, it may be appropriate that low-priority information is forgotten, enabling relatively higher-priority information to remain accessible for retrieval. The point still remains—sometimes just-in-time learning insufficiently prepares learners to remember.

Ideally, we hope that our learning interventions create something like the graph below, where the learners utilize what they've learned soon after the learning event:

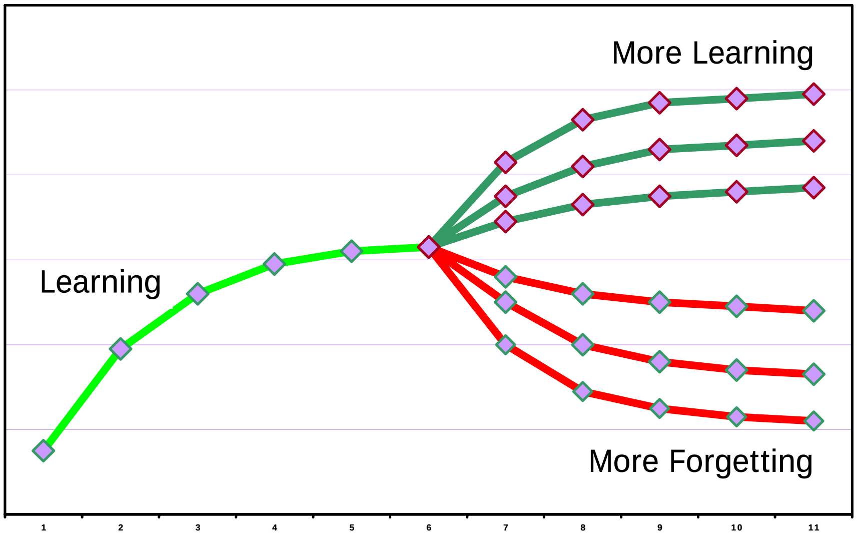
When the Learning is Utilized Afterwards



The graphed lines represent the amount of information that learners are able to retrieve from memory.

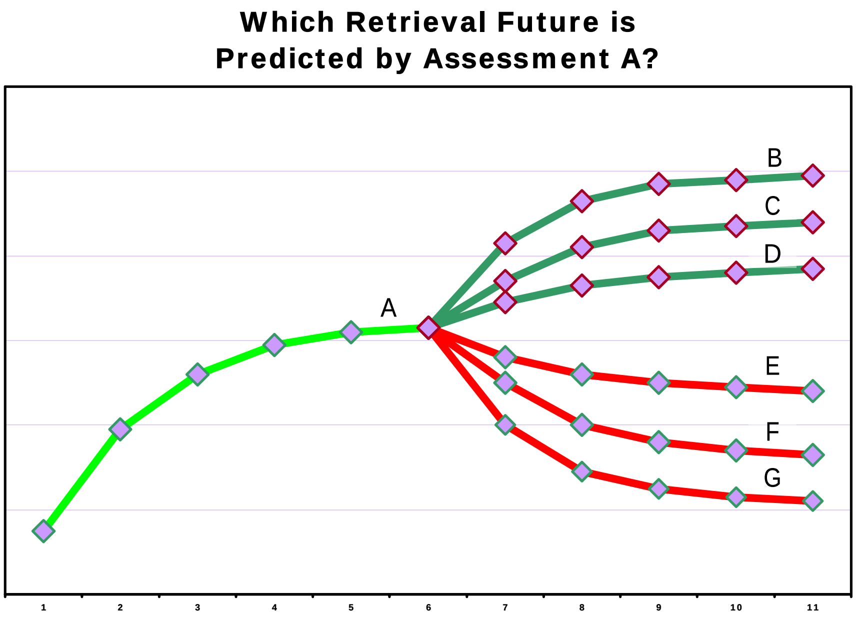
However, the best we can typically hope for is the following more realistic set of curves, where some of the learned information is utilized soon after learning and some is not:

When Some is Utilized & Some Is Not



As we discussed earlier, forgetting is more likely than the additional learning, because so much of what we learn is not utilized routinely.

The graphs depicted on the previous pages show how complicated learning assessment can be. Retrieval performance can rise, fall, or stay the same after the learning events end. One thing is certain: measuring retrieval at the end of the learning event is not necessarily a reliable predictor of future retrieval.


Which Retrieval Future is Predicted by Assessment A?

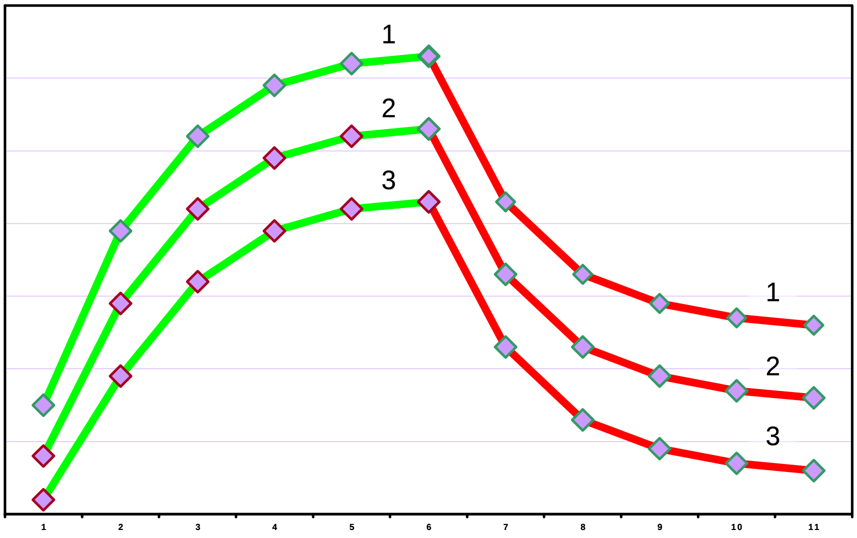
If we design an assessment and deliver it at Point A in the graph above—at the end of the learning event—we may be misleading ourselves about our learners' future ability to retrieve what they have learned.

Since one of our primary goals in developing learning events is to enable long-term retrieval, shouldn't our assessments be reasonably predictive of long-term retrieval?

Learning Methods Differ in their Ability to Minimize Forgetting

If learning methods didn't differ in their ability to minimize forgetting, the above worries would melt away. We could simply measure retrieval at the end of learning and we'd have the best correlate available with which to predict future retrieval. By knowing the top point of the forgetting curve, we could predict any future point. For example, in the graph below, the top forgetting curve is 10 points higher than the middle forgetting curve, and the middle curve is 10 points higher than the bottom forgetting curve. The ten-point difference is true at each and every point on the curves.

If Learning Methods Did NOT Differ in Minimizing Forgetting

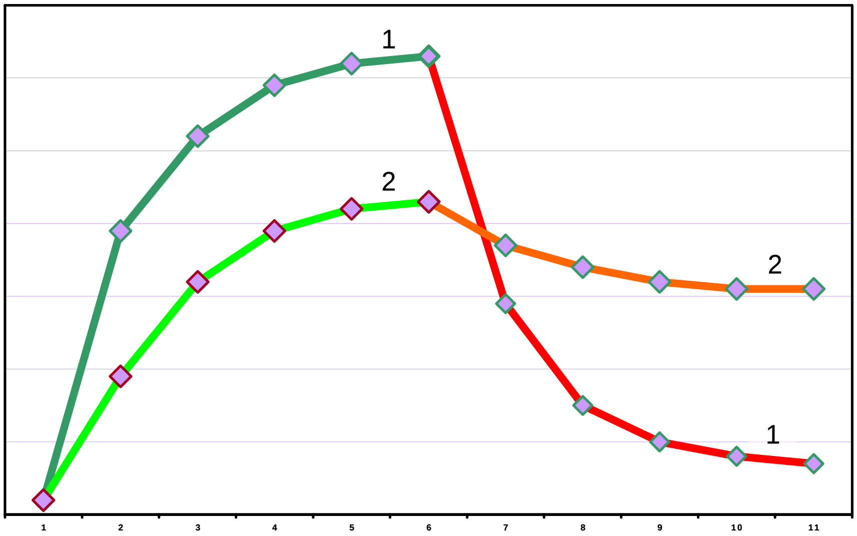


Suppose further that Learning Method 1 produced the top curve and Learning Method 3 produced the bottom curve. If we measured them both at the end of learning, we would find that Learning Method 1 produced an assessment score 20 points higher than Learning Method 3. Based on this end-of-learning assessment, we would know that Learning Method 1 would produce higher levels of retrieval at every point in the future.

Such simplicity would make decision-making easy. It would also make end-of-learning assessments a reliable tool for predicting future retrieval. Unfortunately, human learning is not so simple. Different learning methods produce different forgetting profiles, as you can see in the graphs on the following page.

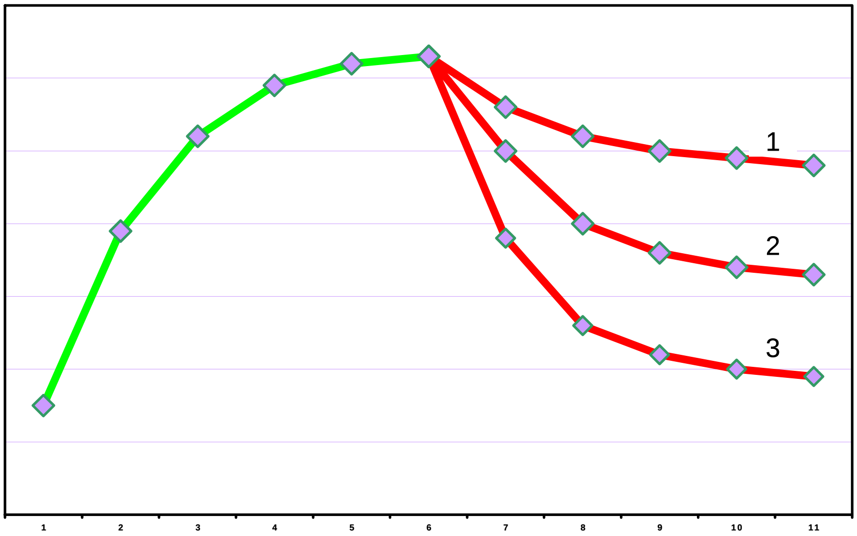
The following graph depicts learning and forgetting curves for two distinct learning methods. Note how Learning Method 1 would produce better end-of-learning assessment scores than Learning Method 2, but would produce much poorer retrieval performance in the long run.

Which is the Better Learning Intervention?



In the graph above, Learning Method 1 prompts massive forgetting, while Learning Method 2 does a nice job in minimizing forgetting. As may be obvious, an end-of-learning assessment would give us very poor information about future retrieval performance. The following graph shows similar difficulties.

Differences in Forgetting for
Three Different Learning Methods



Do Learning Methods Really Produce Different Forgetting Curves?

Yes, learning methods do produce different forgetting curves. The most obvious example may be the *cramming effect.* If you have ever crammed for a test, you probably remember the differential effects it had in the short and long terms. Cramming produces good short-term retrieval and poor long-term retrieval, creating a very steep forgetting curve. Cramming's opposite, *spaced learning*, is particularly good at minimizing the forgetting curve, and is backed up by loads of research (for example, see Ebbinghaus, 1885/1913; Bahrick & Hall, 2005; Bruce and Bahrick, 1992; Donovan & Radosevich, 1999; Lee & Genovese, 1988; Ruch, 1928; Cain & Willey, 1939; Melton, 1970; Crowder, 1976; Hintzman, 1974; Glenberg, 1979; Rea & Modigliani, 1988; Dempster, 1988, 1989; 1996). To read an overview of the spacing effect, see my research report, *Spacing Learning Events Over Time*, available at www.work-learning.com/catalog/.

But the spacing effect is not the only learning method that produces different forgetting curves. Delaying feedback can produce similar improvements in forgetting (see for example, Kulhavy & Stock, 1989; Sassenrath & Yonge, 1968, 1969; English & Kinzer, 1966; More, 1969; Sturges, 1969, 1972; Phye & Andre, 1989; Kulhavy & Anderson, 1972).

Aligning the learning context and the future retrieval context can also produce long-term benefits (for reviews, see Bjork & Richardson-Klavehn, 1989; Smith, 1988; Smith & Vela, 2001; Eich, 1980; Roediger & Guynn, 1996; Davies, 1986). By providing learners with experience in realistic situations—for example, simulations and scenario-based decision exercises—we increase the likelihood that they will be able to retrieve information from memory when they encounter analogous situations on the job or in future learning events. Aligning contexts in this way can have differential effects on short-term and long-term retrieval processes because, in the short term, the information stored in memory is relatively easy to retrieve; thus, retrieval is less likely to require the support of specific contextual cues. In the long term—after forgetting processes have made retrieval more difficult—contextual cues are particularly important in enabling successful retrieval.

Providing learners with variable practice during learning—as opposed to consistent practice—also produces very good long-term retrieval, while it often depresses short-term retrieval (for reviews, see Lee, Magill, & Weeks, 1985; Van Rossum, 1990). When we provide learners with a variety of retrieval cues in this way, they become more prepared to notice relevant cues in future retrieval situations.

To summarize this section, learning methods can produce different retrieval effects in the short term and in the long term. Because end-of-learning assessments only measure short-term effects, using them to predict long-term retrieval is of dubious merit (Ghodsian, Bjork, & Benjamin, 1997).

There is No Perfect Assessment Design

Unfortunately, there is no one-size-fits-all, always-appropriate assessment design—even in specific well-traveled situations. Tradeoffs would still be required even if we had limitless time and resources. In the real world, where binding constraints can severely limit what is possible, we often have to make difficult decisions that reduce the validity and reliability of our assessments.

While we can't build perfect assessments, we can build better ones. This report is designed to help you think deeply about the tradeoffs involved in assessment design. Only by understanding how the fundamentals of learning and memory relate to assessments will you be able to make intelligent choices about which tradeoffs will allow you to reach your assessment goals.

The rest of this document will describe specific methods that can help us avoid the most obvious difficulties presented by the inherent characteristics of human learning.

Avoiding End-of-Learning Assessments

One way to avoid the problems inherent with end-of-learning assessments is to avoid using them. Instead of giving learners an assessment immediately at the end of learning, we could simply give them an assessment a week or two later. In theory, this might give us a more accurate picture of our learners' future ability to retrieve what they have learned. The closer in time an assessment is to the actual retrieval situation, the more closely the assessment results will mirror the actual real-world results.

On the other hand, there are complications. For example, if our learners study for the delayed assessments, then we are not really measuring the potency of the original learning events. We are measuring the effects of the original learning events plus the effects of the additional studying. Surprising our learners with delayed assessments might also make us rather unpopular, especially if those assessments are onerous.

Finally, waiting until later to assess retrieval can make it difficult, if not impossible, for us to diagnose later performance problems (Coscarelli, 2007). For example, suppose we (a) give our learners training on May 1st, (b) assess retrieval on May 15th, and then (c) find that the training is having no effect on May 30th. If retrieval is okay on May 15th, we can rule out retrieval problems. However, if retrieval on May 15th is inadequate, we can't know whether the training program was insufficient in creating understanding or whether it was insufficient in supporting long-term retrieval, or both.

Augmenting End-of-Learning Assessments

We could augment end-of-learning assessments by providing a second assessment after a delay. This would (a) provide us with two data points, (b) indicate whether retrieval is improving or falling and by how much, and (c) provide at least one delayed assessment to capture information about the effects of our learning intervention on forgetting.

Although using an end-of-learning assessment and an additional delayed assessment has a lot of appeal, we may still have difficulty ruling out the effects of additional study. We may still create learner frustration by compelling the additional effort. Finally, doubling the number of posttests we utilize requires additional effort.

Avoiding the Problems of Context

Retrieval is the process of bringing information from long-term memory into working memory. In some sense, our primary goal as learning professionals is to enable our learners to retrieve information in the future—or more specifically, to retrieve the right information in the right situation at the right time.

The retrieval process doesn’t happen in a vacuum. Retrieval is prompted by the contextual cues that people face. For example, the question, “What is the capital of Massachusetts?” is a retrieval cue for “Boston.” Being presented with a realistic decision scenario on a management issue could be a retrieval cue for what was learned in a recent supervisory-skills workshop. Being presented with a geometry problem can be a retrieval cue for a whole host of skills and knowledge learned in a recent geometry class, including metacognitive problem-solving skills and specific geometry rules and theorems. Hearing a political leader defend a vote may act as a retrieval cue for what was learned in school about government. Seeing a fallen coworker could be a retrieval cue for what was learned in a recent CPR course.

As may be obvious, this process of contextually-cued retrieval doesn’t relate only to formal learning. It happens every minute of every waking hour. When we are in a particular situation, the stimuli in that context remind us of what we’ve learned and experienced.

Research has shown very clearly that learners will retrieve more information from memory if they try to retrieve that information in the same room in which their learning took place (e.g., Smith, Glenberg, & Bjork, 1978). Similarly, when scuba divers learn underwater, they recall more underwater than nearby on land (e.g., Godden & Baddeley, 1975). When people learn during a time when they are sad, they’ll remember more when they’re sad, and vice versa (e.g., Bower, Monteiro, & Gilligan, 1978; Eich, 1995; Smith, 1995). When college students learn with loud noise as a background, they do better on tests when those tests are accompanied by loud noise; silent studying improves performance during silent test-taking as well (Grant, Bredahl, Clay, Ferrie, Groves, McDorman, & Dark, 1998)[3]. For reviews, see Bjork & Richardson-Klavehn, 1989; Smith, 1988; Smith & Vela, 2001; Eich, 1980; Roediger & Guynn, 1996; Davies, 1986.

These varied results demonstrate that context—whether environmental, emotional, or physiological—can provide cues that aid retrieval of learned information. The research also suggests clear recommendations to designers of learning and learning assessments:

1. To maximize future retrievability of learned information, the learning context should mirror or simulate the future retrieval context.

2. To maximize the validity of our assessments to predict future retrieval, the items given in our assessments should utilize contexts that mirror or simulate the future retrieval contexts.

3. Because the context of learning affects future retrievability, some learning methods will be better than others in producing future context-generated retrieval.

4. Because the context utilized by assessment items can be more or less aligned with future retrieval contexts, some assessment items will be better than others at predicting future retrieval.

I may be straying away from the central theme of this document by including recommendations for learning design, not just assessment design. I include these recommendations because they are useful on their own, and, once again, to highlight that different learning methods produce different long-term outcomes—requiring us to go beyond end-of-learning assessments.

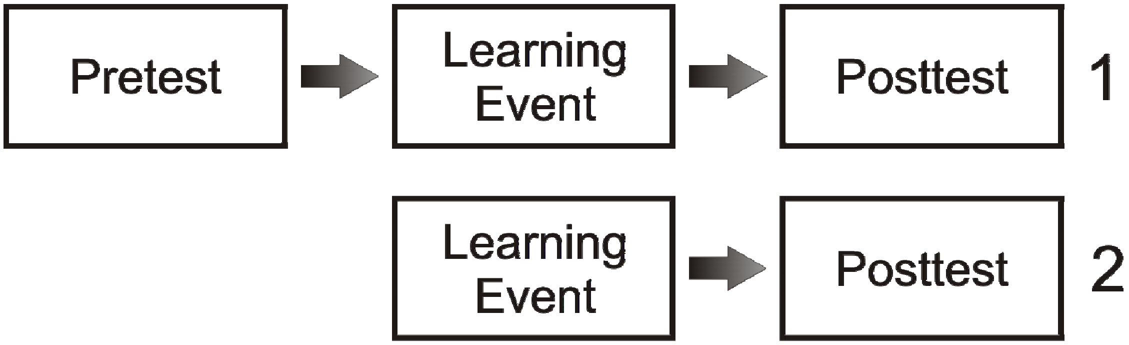
The bottom line for assessment design is to ensure that the questions are relevant to your learners' future retrieval situations.

Avoiding the Problems of Prequestions

Prequestions can produce powerful learning benefits, helping learners know where to focus their attention as they encounter learning material. Unfortunately, this same potency can bias assessment results. If we give learners a pretest and then a posttest, the posttest results will reflect the benefits of the learning AND the benefits of the pretest. Certainly, the following assessment design will be unfairly biased against Group 2—the

[3] This context-alignment effect is a robust effect, having been found in many circumstances. Here are some more examples. When people learn under the influence of alcohol or marijuana, they recall more when tested under the influence, and vice versa (e.g., see studies reviewed by Eich, 1980). When people learn while smelling peppermint, they retrieve more information when smelling peppermint than when smelling osmanthus, and vice versa (Herz, 1997). If people have learned while listening to Mozart, they retrieve more of the learned information while listening to Mozart than they do while listening to jazz (Smith, 1985).

group that doesn't get a pretest. This is true whether the learners get feedback on the pretest or not. There are significant benefits to simply asking learners questions.

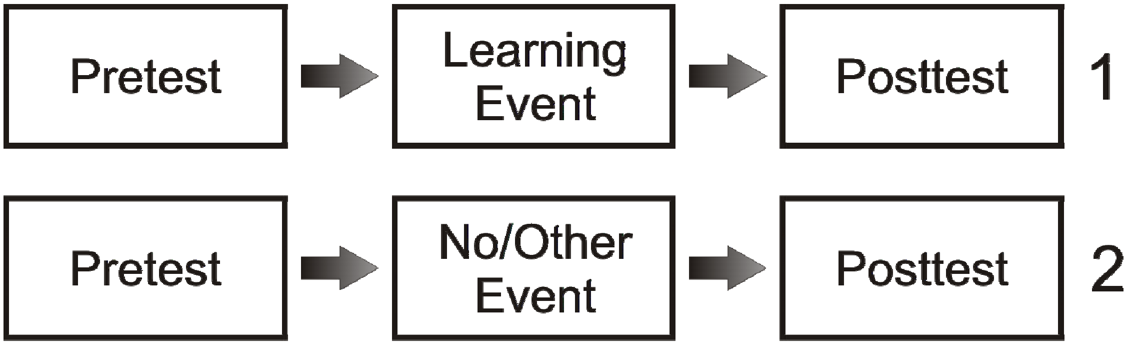


One simple solution to this is to include a pretest in the design of our learning intervention. In other words, the pretest becomes part of the learning intervention, so any benefits of the pretest are benefits inherent in the learning design.

Another way to limit bias due to prequestions is to avoid using them. This, of course, will make it difficult—if not impossible—for us to show that our learning intervention improved learning results.

We can also ameliorate the attention-focusing effects of prequestions by providing them well in advance of our learning events. For example, if you give learners prequestions a month before they begin their learning, the prequestions are unlikely to bias the results.

We can also compare our learning event to a situation where learners get no learning event or where they get another learning event. As illustrated below, this will get rid of bias associated with a pretest. Of course, this design is valuable to the extent that the two learning events are meaningfully comparable.



Regardless of which of the above pretest-posttest designs is used, it is critically important to determine the likelihood that the prequestion's surface characteristics—as opposed to the prequestion's essential learning content—could trigger remembering. For example, suppose you use the following as a prequestion:

> **Question: You're an executive coach and one of your clients is Joe. Joe works for Apple Computer and is a manager in the marketing department. Joe calls you up and tells you about an exciting new idea he has for a marketing campaign. He tells you that he spent a couple of hours developing a presentation so that he could flesh out his ideas. He now wants to get his team on board, but he wants your help in determining the best way to proceed. What will you advise him to do?**
>
> A. Call a meeting immediately to show your team the idea. Direct them to get started implementing the idea.
>
> B. Call a meeting immediately to show team idea. If they like the idea in principle, go to your boss to get her input.
>
> C. Go to see your boss to get her impression. If she likes the idea in principal, then call team meeting to get started.

Can we use this prequestion in its verbatim form on the pretest and the posttest? We can if and only if (a) we didn't give the learners feedback on this question, or if (b) the time between the pretest and posttest is sufficiently long to make it unlikely that the learners would remember the information incorporated into the question[4]. Otherwise, the learners may be able to use the surface characteristics presented in the question to remind them of the correct answer. For this specific question, the name "Joe," or company "Apple Computer," or the title "marketing manager" may remind the learner and make it more likely they'll get the answer correct. Such reminding doesn't have to be completely potent or even conscious to bias the assessment results. Even if it gives learners a 10% better chance of getting the correct answer, the assessment is biased.

This problem can be fixed by creating two versions of each assessment item, each with different background information. Instead of "Apple Computer" it could be "Fidelity Investments". To be sure that the pretest and posttest are equally difficult, you can pilot-test both by asking a group of people to take both tests and then comparing their responses on the paired questions. Similarly, you can randomly assign each member of your question pairs to the pretest or posttest.

[4] Bill Coscarelli told me that he and Sharon Shrock (his co-author for the book *Criterion-Referenced Test Development*) do not recommend the reuse of pretest questions on posttests, and I suggest you consider their wisdom. In situations where logistics or resources make this ideal difficult, I am comfortable reusing pretest questions when no feedback is given or when the length of time between pretest and posttest is sufficient to promote massive forgetting. Note, too, that it is even safer to reuse pretest questions when we use no feedback *AND* we allow time for the pretest questions to be forgotten.

Avoiding the Problems of Inauthentic Assessment Items

Authentic assessment items provide learners with realistic scenarios, requiring them to make decisions about what to do. There are obvious gradations of authenticity. The most authentic assessment items will put learners in the real-world situations they will face after the learning ends. The most inauthentic will ask trivial questions about trivial information.

Expanding on the excellent work of Shrock and Coscarelli (In press, 2007)[5], the following list provides a reasonable approximation of levels of authenticity, starting with the highest level of authenticity and ending with the lowest level of authenticity[6].

A. Real-World Performance
B. High-Fidelity Simulations
C. High-Fidelity Decision-Making Scenarios
D. Low-Fidelity Simulations
E. Low-Fidelity Decision-Making Scenarios
F. Memorization of Critical Information
G. Memorization of Perfunctory Information

Unfortunately, many assessments require performance only at the memorization levels. By assessing only low-level information at the memorization level, we fail to evaluate more important competencies. Moreover, we bias our assessments away from competencies that would more authentically predict real-world performance. In addition, focusing on the retrieval of low-level information can bias our results in both directions—making our learning interventions look good or bad, depending on the questions asked.

If our memorization questions focus on esoteric information, low-level assessments can make our learners and our learning interventions look worse than they actually are. For example, if we ask our supervisory trainees the definition of the term “Work Breakdown Structure,” the results will be biased against those who focused their attention on what to do rather than on the terminology presented.

On the other hand, if low-level questions are relatively easy compared with the actual retrieval performance required, the results will make the learners and the learning intervention appear better than they are. For example, if we ask a police recruit whether a

[5] Shrock and Coscarelli (In press, 2007) use the following categories: Level A—Real World, Level B—High fidelity simulation, Level C—Scenarios, Level D—Memorization, Level E—Attendance, Level F—Affiliation. While these categories are perfect for use in assessment certification decisions, I modify them because more gradations are helpful in talking about authenticity of assessment items.

[6] While the list puts high-fidelity decision-making scenarios above low-fidelity simulations, this ordering could be reversed depending on the level of fidelity of each in comparison with what is most important for the learner to be able to do.

handgun is a pistol or a rifle, the results will suggest competence when the recruit may not even know how to clean his gun, let alone use it in a real-world situation.

Of course, memorization is not always inauthentic. Where memorized retrieval is the goal, memorization questions provide very authentic assessment. For example, if we are preparing our learners to pass a drivers' test that will require them to know how many feet it takes a car to stop on dry pavement after the brakes are applied at 50 miles per hour, then a memorization-level question about stopping distance is very authentic. Certainly, it is *not* authentic to the task of driving, but it is authentic to the goal of the instruction, which in this case was passing the paper-and-pencil portion of the driver's test. Authenticity depends on the goal of learning.

Low-level memorization questions are dangerous for another reason, as well. They send a clear message to learners about what is important. In so doing, they prompt learners to focus on low-level information instead of focusing on more relevant information.

To summarize the recommendations from this section, the more our assessment items move up the authenticity hierarchy listed above, the more they will assess the performance we really want to measure.

Going Beyond Measures of Retrieval

Measuring retrieval is what we typically do when we give assessments, but it is not the only method available. In the employee-training field, in 1959, Donald Kirkpatrick proposed his now-famous four-level model of training assessment.

- Level 1 – Reaction: What the learners thought of the training.
- Level 2 – Learning: Whether the learners learned from the training.
- Level 3 – Behavior: Whether job behaviors changed as a result of the training.
- Level 4 – Results: Whether business results improved because of the training.

Under Kirkpatrick's system, retrieval is represented in Level 2, the level that measures learning. I present Kirkpatrick's formulation not because it is perfect or all-encompassing—it is not. In fact, it is particularly problematic for education situations. I present it to highlight the fact that retrieval is not all that can be measured.

Here is a more expansive, yet not exhaustive, list of learning-assessment options:

1. Learner satisfaction with the learning experience.
2. Learner estimation for how much they have learned.
3. Learner estimation of how much they will use the information.
4. Learner after-learning report on the usefulness of the learning experience.
    - What information have you used?
    - What information have you found helpful?
    - What information did you not find helpful?
    - What obstacles have you encountered to using what you learned?
    - Have you shared what you learned with others?
5. Learner retrieval at end of the learning event.
    - Recognition, Cued recall
    - Decision making, Scenario-based decision making
6. Learner retrieval after the learning event has ended.
    - Recognition, Cued recall
    - Decision making, Scenario-based decision making
7. Learner after-learning performance (requiring retrieval and application)
8. Learner's environment after-learning performance
    - Team performance and satisfaction
    - Family performance and satisfaction
    - Organizational unit performance and satisfaction
    - Organization performance and satisfaction
    - Community performance and satisfaction
    - Etcetera, expanding to larger or smaller units of analysis
9. Learner propensity for future learning
    - Measuring how the learning intervention enables the learner to learn about the targeted topic in future situations.

Depending on your mood or temperament, the list above can be either daunting or exhilarating. I include it to reiterate the point that retrieval is not the only assessment option.

One warning. Asking learners their opinions about learning events is fraught with difficulties. Learner smile-sheet ratings, even when learners are asked to assess the value of the learning, have been found to correlate at a very low level (with an *r* of less than .2) with learners' ability to retrieve information or apply that information (Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997). When learners are asked to predict how much they'll be able to remember, they are typically overconfident about their ability to retrieve information (Zechmeister, & Shaughnessy, 1980). Measuring retrieval is a much better method than looking only at learner opinions.

Measuring the after-learning effects of learning can be particularly valuable, though doing so is often logistically difficult. The advantage of measuring after-learning effects is pretty obvious when we ask ourselves the following questions:

- What happens if our learners can retrieve the information they learned but they can't apply it?

- What happens if they can retrieve the information and apply it, but that their actions cause harm?

- What happens if our learners can retrieve the information but they then have difficulties learning new information about the topic?

- What happens if our learners can retrieve and apply the information they learned, but by so doing, it hurts team morale and performance?

The list of questions could be expanded, but the point is obvious. Retrieval is only the first step on our learners' way to successful application of learning. It is a necessary step, but it is not sufficient on its own. If we really want to understand the full effects of our learning interventions, we need to go beyond the measurement of retrieval.

On the other hand, because retrieval is necessary to enable application—and because it is so much easier to measure than actual application—retrieval is often our best practicable measurement. Knowing its strengths and limitations is helpful because it enables us to design our retrieval assessments to maximize their authenticity and relevance. This report has been designed with that intent—to provide you with a deep understanding of how to design authentic and relevant retrieval-based assessments.

Final Recommendations

The following list of recommendations is designed as wisdom to be considered, not as a recipe to follow.

1. Figure out what learning outcomes you really care about. Measure them. Prioritize the importance of the learning outcomes you are targeting. Use more of your assessment time on high-priority information.

2. Figure out what retrieval situations you are preparing your learners for. Create assessment items that mirror or simulate those retrieval situations.

3. Consider using delayed assessments a week or month (or more) after the original learning ends—in addition to end-of-learning assessments.

4. Consider using delayed assessments instead of end-of-learning assessments, but be aware that there are significant tradeoffs in using this approach.

5. Utilize authentic questions, decisions, or demonstrations of skill that require learners to retrieve information from memory in a way that is similar to how they'll have to retrieve it in the retrieval situations for which you are preparing them. Simulation-like questions that provide realistic decisions set in real-world contexts are ideal.

6. Cover a significant portion of the most important learning points you want your learners to understand or be able to utilize. This will require you to create a list of the objectives that will be targeted by the instruction.

7. Avoid factors that will bias your assessments. Or, if you can't avoid them, make sure you understand them, mitigate them as much as possible, and report their influence. Beware of the biasing effects of end-of-learning assessments, pretests, assessments given in the learning context, and assessment items that are focused on low-level information.

8. Follow all the general rules about how to create assessment items. For example, write clearly, use only plausible alternatives (for multiple-choice questions), pilot-test your assessment items to improve them, and utilize psychometric techniques where applicable.

References

Alliger, G. M., Tannenbaum, S. I., Bennett, W. Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50,* 341-358.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*, 566-577.

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.) *Current Issues in Cognitive Processes: The Tulane Floweree Symposium on Cognition* (pp. 313-344). Hillsdale, NJ: Erlbaum.

Bower, G. H., Monteiro, K. P., & Gilligan, S. G. (1978). Emotional mood as context for learning and recall. *Journal of Verbal Learning and Verbal Behavior, 17,* 573-585.

Bruce, D., & Bahrick, H. P. (1992). Perceptions of past research. *American Psychologist, 47,* 319-328.

Cain, L. F., & Willey, R. (1939). The effect of spaced learning on the curve of retention. *Journal of Experimental Psychology, 25,* 209-214.

Coscarelli, W. (2007). *Personal communication.* April 6, 2007.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Davies, G. (1986). Context effects in episodic memory: A review. *Cahiers de Psychologie Cognitive, 6*, 157-174.

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.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1,* 309-330.

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, CA: Academic Press.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don’t. *Journal of Applied Psychology, 84,* 795-805.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*, (Translated by H. A. Ruger and C. E. Bussenius). New York: Teachers College, Columbia University. (Also available 1964 and 1987, New York: Dover Publications. Original published in 1885).

Eich, E. (1995). Mood as a mediator of place dependent memory. *Journal of Experimental Psychology: General, 124*(3), 293-308.

Eich, J. E. (1980). The cue dependent nature of state dependent retrieval. *Memory and Cognition, 8,* 157-173.

English, R. A., & Kinzer, J. R. (1966). The effect of immediate and delayed feedback on retention of subject matter. *Psychology in the Schools, 3*, 143-147.

Ghodsian, D., Bjork, R. A., & Benjamin, A. S. (1997). Evaluating training during training: Obstacles and opportunities. In M. A. Quiñones & A. Ehrenstein (Eds.) *Training for a rapidly changing workplace: Applications of psychological research* (pp. 63-88). Washington, DC: American Psychological Association.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing and repetitions on recall and recognition. *Memory & Cognition, 7,* 95-112.

Godden, D. R., & Baddeley, A. D. (1975). Context dependency in two natural environments: On land and underwater. *British Journal of Psychology, 91,* 99-104.

Grant, H. M., Bredahl, L. C., Clay, J., Ferrie, J., Groves, J. E., McDorman, T. A., & Dark, V. J. (1998). Context-dependent memory for meaningful material: Information for students. *Applied Cognitive Psychology, 12,* 617-623.

Herz, R. S. (1997). The effects of cue distinctiveness on odor-based context-dependent memory. *Memory & Cognition, 25*(3), 375-380.

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). Potomac, MD: Erlbaum.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63,* 505-512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279-308.

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Different effects for discrete and continuous tasks. *Research Quarterly for Exercise and Sport, 60,* 59-65.

Lee, T. D., Magill, R. A., & Weeks, D. J. (1985). Influence of practice schedule on testing schema theory predictions in adults. *Journal of Motor Behavior, 17*, 283-299.

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.

More, Arthur, J. (1969). Delay of feedback and the acquisition and retention of verbal materials in the classroom. *Journal of Educational Psychology, 60*, 339-342.

Phye, G. D., & Andre, T. (1989). Delayed retention effect: Attention, perseveration, or both. *Contemporary Educational Psychology, 14*, 173-185.

Rea, C. P., & Modigliani, V. (1988). Educational implications of the spacing effect. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.) *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 402-406). New York: John Wiley & Sons.

Roediger, H. L., III, & Guynn, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (eds.), *Memory* (pp. 197-236). San Diego, CA: Academic Press.

Ruch, T. C. (1928). Factors influencing the relative economy of massed and distributed practice in learning. *Psychological Review, 35*, 19-45.

Sassenrath, J. M., & Yonge, G. D. (1968). Delayed information feedback, feedback cues, retention set, and delayed retention. *Journal of Educational Psychology, 59*, 69-73.

Sassenrath, J. M., & Yonge, G. D. (1969). Effects of delayed information feedback and feedback cues in learning on delayed retention. *Journal of Educational Psychology, 60*, 174-177.

Shrock, S., & Coscarelli, W. (in press). *Criterion-Referenced Test Development (Third Edition)*. San Francisco: Pfeiffer.

Smith, S. M. (1985). Background music and context-dependent memory. *American Journal of Psychology, 98*, 591-603.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies & D. M. Thomson (eds.) *Memory in Context: Context in Memory* (pp. 13-34), Chichester, UK: Wiley.

Smith, S. M. (1995). Mood is a component of mental context: Comment on Eich (1995). *Journal of Experimental Psychology: General, 124*(3), 309-310.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8*, 203-220.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory and Cognition, 6,* 342-353.

Sturges, P. T. (1969). Verbal retention as a function of the informativeness and delay of informative feedback. *Journal of Educational Psychology*, *60*, 11-14.

Sturges, P. T. (1972). Information delay and retention: Effect of information in feedback and tests. *Journal of Educational Psychology*, *63*, 32-43.

Thalheimer, W. (2006, February). *Spacing Learning Events Over Time: What the Research Says.* Retrieved March 21, 2007, from http://www.work-learning.com/catalog/.

Van Rossum, J. H. (1990). Schmidt's schema theory: The empirical base of the variability of practice hypothesis: A critical analysis. *Human Movement Science, 9(3-5)*, 387-435.

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.