JUDGING THE CREDIBILITY OF WEBSITES: AN EFFECTIVENESS TRIAL OF THE SPACING EFFECT IN THE ELEMENTARY CLASSROOM

VANESSA LAUREN FOOT

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

August 2019

Abstract

Spaced learning—the spacing effect—is a cognitive phenomenon whereby memory for to-be-learned material is better when a fixed amount of study time is spread across multiple learning sessions instead of crammed into a more condensed time period. In an educational context, this means that long-term retention is enhanced when students begin to review subject material several days leading up to a test instead of cramming right before the test. The spacing effect has been shown to be effective across a wide range of ages and learning materials, but no research has been done that looks at whether spacing can be effective in real-world classrooms, using real curriculum content, and with real teachers leading the intervention. The current study was the next step in determining whether spacing can and should be implemented across the curriculum. Lesson plans for teaching website credibility was distributed to homeroom elementary teachers with specific instructions on how to manipulate the timing of the lessons for either a massed (one-per-day) or spaced (one-per-week) delivery, and after one month, students were asked to apply their knowledge on a final test, where they evaluated two new websites. Students in the spaced condition remembered more facts from the lessons but showed no spacing advantage on the critical thinking measures where they had to explain their ratings in a paragraph. There was no difference in the actual rating scores during the lessons or at final test. These results indicate that when lesson plans are released to homeroom teachers, variability between teachers and classrooms may result in an overall reduction or elimination of a traditional spacing effect. Future recommendations for spacing studies are made. Keywords: *spacing, distributed practice, critical thinking, website evaluation.*

Acknowledgements

First and foremost, thank you to the principals, teachers, and students who, despite their crazy workloads, took the time to participate in this research. If you know a teacher, you may already know that it's not an easy gig. It's one of the only jobs where you have to plan every second of the day before it starts, and there is a shockingly low probability that the day will ever go as planned. Not only are teachers responsible for acting in-loco-parentis, but they handle all of the social, emotional, and intellectual needs of their students, while simultaneously delivering large amounts of curriculum content. There are so many tips and tricks that cognitive psychology offers that teachers could implement in the classroom, but for some reason, many teachers don't know about them. It has become my professional goal to show teachers that there are evidence-based strategies that they can use to make them more effective teachers—and their students more effective learners—by doing *less* work.

Thank you to my supervisor Melody for always being supportive, for mentoring me but always letting me chart my own path, and for constantly making me feel like I have superpowers. Thank you to her wife, Sandi, for taking interest in my work and always reminding me of its value. Thank you also to my lab mates for your support—Tina, Annalise, Katie, Justeena, Vanessa, and all of my wonderful study volunteers. To my mom June (a retired teacher), for contributing to this research in 2015 and for being the sounding board for every hurdle that I faced, and to the rest of my family, Aaron, Tala, Aunt Marion, and my best friend Sarah who all helped to push me when I needed it. A special thank you to my dad George who is no longer with us but still somehow encourages me to "get that A+". Last, but definitely not least, thank you to my incredible husband Mike, my cheerleader who inspires me to be my best every day, and to our future baby boy, who literally gave me the kick I needed to get this paper finished.

Table of Contents

List of Tables

List of Figures

Introduction

The goal of the learner is seemingly simple—he or she strives to acquire new knowledge so that they can understand, retain, and retrieve it when needed. Classrooms are full of students who are trying to do just that. But each year, they remember some concepts and forget others. Teachers then are given the task of figuring out which students remember what, how well they remember it, whether they can apply it, and how much they will need to re-teach it to bring students back up to mastery before adding on new knowledge that builds on existing content. This can be a daunting task for any teacher.

In order to ensure that students are on track with their learning goals, teachers are required to follow current curriculum guidelines, which include expectations and outcomes for their students during and at the end of each school year. Although teaching the content listed in these guidelines remains a central task for teachers, there are still unanswered questions about how to *implement* the content to enhance student retention.

The Ministry of Education in Ontario is the body responsible for designing the different subject curricula that are delivered to teachers. When designing a curriculum document, there are many aspects of learning that need to be addressed. Content is of the utmost importance, as are the aims and objectives for learning, assessment, and educational strategies for implementation (such as problem-based learning). Roles for parents, teachers and students are discussed, as are best practices for delivering the many expectations that are listed as both specific and overall goals. However, a review of the most current curriculum documents (Ministry of Education, 2005) demonstrate that although the curriculum provides some further guidance and information to support its implementation (especially when it comes to modifying and accommodating programming for students), there is no direct connection to strategies from the field of cognitive

science which may help to boost retention (those that have been recommended by the National Council of Teacher Quality, 2016). The NCTQ is an American research and policy group that was founded by the Thomas B. Fordham Institute, who serve to overhaul education and challenge its current system. One of the documents produced by the NCTQ recommended that in order to boost student retention, teachers should use any and all of the following strategies: pairing graphics with words; linking abstract concepts with concrete representations; posing probing questions; interleaving problems; assessing students; and distributing practice (i.e., spacing out review sessions). Their recent review of hundreds of relevant teacher education textbooks demonstrated that almost 60% of these texts fail to mention even one of these six fundamental instructional strategies, and those who do fall short in explaining that strategy properly (NCTQ, 2016). They claim that, "textbook publishers and authors are failing the teaching profession, students and the public by neglecting to provide our next generation of teachers with the fundamental knowledge they need to make learning stick" (NCTQ, 2016, p. 30).

The NCTQ's review also found that some strategies *are* being used in the classroom but adjusting timing of lessons via practice to boost retention is still a relatively unexplored and underutilized area (Harden, 1999). The NCTQ's recommendation for distributing practice is below (Figure 1). Practice is a standard part of most teachers' lesson plans, seen most commonly in homework review and daily activities such as the *Daily Five* (Boushey & Moser, 2014), where students rotate through math and literacy centers on a regular basis until they have mastered the required skills. A detailed explanation of this topic is beyond the scope of this paper. However, the caveat is that practice is not the only required piece of the puzzle. Depending on the interval

between instruction and practice, the *timing* of practice can have massively different effects on student learning and retention.

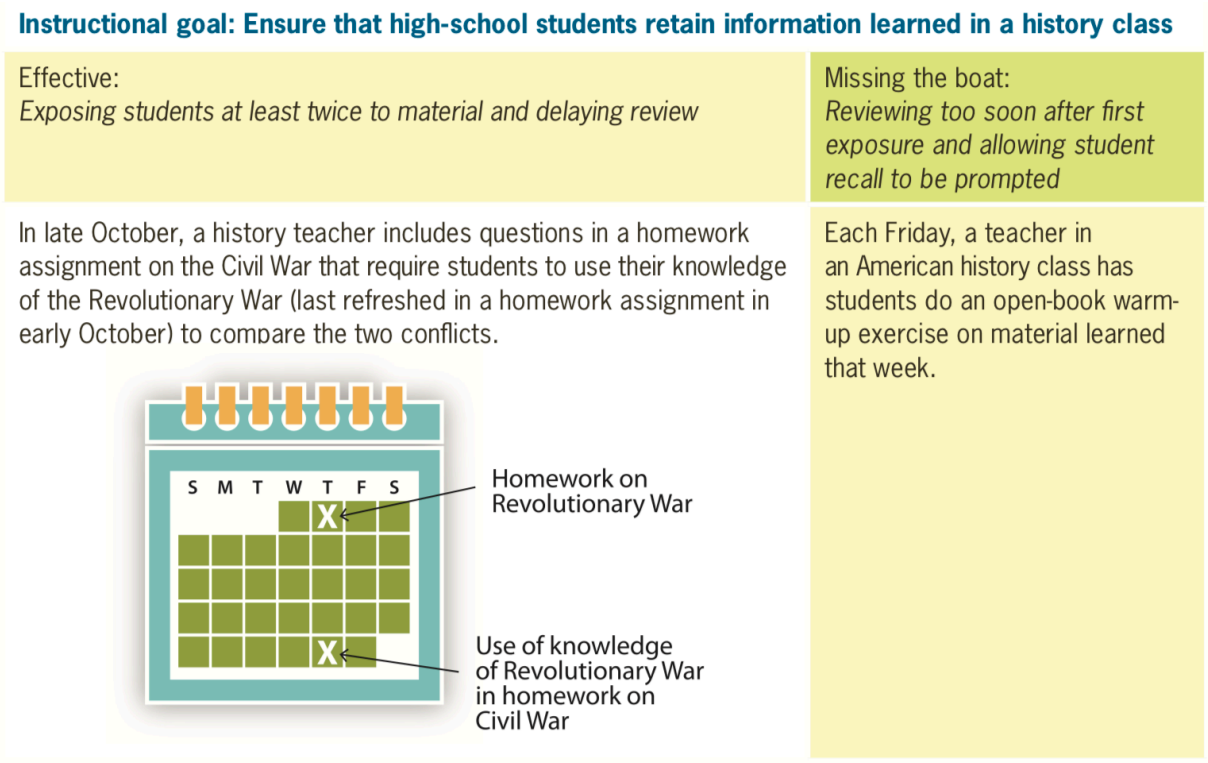


Figure 1. Distributed practice example from the NCTQ (2016) document, "Learning about Learning: What Every New Teacher Needs to Know."

Thankfully, if teachers want to adjust timing of practice to boost retention, there is an abundance of scientific data on how to implement a successful distributed practice intervention (for a review, see Wiseheart, Kupper-Tetzel, Weston, Kim, Kapler, & Foot-Seymour, 2019). However, a barrier to wide-scale implementation is that there is insufficient evidence to show that systematically modifying lesson timing is worth the extra effort that it may entail. Therefore, the focus of this dissertation is to (1) discuss the vast existing literature on the spacing effect, (2) point out the holes in the existing literature that may be the reason for the delay in school implementation (i.e., lack of existing classroom studies and a focus on fact learning), and (3)

attempt to fill those holes with a new, large scale effectiveness study on lesson timing in the classroom, using both fact and critical thinking materials.

The Spacing Effect

In the psychology literature, the *spacing effect* (often referred to as the *distributed practice effect*), refers to the boost in retention that occurs after newly-learned information is relearned or restudied across multiple smaller chunks of time, as opposed to learned once in a longer chuck of time. Given equal amounts of time spent studying, spacing has been shown to boost long-term memory (for a review, see Cepeda, Pashler, Vul, Wixted & Roher, 2006).

In a typical spacing paradigm, the learner is given some new information to memorize (e.g., a list of words). Learners are often separated into two groups: massed and spaced. The massed learner spends some time learning the information, and then restudies (practices) the same information in repetitive blocks (occurring one immediately after the other with little or no time in between). The spaced learner is given the same information to learn but instead of having the blocks appear one immediately after the other, he or she is given some time in between the blocks before restudying (this is called the *inter-study interval).* In the literature, these are also called *practice* or *review* sessions, because in order to use spacing effectively, repetition of the same items is key (as seen as early as Ebbinghaus, 1885/1964). After equal amounts of time following the final learning episode have passed for these two learners (called the *retention interval)*, they are tested on the information to see how much they remember (Figure 2).

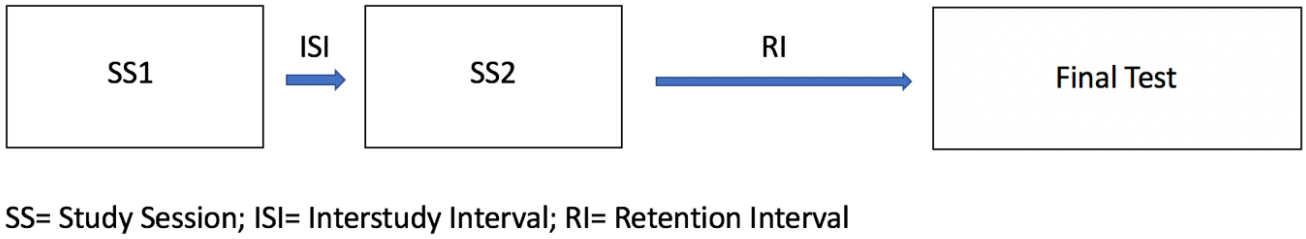


Figure 2. A traditional spacing design with two study sessions.

How long should the intervals be between study sessions (lessons) in order to maximize retention and pursue mastery of concepts? The answer is that it depends on how long an individual is required to remember the information (before a test, for example), or in some cases how long before it is presented again and extended upon. Over the years, spacing researchers have demonstrated that there is a relationship between the interstudy interval (the time one spends between restudying the information) and retention interval (how long someone needs to remember it) (Glenberg, 1979), and that there is an optimal interstudy interval that one can choose in order to be able to access the information later on with greater success (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008).

An example of this might be in school when studying new language vocabulary (e.g., French) for an end-of-unit test occurring one month after the new vocabulary has been presented initially. In order to maximize retention, students should review the words with an optimal amount of time in between. According to the data provided by Cepeda et al. (2008), if a learner wants to recall something after a month has passed, the optimal spacing schedule will be about one week. If a person wants to retain information for a longer period of time (e.g., several years), they could do that by adjusting this interval, delaying review for a longer long-term memory store which could potentially double the amount remembered (compared to a less temporally distributed study scale). Although there are costs to using a gap that is longer than the optimal

value, because forgetting may happen, it is still better than having a temporal gap that is too short (Cepeda et al., 2008).

Spacing Benefits to Learning

Spacing has deep roots in cognitive psychology, starting with Ebbinghaus (1884/1964), who noticed that when study sessions were spaced apart in time (as opposed to massed together), it was easier to retrieve the information. Since then, researchers have demonstrated its reliability over and over again. Many reviews and meta-analyses have been conducted on the spacing effect (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010; Dempster, 1996; Janiszewski, Noel, & Sawyer, 2003; Kupper-Tetzel, 2014; Maddox, 2016; Toppino & Gerbier, 2014; Wiseheart et al., 2019). These reviews explain that spacing has been shown to improve memory for many different types of content, such as basic vocabulary (Bloom & Shuell, 1981), random facts (DeRemer & D'Agostino, 1974), textbook concepts (Reder & Anderson, 1982), word lists (Zechmeister & Shaughnessy, 1980), addition (Reed, 1924), multiplication (Rea & Modigliani, 1985) and geometry (Rohrer, 2009; Rohrer & Taylor, 2007; Taraban, Rynearson & Stalcup, 2001). Spacing effects have been seen across age groups. Benefits have been seen in infants (Rovee-Collier, Evancio, & Earley, 1995), elementary school children (Carpenter, Pashler, & Cepeda, 2009; Foot-Seymour, Foot & Wiseheart, 2019; Sobel, Cepeda, & Kapler, 2011), high school students (Bloom & Shuell, 1981; Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014), and healthy adults, including older adults (Cepeda et al., 2008; Simone, Bell, & Cepeda, 2013).

Reported effect sizes for spacing in the verbal literature are the largest ($d = 0.85$: Cepeda et al., 2006; Moss, 1995), and accumulating evidence suggests that the magnitude of spacing effects may depend on type of content or the skill that is being learned (for a full review, see

Wiseheart et al., 2019). For example, the estimated effect size for non-verbal realms is predicted to be a bit lower (d = .5) based on the studies that have been conducted using this type of material (Foot-Seymour et al., 2019; Gluckman, Vlach and Sandhofer, 2014; Kapler, Weston and Wiseheart, 2015; Vlach & Sandhofer, 2012).

Spacing in the Classroom

The K-12 school year in Ontario runs from September to June, with an 8-week summer break. When students return to the classroom in September, the extent of forgetting is often strikingly obvious to teachers (Allinder, Fuchs, Fuchs, & Hamlett, 1992). One early study showed that 8-weeks was enough for knowledge of psychology course content to drop from 62% to 23% (Gustav, 1969). If students are to remember what they learned the year previous, they need to be repeating it throughout the school year in order to maximize their chances of retaining previously learned concepts.

Though spacing is well recognized by the scientific community, why don't teachers know about it? Why haven't specific recommendations been made? This could be because spacing studies have traditionally focused on teaching rote memorization in controlled laboratory settings, which is extremely different from the wide array of learning that takes place in the classroom. As well, there aren't yet enough classroom studies to provide sufficient evidence supporting the spacing effect in an educational context. Some of the applied classroom-based studies that have been done with verbal/factual material show spacing benefits of word and phonics learning (Seabrook, Brown & Solity, 2005), word and fact learning (Carpenter et al., 2009; Sobel et al., 2011), second language learning (Bloom & Shuell, 1981; Küpper-Tetzel et al., 2014), and text comprehension (Rawson & Kintsch, 2005; Verkoeijen, Rikers & Ozsoy, 2008). These studies all showed benefits of spacing.

Previous classroom studies have shed some light on the difficulties of planning and conducting real-world classroom research, and why this type of research might be so scarce in the literature. In a spacing study by Foot-Seymour et al. (2019), there was a delay in the starting of the research due to a school board-wide strike, and then when the classroom learning began, lessons were interrupted by several snow days, fire drills, and alternate class programming. The noise of the classroom setting might be an intimidating place for researchers to venture, especially with studies that require more rigorous control. Regardless of these points, the research needs to be done, and the spacing effect has been such an area of interest over the past century that it is now time to investigate whether spacing benefits can survive in typical classroom settings and with all of the challenges that come with conducting research in a naturalistic setting.

It is true that classroom studies can be difficult to implement, but additionally, even if conducted more often, there could be criticism that students need to start going beyond simple fact learning to integrate critical thinking (i.e., higher-order thinking) skills. Rote memorization and fact learning has its place in the classroom, of course—student success is heavily dependent on a foundational knowledge base in every subject. The problem arises when students are asked to go beyond the basic facts and apply them in problem solving situations where they need to interpret, analyze, evaluate, explain and make inferences. In order to do this, it is essential that students think critically.

Spacing and Critical Thinking

Critical thinking is an important tool if we are to maintain our roles within a democratic society (Dewey, 1909). As such, breaking down its components is necessary so that we can train our newest generation of thinkers. As citizens, students need to obtain a critical view of the

world instead of simply accepting the thoughts and opinions that are placed upon them—especially now that they are being constantly exposed to information not only through school, but at home via the Internet (Pearson, 2013). Given that critical thinking is at the forefront of many education policy documents (Fullan, 2013) and is of vital importance to the Ministry of Education (Ontario Ministry of Education, 2015), it is surprising that it has largely been ignored by the scientific community. To date, fewer than 1% of spacing effect research studies have examined the learning of critical thinking skills using anything similar to the above criteria. When compared to verbal learning, with 839 effect sizes reported in Cepeda et al. (2006) and 269 effect sizes reported in Janiszewshi et al. (2003), there have only been 11 critical thinking effect sizes reported in Donovan & Radosevich (1999) and 8 effect sizes reported in Moss (1995).

There have only been a few studies that have looked at spacing and critical thinking content in the classroom. The first looked at spacing in a simulated undergraduate classroom and investigated long-term benefits for factual and higher-level learning (Kapler et al., 2015). The researchers hosted a university lecture where they presented students with science material. Students were asked to review the material after either one (massed) or eight days (spaced), and five weeks after the last lesson, students completed a final test. Final test questions consisted of either factual or higher-level application questions. Reviewing the material in the spaced condition was more beneficial for both factual and higher-level questions on the final test.

Vlach and Sandhofer (2012) looked at spacing and critical thinking in students aged 5-7. Researchers taught and tested students in their university laboratory school. The researchers asked students to study facts about food chains and then tested their ability to generalize about the consequences of what happens when that food chain is disrupted. Students were in one of

three conditions: massed study sessions where all learning and reviewing occurred in one day; spaced study sessions that were spread across two days (which they referred to as *clumped*); or spaced study sessions that were spread across four days. Children were tested after a one-week retention interval. Students who were in the spacing conditions showed retention benefits for both the factual materials and the ability to generalize from what they had learned. Gluckman et al. (2014) replicated this study but added a memory component that added more fact learning content in addition to the generalization content. Memory for facts surrounding food chains was significantly better in the spaced conditions than in the massed conditions, and as expected, there also was a spacing advantage for both simple and complex generalizations of concepts.

Foot-Seymour et al. (2019) is the most recently published spacing and critical thinking study. This was the first known study to see whether effects typically expected from a controlled laboratory study could withstand the noise of the classroom in a non-verbal (critical thinking) realm. In this efficacy study of the spacing effect, researchers implemented and taught a critical thinking curriculum unit on website evaluation, which was based off of the standard media literacy curriculum. A total of 558 students in grades 4-6 were randomly assigned to either a massed condition (three days in a row), or spaced condition (three lessons one week apart). As expected, students who took part in the weekly lessons had a statistically significant spacing advantage on the final test 35-days later for both the fact and critical thinking measures (Figures 3 and 4). More specifically, students in the spaced condition remembered more from the website credibility lessons and were better able to explain their website ratings than students in the massed group.

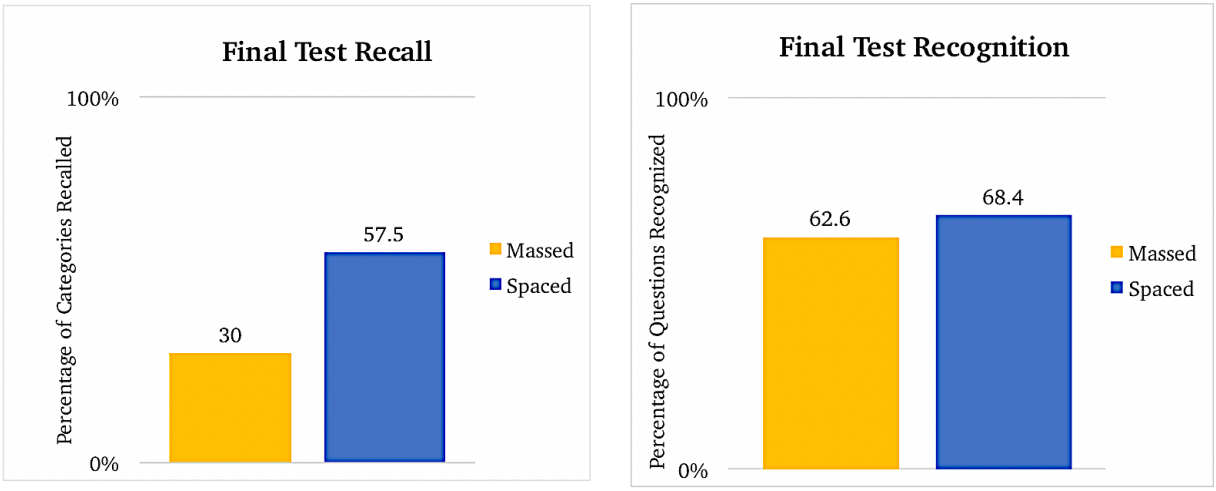


Figure 3. Fact learning measures in Foot-Seymour et al. (2019). The recall measure asked students to remember the four categories of website evaluation, and the recognition question asked students to remember which of the 17 questions they saw during the lessons.

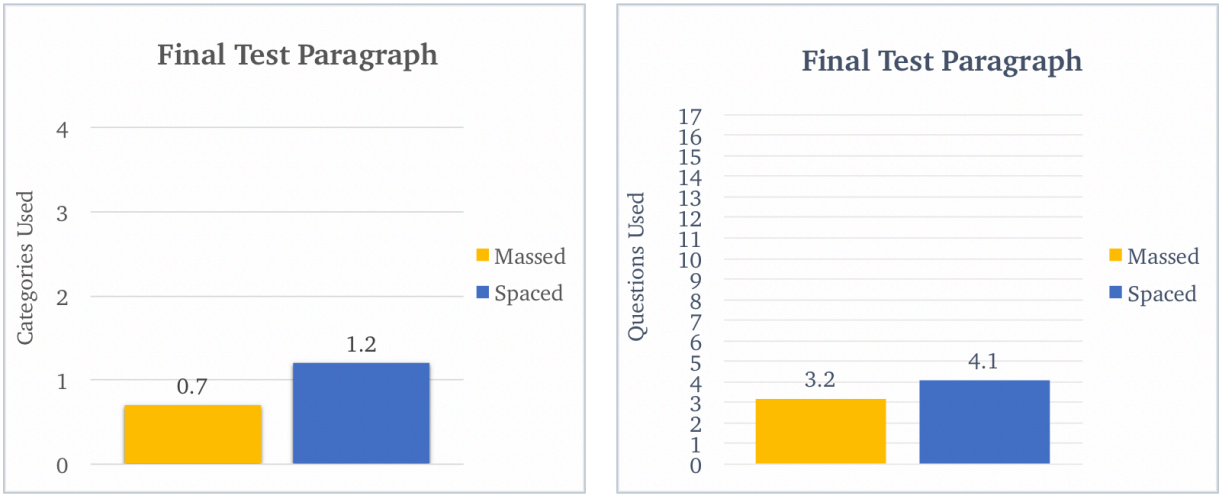


Figure 4. Critical thinking measures in Foot-Seymour et al (2019). The paragraph asked students to rate a website and measured the categories and questions that they spontaneously used to support their rating.

Critical Thinking and Website Evaluation

Robert Ennis defines critical thinking as "reasonable, reflective thinking that is focused on what to believe or do" (Ennis, 1987; Ennis, 2018). Other definitions of critical thinking exist (Facione, 1990; Kuhn, 1999; Siegel, 1988), but Ennis describes them as smaller pieces of a larger conceptual pie. Each definition is closely related to each other in the ways that count. A

key commonality is that critical thinking is goal-oriented—a good critical thinker evaluates their options before coming to a well-reasoned decision (Figure 5). Critical thinking is also best when the individual has some background knowledge and experience in the field in order to be able to engage in the full process.

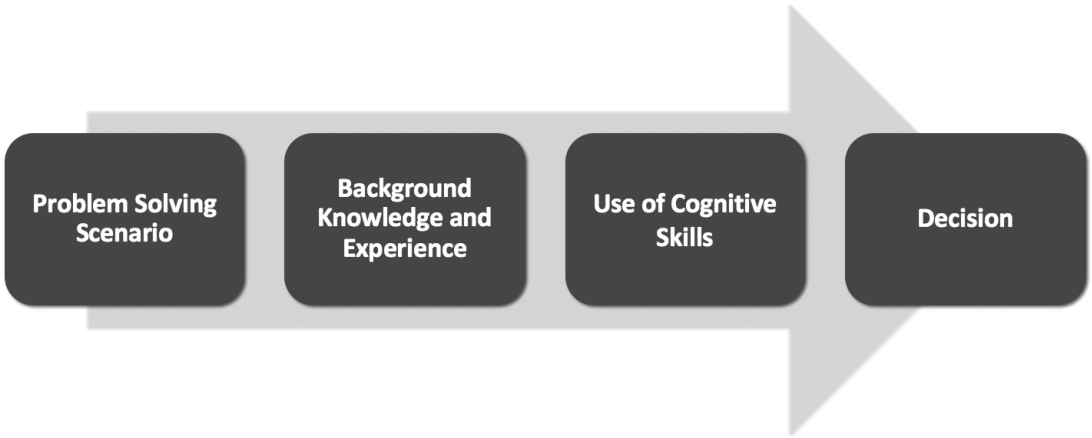


Figure 5. A summary of the critical thinking process by Robert Ennis (1987; 2018).

Experts within the field of decision making suggest that another way to view critical thinking is under the larger umbrella of rationality. The concept of rationality includes informed decision-making, reducing the chances of errors, assumptions, guesswork, subjectivity, and all of the other biases that might lead someone to making poor judgments. The advantage to viewing thinking in this way is that it has been studied successfully in the cognitive science literature for years (for a review, see Toplak, West & Stanovich, 2012). According to Toplak et al. (2012), "thinking" in and of itself is not a domain of knowledge. Students already know how to think—they just need to think *better*. How does one measure *better* thinking? A quality assessment would need to include practical scenarios where students use verbal reasoning, argument analysis, hypothesis testing, using likelihood and uncertainty, and decision making which would then be recognized and used appropriately (Halpern and Butler, 2019). Although much of that

assessment is beyond the scope of this research study, it should be considered by anyone designing or planning a critical thinking curriculum for students.

In order to define the specific skills and definitions of critical thinking, the American Philosophical Association (APA) brought together 46 leading scholars (including Robert Ennis) in the hopes of formulating a consensus. According to Ennis, critical thinking involves a set of pertinent skills and dispositions that should be taught explicitly and infused to everyday life in order to create an implicit understanding. These skills and definitions are listed in the Alpha Conception Report—a report outlining a list of six cognitive, or critical thinking skills: interpretation, analysis, inference, evaluation, self-regulation, and explanation (Table 1). The report (Facione, 1990) stated that:

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. Critical thinking is essential as a tool of inquiry. As such, critical thinking is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, critical thinking is a pervasive and self-rectifying human phenomenon. The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit. Thus, educating good critical thinkers means working toward this ideal. It combines developing critical thinking skills with nurturing those dispositions which consistently yield useful insights, and which are the basis of a rational and democratic society (p. 2).

This definition of critical thinking, although seemingly outdated, has stayed persistent in the literature (e.g., Abrami, Bernard, Borokhovski, Wade, Surbes, Tamim & Zhang, 2008; Boyd, 2019; Ennis, 2018; Facione, 1998; Lai, 2011) and has been cited over 2100 times. It has been critiqued for being too broad (Alston, 1991), but researchers in the field have responded by saying that critical thinking dispositions and abilities are analogous to defining what a “sport” is. Many

activities qualify as a sport, but there are no characteristics that define every sport perfectly. The same goes for critical thinking. A thinker is not required to engage in every single aspect of critical thinking, or have every single disposition, in order to be successful (Byrnes & Dunbar, 2014).

Critical thinking is also easier to conceptualize when it is put in the context of a problem, especially since it is goal-directed. By exemplifying and discussing critical thinking in the context of a real problem that is embedded in subject content, students can understand what it is and how to use it. In a meta-analysis of critical thinking skills in the classroom, Abrami et al. (2008) found that instruction of critical thinking was in fact most effective when students were taught critical thinking instruction and subject content in approximately equal parts. This led to their recommendation that teachers should be teaching critical thinking skills so that students are able to put them into context and learn to use them before transferring them to other disciplines. Students should be given practical and relevant examples of when they might use their developing critical thinking skills, such as website evaluation. In line with this approach are the findings from the APA document (Facione, 1990) and other research suggesting that subject matter should be taught with critical thinking skills training, as opposed to the latter taught separately (Angeli & Valanides, 2009; Ennis, 2018; Facione, 1990).

This is one of the primary reasons that we chose to teach critical thinking in conjunction with website evaluation in the current study. We also decided to teach website evaluation for value-based reasons. We felt as if teachers would gain more value from participating in a research study that targeted a specific and desirable skill, which would be fun for students to do, and has historically been difficult for teachers to teach. A qualitative study by Descours (2013) found that teachers in Canada and the United Kingdom do not have a united definition of critical thinking. Teachers who were surveyed in the study agreed that critical thinking is a skill, that it

can be taught, and that it should be infused within the curriculum, but the majority had conflicting ideas on how to achieve these goals. Group work, class discussion, the use of open-ended questions, and the willingness to accept multiple perspectives from students were some of the most common suggestions. Currently, some educators in Ontario are trained to use Bloom's taxonomy of educational objectives (Bloom, 1956) to teach critical thinking to their students. Bloom's taxonomy is a hierarchical framework that is intended, and even recommended by the Ministry (2007), to help teachers formulate test questions across the curriculum, ranging from specific fact retention to more practical complex reasoning. However, it has been suggested that Bloom's taxonomy is not appropriate to use in the classroom because there is little evidence that thinking is hierarchical in nature—learning facts and complex thinking are not completely distinct processes (Ennis, 1996).

Traditional spacing paradigms require that students who participate in different timing manipulations (spaced, massed) have a similar knowledge base in the first session so that they can build memory traces with each review session, working up to mastery. When balanced with the idea that critical thinking is best done with a known subject area, the topic was an important consideration. Website evaluation, although it intuitively seems as if it should be done on a regular basis, is not addressed in the classroom as much as it should be. This is surprising given that a high percentage of students in North America use the Internet for research purposes (Pearson, 2013). There has been a lot of research suggesting that despite its high frequency use, students are still largely uncritical users of websites as sources of information, especially in elementary school (Zhang, Duke & Jimenez, 2011). In fact, a self-report poll, although relatively outdated, indicated that as little as 4% of students check the accuracy of online information at school, and as little as 2% do so outside of home (New Literacies Research Team & Internet

Reading Research Group, 2006). This is a major issue since the Internet has become such a major part of students' lives. Critical literacy is closely tied, if not synonymous, with critical thinking, since it includes the need for readers approach the Internet with a selective, evaluative, and questioning stance (Burbules & Callister, 2000). See Table 2 for a chart of the APA's skills and dispositions and how they tie into website evaluation.

Table 1.

The Alpha Conception Report, outlining brief general critical thinking dispositions and abilities (Facione, 1990).

Dispositions. Ideal critical thinkers are disposed to:
Seek and offer clear statements of the conclusion or question
Seek and offer clear reasons, and be clear about their relationships with each other and the conclusion
Try to be well-informed
Use credible sources and observations, and usually mention them
Take into account the total situation
Keep in mind the basic concern in the context
Be open-minded
Seriously consider other points of view
Withhold judgment when the evidence and reasons are insufficient
Take a position and change a position when the evidence and reasons are sufficient
Seek as much precision as the nature of the subject admits
Seek the truth when it makes sense to do so, and more broadly, try to "get it right" to the extent possible or feasible
Employ their critical thinking abilities and dispositions
Abilities. Ideal critical thinkers have the ability to:
(Basic Clarification)
Focus on a question

Analyze arguments

Ask and answer clarification questions

Understand and use elementary graphs and maths

(Bases for a decision)

Judge the credibility of a source

Observe, and judge observation reports

Use existing knowledge

Background knowledge, including (with discrimination) internet material

Their knowledge of the situation

Their previously-established conclusions

(Inference)

Deduce, and judge decisions

Make, and judge inductive inferences and arguments

Enumerative induction

Argument and inference to best explanation

Make, and judge value judgements

(Advance clarification)

Define terms, and judge definitions

Handle equivocation appropriately

Attribute and judge unstated assumptions

Think suppositionally

Deal with fallacy labels

Be aware of, and check the quality of, their own thinking ("metacognition")

Deal with things in an orderly manner

(Non-Constitutive, But Helpful)

Employ rhetorical strategies

Table 2.

Overview of how critical thinking skills connect to evaluation of websites.

	Design	**Authority**	**Content**	**Purpose**
Interpretation: To comprehend and express the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, procedures or criteria.	Decode the significance of the website design, detecting and attending to all parts of the website.	Recognize the authorship of the website and decide, without using prejudice or bias, whether they can be trusted.	Detect, understand, describe, and characterize information from the website content.	Interpret the information displayed to decide why the website has been created (to inform, to persuade, to sell).
Analysis: To identify the intended and actual inferential relationships among statements, questions, concepts, descriptions, or other forms of representation intended to express beliefs, judgments, experiences, reasons, information or opinions.	Examine individual aspects of the website design, to make a decision about the site as a whole.	Identify frames of reference and perspectives of website author(s).	Compare and contrast ideas presented on the website, identifying their parts to come to a decision about the whole site.	Identify intended and actual inferences from information provided on the website, to determine why it was created.
Inference: To identify and secure elements needed to draw reasonable conclusions; to form conjectures and hypotheses; to consider relevant information and to deduce the consequences flowing from data, statements, principles, evidence, judgments, beliefs, opinions, concepts, descriptions, questions, or other forms of representation.	Identify and secure the elements of the website's design to draw conclusions about its credibility.	Use evidence and prior knowledge to draw conclusions from the website authority.	Draw conclusions from the website's information, using evidence and conjecturing alternatives.	Use evidence and prior knowledge to draw conclusions about why the website was created.
Self-Regulation: Self-consciously monitor one's cognitive activities, the elements used in those activities, and the results educed, partially by applying skills in analysis and evaluation to one's own inferential judgments with a view toward questioning, confirming, validating, or correcting one's own reasoning or one's results.	Identify self-biases and question, confirm, validate, or correct their ideas about the website's design.	Identify self-biases and question, confirm, validate, or correct their ideas about the website's authority.	Identify self-biases and question, confirm, validate, or correct their ideas about the website's content.	Identify self-biases and question, confirm, validate, or correct their ideas about the website's purpose.
Evaluation: To assess the credibility of statements or other representations which are accounts or descriptions of a person's perception, experience, situation, judgment, belief, or opinion; and to assess the logical strength of the actual or intended inferential relationships among statements,	Formulate an assessment of the website's design, looking at factors that may increase or decrease the	Assess the credibility of the website author, and whether they are a credible source.	Assess the statements and arguments on the website, identifying judgments, beliefs and opinions.	Assess the claims and arguments to determine why the website was created.

descriptions, questions or other forms of representation.	credibility of the site.			
Explanation: To state the results of one's reasoning, to justify that reasoning in terms of the evidential, conceptual, methodological, criteriological and contextual considerations upon which one's results were based; and to present one's reasoning on the form of cogent arguments.	Justify the procedure and present arguments that led students to their final credibility decision.	Justify the procedure and present arguments that led students to their final credibility decision.	Justify the procedure and present arguments that led students to their final credibility decision.	Justify the procedure and present arguments that led students to their final credibility decision.

Spacing Effect Theories

There are several theories that attempt to explain the spacing effect—most notably, encoding variability (e.g., Glenberg, 1979) and study-phase retrieval (e.g., Thios & D'Agostino, 1976). The encoding variability theory suggests that each item is stored in memory along with the specific context in which it was learned, and that context changes over time. The greater the number of distinctive contexts that are associated with each item, the larger the probability that the item can be accessed and successfully retrieved (i.e., the spacing group would have a better chance of increasing their contextual cues than the massed group). An alternative theory, study-phase retrieval, suggests that learning of an item will be superior if the first memory trace can be retrieved from memory and that initial memory trace strengthened. For items that are retrieved soon after the first learning session, the reconstruction process will be easy, leading to little additional memory trace strengthening. For items that are retrieved later, after a spacing interval, retrieval will be more effortful and greater reconstruction will occur. This study phase retrieval theory ties into research on desirable difficulties (Bjork & Bjork, 2011), which explains that it is often slightly *more difficult learning* (e.g., spaced out study sessions) that can lead to better retention and transfer later on. These two theories differ in terms of the key mechanisms responsible for driving learning benefits. However, some researchers have recently suggested

that there are multiple mechanisms at play, and that spacing effects are due to a combination of encoding variability and study-phase retrieval effects (Delaney et al., 2010; Karpicke, Legman & Aue, 2014; Mozer, Pashler, Cepeda, Lindsey & Vul, 2009).

Current Study

The next step when conducting classroom research is to see whether similar effect sizes could be seen with even less control than in the efficacy study by Foot-Seymour et al. (2019), where the lead researcher taught each lesson herself and maintained as much classroom control as possible. The current study was an effectiveness trial to see whether spacing benefits could be robust in ecologically valid settings, with homeroom teachers executing pre-designed lessons.

An understanding of the distinction between the terms *efficacy* and *effectiveness* study is crucial when conducting research but also when interpreting results from studies and making assumptions about their generalizability (Signal & Waljee, 2014). These types of studies sit on a spectrum, with efficacy studies being defined as performance under ideal and controlled circumstances, while effectiveness studies occur in real-world conditions. If the criteria are as in *Table 3* below, this study would lie somewhere in the middle. The term effectiveness study was chosen because although the study meets some of the criteria for efficacy research, the classroom is not ideal or controlled in the best of times. Effectiveness research has the added benefit of accounting for external factors that may decrease intervention's effect, thereby lessening the effect size. It can therefore be more relevant for making decisions when it comes to the generalizability of a finding, since there is less control than would be expected from an efficacy trial.

Table 3.

Differences between efficacy and effectiveness studies (Singal, Higgins, & Waljee, 2014).

	Efficacy Study	Effectiveness Study
Question	Does the intervention work under ideal circumstances?	Does the intervention work in real-world practice?
Setting	Resource-intensive 'ideal setting'	Real world everyday setting
Study Population	Highly selected, homogenous population Several exclusion criteria	Heterogenous population Few to no exclusion criteria
Providers	Highly experienced and trained	Representative usual providers
Interventions	Strictly enforced and standardized	Applied with flexibility

If spacing is to be used in the classroom with no researcher support, more evidence is needed to see if teachers can lead the intervention on their own with traditional lesson plans and minimal instruction. That is what the current study set out to do. With guidance from the lead researcher (VF), teachers were given lesson plans and manipulated the timing of when lessons were taught, reviewed and tested. Teachers were able to execute these lessons during their usual literacy block, since lessons were embedded with curriculum-based content and taught by the participants' homeroom teacher. Students participated in the same three lessons but took part in either a spaced learning schedule (*weekly* lessons: 7-day interstudy interval [ISI]), or a massed learning schedule (*daily* lessons: 1-day interstudy interval [ISI]). All students were given a final test approximately 35 days later. 35 days was chosen as the retention interval because it's the optimal retention interval (RI) for a 7-day spacing condition (Cepeda et al., 2008), and because it was the ISI and RI combination in the most related spacing effect study (Foot-Seymour et al., 2019). Additionally, it is feasible for a teacher to plan their lessons with a one-week spacing

design with a recommended testing date of one month from the last lesson. Since spacing benefits are present across a very wide range of retention intervals, there would be essentially the same results even with a slightly different ISI and RI combination (Cepeda et al., 2006, 2008).

Hypotheses

The hypotheses were as follows: First, the spacing effect would benefit fact learning. Students in the spacing condition, when cued, would *recall* more information from the lessons (the four categories of website evaluation: design, authority, content and purpose; Table 4) than students in the massed condition. This will be prompted by asking students, "What are the four categories of website evaluation?" Second, the spacing effect would benefit critical thinking: Students in the spaced condition would spontaneously *use* more information in an open-ended paragraph, by giving details taught in the lessons to explain their website ratings, than students in the massed condition. This will be prompted by giving students two different websites and simply asking for each, "Is this website credible? Please explain using evidence from the website."

Websites were designed by VF and student volunteers with a rating in mind. Following the above checklist, each of the five websites that students were asked to rate from 0-10 had a specific answer (e.g., the pre-test website should have been rated a *5* since 50% or 7 out of 14 of the answers to the above questions were *yes* and the other half were *no)*.

Table 4.

Website evaluation checklist (adapted from Bronstein [2007] and Foot-Seymour et al. [2019]).

Design
Do the photos and colour choices look professional?
Is the website nicely organized and easy to navigate?
Are there any obvious spelling errors or typos?
Is the layout consistent from page to page?
Authority
Is the author/creator of the website clearly identified?
Is the author of the website an expert in their field?
Is there a way to contact the author by phone, mail or e-mail?
Content
Does the website say when it was created?
Does the website say when it was last updated?
Can you confirm that the information is correct by doing a Google search?
Are the links relevant to the subject? In other words, do the links take you somewhere that makes sense if you click on them?
Purpose
Is the website trying to educate you with real information?
Is the author trying to sell you something?
Do you think the author has intentionally left out any important information that could help you decide if it's real or fake?

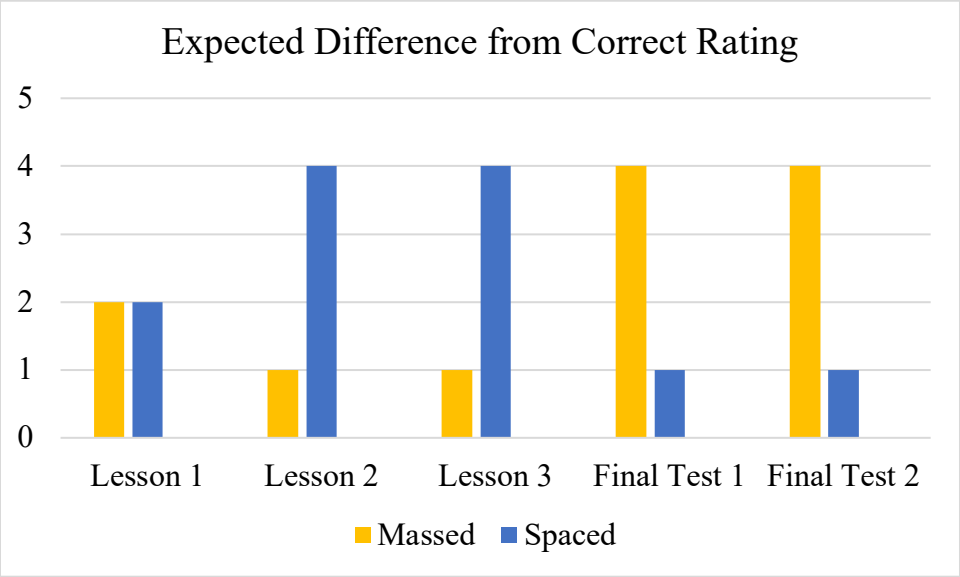


Figure 6. Expected difference from correct rating

Based on the pre-existing verbal literature, we had additional hypotheses tied to ratings: Students in both groups would rate the pre-test similarly at baseline, but students in the massed condition (represented by the blue line in Figure 6), would be better at rating websites *during* the daily lessons than spaced students in the weekly lessons. This can be seen in Figure 6 with the massed group staying closer to the correct rating (represented by a smaller difference score) than the spaced group. At final test, the trend would shift, and the spaced group would rate the websites more accurately than the massed group. These hypotheses are tied to the desirable difficulties theory (Bjork & Bjork, 2011) described earlier, as well as evidence from the verbal spacing literature that suggests that massed learners are more successful at retrieving information in the short-term during lessons that are massed together in time, whereas they forget more of the content in the long-term. Spaced learners, on the other hand, have a more difficult time *during* practice sessions but have an easier time retrieving it after the retention interval (Wiseheart et al., 2019).

We also had an exploratory hypothesis, to see whether student responses would have any connection to the categories and questions that were focussed on. This is because certain categories may have swayed responses. For example, one student mentioned at final test, “I didn't have any issues with purpose, but the problems in the other areas were so severe that I had to rate this site a zero. I think content is the most important part of website credibility, and since this site's information was incorrect, there was basically nothing useful about this site.”

Method

This study explored whether spacing improves learning and retention of both fact-learning and critical thinking skills in the elementary classroom. Fact learning was defined as *cued recall* of the information provided in the lessons (the four categories described above), and critical thinking was defined as the spontaneous *use* of the information (the four categories and specific questions) that had been provided in the lessons one month earlier. In order to address this research question, a typical spacing paradigm was used. Students had three study sessions (lessons) covering the same conceptual information, separated by a period of time referred to as the inter-study interval (ISI). Three lessons were taught to mimic standard teaching practice and to give students more of an opportunity to learn the skill. After a 35-day retention interval, students completed a final test assessing their retention and ability to use the information from the lessons (Figure 7).

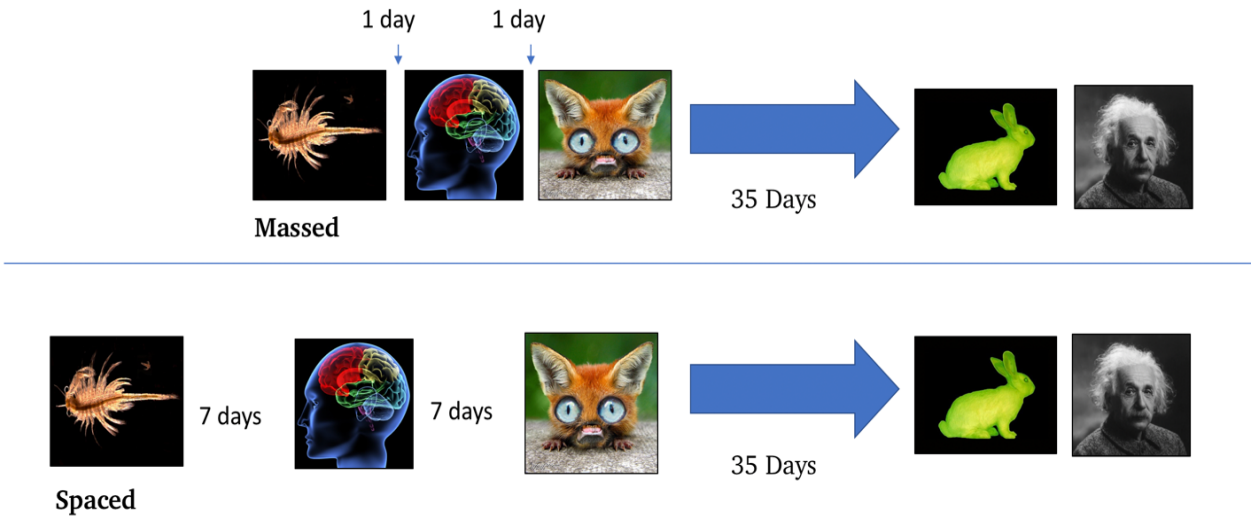


Figure 7. Visual representation of the current study (massed, spaced). Each photo represents a new website where students practiced their website evaluation skills.

Participants

Elementary school students from York Region District School Board, aged 10-14 years, participated in this study. This age group was chosen because the Ontario Curriculum asks that, starting at approximately 10 years old, students must begin to value critical literacy and "differentiate between fact and opinion; evaluate the credibility of sources, and recognize bias" (Ontario Ministry of Education, 2005, p. 89). After this learning begins, there is no formal curriculum material that asks students to draw upon their critical thinking skills as we have described in this paper, and none which ask them to evaluate websites. It is often the responsibility of the teacher to implement this training in their program, and it varies widely between teachers. In order to randomize these effects, we focussed on recruiting a large sample size with an even distribution of student ages across both conditions.

A total of 1054 students were recruited for the website credibility lessons, from 16 participating schools across York Region District School Board. There were 42 participating classrooms, each with their own homeroom teacher. Of the students who were recruited, three

did not receive parental consent to participate and were removed from the classroom for the lessons. One full classroom was excluded from the final data set due to a teacher-altered spacing schedule (this classroom created a spacing schedule of 4-5-4 days instead of the requested 7-7-7 days). A further 36 students were given parental permission to participate in the lessons but asked researchers not to use their data for the analyses. Since there were four lessons including final test, and all were necessary to collect a full data set from each student, a total of 160 students were excluded from data analyses due to missing a lesson (e.g., due to missing a day at school for illness or school activity). Since teachers were encouraged to include all students including those who were on an IEP (individualized education plan) who would have typically been removed from class and placed into their SERT (special education resource) room with a special education resource teacher, is it possible that some of these students left for their regular programming on at least one of the days which would have created some missing data. Some students (n = 15) were excluded from the analysis because they were English Language Learners (ELL) and did not read and write English without full support—however, ELL students were given the help they needed via a one-on-one teacher or Google Translate so that they could still participate in the lessons as much as possible. All efforts were made to ensure that our research practices were fair and equitable. The final sample consisted of 716 (n = 349 spaced; n = 367 massed), with a mean age of M = 11.77 (SD = 1.13) for the spaced group and M = 11.97 (SD = 1.07) for the massed group. See Figure 8 for the RCT flow diagram, and for a more comprehensive overview of participants from the final sample, see Table 5.

Our sample size surpassed our minimum recruitment aim, which was n = 114 per group at analysis. We based sample size on an effect size of d = 0.48 and 95% power, our estimate of the effect size for critical thinking and spacing based on a the most related prior classroom study

(Foot-Seymour et al., 2019). We aimed for a much larger sample to account for differences in teacher effectiveness, aiming for a sufficient sample size that mean teacher effectiveness would be approximately equivalent across groups. Since we do not know the distribution of teacher effectiveness, and thus cannot determine the minimum required number of classrooms, we aimed to recruit as many classrooms as possible during the time available for data collection. Given our large sample size, our power to detect an effect was 99.99%.

Since it is not standard procedure within schools to collect equity and identity-based data from students due to ethical considerations, census data for York Region were reported instead. Demographic census data demonstrates that 51% of York Region's population are Caucasian and 49% are from a visible minority. Out of those identifying as a visible minority, 45% self-identified as Chinese, 22% as South Asian, 8% as West Asian, 5% as Black, 5% as Filipino, 3% as Korean, 3% as Southeast Asian, 3% as Latin American, 2% as Arab, 1% as Japanese, and 4% as multiple or another visible minority. More details on York Region demographics are available on the Public Tableau website (Regional Municipality of York, 2018).

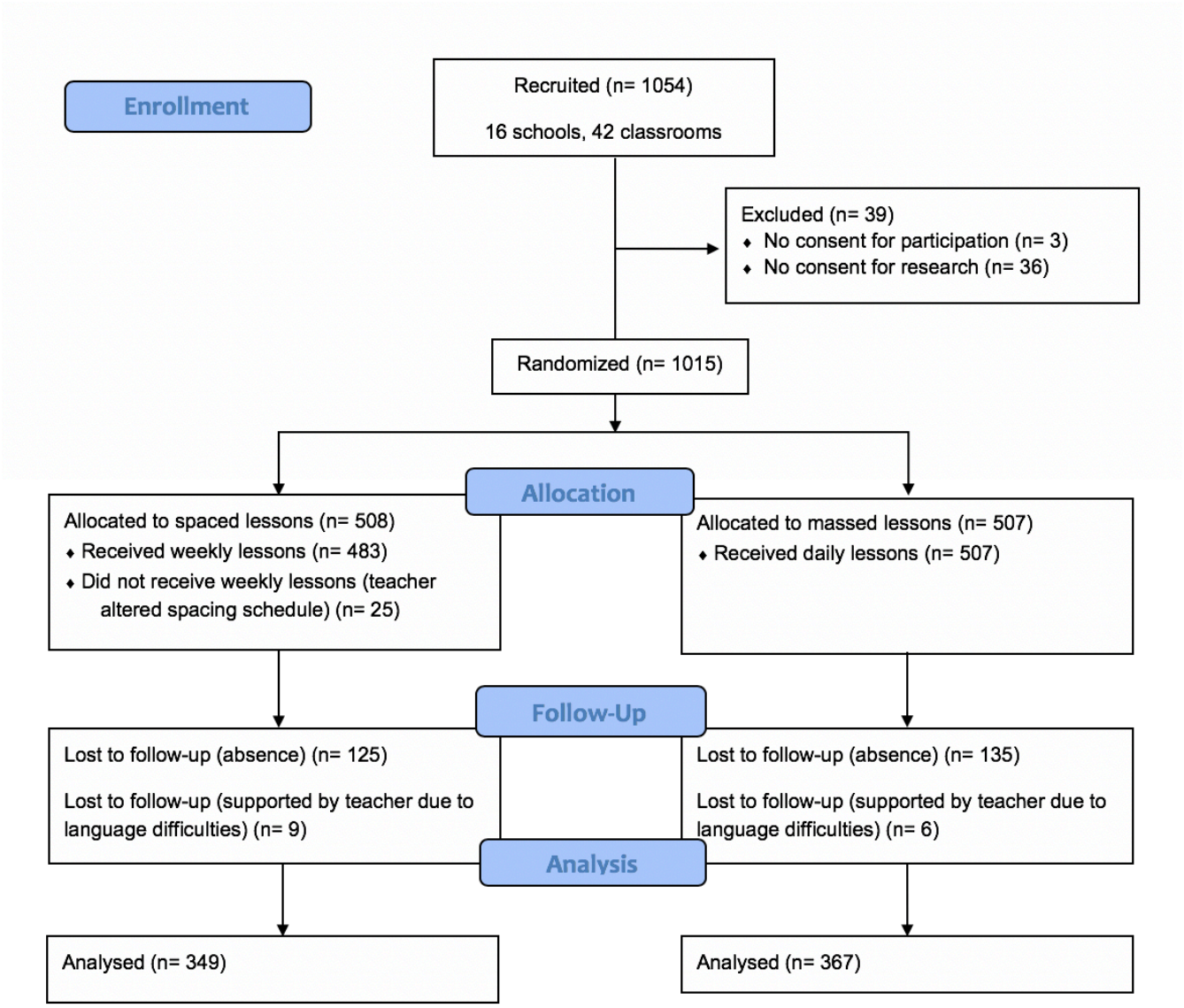


Figure 8. RCT flow diagram of sample.

Table 5.

*Overview of final sample. *As reported by participants.*

		Spaced (n = 349)	Massed (n = 367)	Overall (n = 716)
Gender Identity*	Male	175	197	372
	Female	165	169	334
	Other	1	0	1
	Prefer not to answer	8	1	9
Grade	5	56	39	95
	6	125	101	226
	7	85	122	207
	8	83	105	188
Age	10	41	31	72
	11	122	94	216
	12	87	120	207
	13	74	98	172
	14	25	24	49

Design

A between-subjects design was used, where classes were randomly assigned to either the spaced or massed condition, stratified to ensure that there were an equal number of grades and locations for each condition. The massed condition was used as a control. Students in both conditions were given an identical set of online lessons but received the lessons daily (massed: three days in a row) or weekly (the same day of the week for three weeks). Although traditional spacing effect studies have only two study sessions (see Cepeda et al., 2006), three study sessions (lessons) were taught to mimic the methodology in Foot-Seymour et al. (2019), who

made the decision to add a lesson so that students could experience more variability of websites and have an additional chance to review the content.

Classes were taught on each day of the week, with specific days varied across classrooms, and there was a mixture of days in each condition. Volunteers were sent to classrooms on each day of the study to assist with students who needed extra support, and to ensure that teachers were carrying out the intervention in the agreed upon schedule (the only class to run without a volunteer was the class that was excluded due to their self-made spacing schedule that differed from the rest). The volunteers' presence was non-intrusive and did not affect the teacher's ability to implement the lessons as per their usual teaching practice.

Procedure

After York University Human Research Participants Committee ethics approval and York Region External Research permissions were obtained, schools and classrooms were contacted in person and via e-mail and selected based on principal and teacher interest. When a teacher agreed to participate and the principal gave approval, lesson plans were sent (Appendix A), and dates were pre-selected by the researcher. Teachers could send alternative date options but were required to participate in the condition that they were assigned to. Communication with teachers was frequent to ensure that all aspects of the research were going smoothly. Consent forms were collected by teachers before the lessons began. Consent forms had three options: (1) students could participate in all aspects of the lessons and have their work used anonymously for research purposes; (2) students could participate in all aspects of the lessons but could not have their work used anonymously for research purposes (responses could be recorded initially but needed to be deleted before the analysis); and (3) students could not participate in any aspect of the lessons. If parents chose the third option and did not consent, they were contacted by the teacher to confirm

and ask whether students could do alternative programming in the room or if parents would like to send them to another classroom. Only three parents chose this option (n = 3): these students were removed and placed in another class to complete a task assigned by their homeroom teacher. The 36 students who did not receive consent to have their work used for research purposes (n = 36) participated in all aspects of the lessons (in-person and online), but when the survey prompted them to give their identification, they were told to mark the box with an "X" so that their responses could be deleted from the system. Consent forms were also collected to ensure that these students did not have their responses recorded and saved. The in-class consent management was done quietly and efficiently so that other students were not aware of who didn't get consent, in an attempt to keep classrooms inclusive and equitable.

All lessons were online (Appendix A contains the links to all online lessons), with videos and questions programmed on Qualtrics. On day one, the researcher was present to meet the students and teacher and show them how to access the survey via URL. Headphones were provided to students who did not bring them from home. The researcher led students through lesson one, but minimal formal teaching was required since all lessons were accessed through the URL. The researcher was also there to troubleshoot if necessary and answer any questions about the lesson plan in person.

Teachers varied in their preparedness for the first lesson. Some had lesson plans printed off and highlighted, in an effort to fully understand all aspects of the study. Others had not prepared for the first lesson, asking for a copy of the lesson plan that had been sent to them previously. Since teachers were recruited several months before the first lesson was scheduled, some reflected that their busy teaching schedules prevented them from being fully prepared for the lesson when it came up. As such, it was necessary for the lead researcher to come in to model

the lessons. After this was done, the remaining review lessons were executed exactly as the lesson plans required. Each teacher was asked if they would be able to lead lesson one without the researcher, and each teacher indicated that they could. For the remaining lessons, a volunteer was sent out to classrooms to ensure that the lessons were carried out properly and were told to notify the researcher of any changes to lesson plans. There were a few slight changes in regard to lesson timing (one class lost Internet for 20 minutes; another ran out of time and needed to have the final discussion over lunch). However, this was not measured in any way and each situation was addressed in the moment to determine whether the class should proceed as usual. As long as students finished their assigned tasks within a reasonable time frame, they were included in the results.

A total of 80-100 minutes was allocated per class for the first day. At the beginning of the lesson, students watched a short introduction video which contained the definition of credibility and completed a pre-test website evaluation on the Sea Monkey website (www.seamonkeyonline.net). For this evaluation, students were asked to explain why they thought the website was credible or not prior to taking part in the credibility lessons. Responses were required to have a minimum of 150 characters (approximately three sentences). The pre-test was included to measure student responses at baseline.

After the pre-test, students watched a YouTube video (pre-recorded by the researchers) which led them through the credibility checklist using an example from a National Geographic website. After the video finished, they were asked to go through the checklist for the Sea Monkey website again (www.seamonkeyonline.net), give another rating out of 10 now that they could make a more informed decision, and explain their rating using the four categories (design, authority, content, and purpose) in a paragraph. After about an hour, students verbally shared

their answer as a class in a group discussion format. The first discussion showed that students were already thinking critically—they were engaged in discourse about the categories of website evaluation (e.g., for the design category, some students commented that "the website looks like a blog so it can't be credible," while others said, "it had good contrast with the white background and black font. It could have been worse"). The discussion was held for approximately 10-15 minutes, or until students were done sharing their ideas. Responses were recorded on the board in point form but were later erased to prevent students from any additional studying.

Lesson two was led by the homeroom teacher either one day or one week later. Students were randomly placed into small groups and were asked to brainstorm the four categories and 14 questions that they had previously learned. Responses were recorded on a chart paper that was later destroyed to prevent studying. Then, students went on the second website (www.brain-science.ca) and used the online website evaluation survey to record their responses. Lastly, students verbally shared their answer as a class in a group discussion format as they had done in the first lesson.

Lesson three was identical to lesson two, but with a new website (www.bizarre-animals.ca). Students completed the small group activity, then the online website evaluation survey, and then had their class discussion.

For the final test, which took place 35 days after lesson three, students were asked to complete three tasks. First, they were asked to recall the four categories of website evaluation (design, authority, content, and purpose). Next, they were given two websites (www.researchscience.net, www.associationofgeniuses.org) and asked to evaluate them one at a time, give them a rating out of 10, and write a paragraph (without the checklist) supporting their rating. This was an identical format to what students completed during the pre-test. Written

responses were required to have a minimum of 300 characters. In order to do well, students needed to spontaneously use the four categories of website evaluation and the 14 questions within those categories that they were taught during the lessons.

Lastly, students were asked which of the four categories they thought was the most important to determine website credibility. We had no a priori hypotheses about what would be said but looked to see if there was a relationship between what was used in their paragraphs and what students claimed to put value on the most. We also used this measurement to determine whether each category was equally reliable and useful, as the websites were designed, or if students were basing their decisions on a smaller number of categories. For example, if a website did not have a credible design but the content was credible (e.g., the content could be confirmed with a Google search, the links were all relevant and up to date, and the content was regularly updated) we were interested to know whether students were consistently rating the websites higher than if the reverse was true (the design was excellent but the content was poor).

After the final test was completed and student data was finalized, students were matched across the pre-test, teaching sessions, and the final test, in order to make sure that each student who had their data analysed was present on all days of the lessons. Any student who missed a day of the lesson was removed from the data analysis.

Materials

Websites. All websites were created by the researchers on WordPress. Each was designed to have a specific level of credibility (3, 5, or 7 out of 10), with at least one of the four categories scoring very low (Table 6). Red flags (deliberate errors) were embedded throughout the websites to encourage a scavenger-hunt feel while students were going through the checklist.

Table 6.

Websites Used During Credibility Lessons. Rating values represent correct rating out of 10, and individual category numbers show which categories scored high/low to lead students to that decision.

Website	Rating (/10)	Design (/4)	Authority (/3)	Content (/4)	Purpose (/3)	Total (/14)
Sea Monkey Online	5	3	0	1	3	7
Brain Science	3	4	0	0	0	4
Bizarre Animals	7	4	0	4	2	10
Glowing Bunnies	7	1	3	3	3	10
Association of Geniuses	3	2	0	1	1	4

Sea Monkey Online (www.seamonkeyonline.net)

"This website is all about the marvellous creature, the sea monkey! Feel free to browse, search and comment." This website taught students about the true history of the sea monkey, in an error-ridden blog formatted website. There were distracting spelling errors all over the website that varied in word difficulty. The information provided was true, but the authorship was up for debate. The website claimed to be written by "The Office of Science and Society" at "MacGill University" but gave no author name or credentials, and the hyperlink that was connected to that name took students to a different website run by the real McGill University. The author name was still embedded deep within the McGill University website. Some students noticed the error in the university name on the Sea Monkey website immediately and others did not, but regardless, all students were taught that they needed to pay attention to small details and trust their gut when it came to making decisions about specific website items to see if they were red flags (errors) or not.

Brain Science (www.brain-science.ca)

This website was based on the myth that people only use 10% of their brains, and there was an expensive pill that could change that. The purpose of the website was in question, asking for large sums of money in exchange for this super pill. *"Neuroflex is the first ever pill that allows humans to improve their brain power! It allows you to activate more regions of your brain and guarantees obtaining the maximum results with the minimum amount of effort. Neuroflex consists of a few essential ingredients that are important in enhancing brain function. It is 100% natural, with all ingredients extracted from plant sources.*" One of the defining features of this website was that it was visually pleasing and very professional looking. There was an author name on the website (Dr. Daniel Reid), and the site header gave credit to the "International Journal of Brain Science." It also appeared to have very scientific-sounding information. However, this was the least credible website of the lessons, identified by the false information that students would have noticed while doing a Google search of the content. Also, most students noticed that the website was trying to sell them a very expensive pill (in the currency of British pounds) which was sold in a bottle that, unlike the rest of the website, did not look professional.

Bizarre Animals (www.bizarre-animals.ca)

This website took students through several strange animals, like the giant squid. It taught real content about seemingly bizarre creatures. *"The giant squid lives in the depths of the ocean. Giant squids can grow to a tremendous size due to deep-sea gigantism. Recent estimates put the maximum size at 13 m for females and 10 m or males from the back fins to the tip of the two long tentacles (second only to the colossal squid at an estimated 14 m (46 ft), one of the largest living organisms."* Inspiration from this website's content came from the story of the Gulper Eel, a rare deep-sea creature that can stretch its jaws in a remarkable way. Other animals were added to the website that were equally unusual. Students might have had some prior knowledge of at least one

of these animals, but others were likely so rare that they would have needed to check their credibility before claiming certainty. There were also red flags in the author category (each post lists the author as "staff"), which at this point the students would have known and needed to review in order to successfully rate the website.

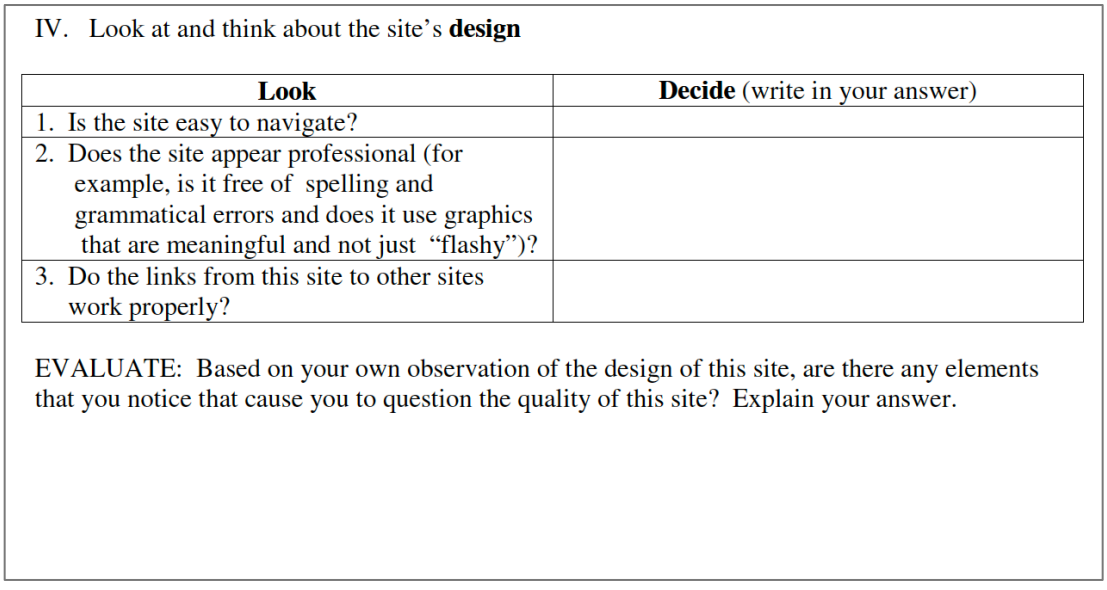
Glow-in-the-Dark Bunnies (www.researchscience.net)

"In normal light, these rabbits all look normal—cute, fluffy, and white. But wait until you turn off the lights. The rabbits glow fluorescent green!" This website shared real research about glow-in-the-dark animals, but the design looked unofficial, with a lime green background, blurry photos, and red text. Many students noticed that although the information on the website was true, there was not much content listed that could help them make their decision. Most of the content listed were hyperlinks that brought them to other sites. There was also a red flag in the form of a picture of a regular rabbit, with the heading, "to compare, here is a photo of a normal rabbit." Many students noticed that this was out of place on this scientific-seeming website.

Association of Geniuses (www.associationofgeniuses.org)

"Sharing biographies of geniuses around the world." This website was part Albert Einstein biography and part advertisement for an association that supposedly aimed to help young geniuses discover their full genius potential. The purpose of this website was two-fold, and the content was completely false. This website had a very simple design but had inconsistent fonts on every page. Students were also told that they could donate to provide an hour of tutoring for a child, but until that point the website was not convincing enough to give them confidence that this would have been a good idea (as noted from the website responses students gave during this lesson). The donation button was not connected to any sort of account, so if students tried to see whether they could have donated to the cause, they would not have been able to.

Website credibility checklist. The website credibility checklist (Table 4) was originally adapted from Bronstein (2007) for use in our prior (Foot-Seymour et al., 2019) website credibility and spacing effect study. Bronstein created a credibility checklist with the assistance of a Delphi panel of experts and explored the reliability and validity of this checklist for classroom use. She summarized a variety of commonly used checklists and designed her own based on a mixture of best-practices by other educators. The checklist was designed so that students could have little to no background knowledge or critical thinking vocabulary and could be encouraged to respond with more than a simple "yes" or "no" assessment while proceeding through the list. She argued that instead of checklists with only binary options, continuous scales should be used, since critical evaluation is an ambiguous process that involves many different options for premises and different forms of reasoning that are equally legitimate. The goal was to gain deeper insight into students' thought processes. Instead, students would look at the category (design, authority, content and purpose) and then write in a response to explain their thinking (Figure 9). The full checklist was finalized by the Delphi panel and aimed for delivery to high school students.



IV. Look at and think about the site's **design**

Look	**Decide** (write in your answer)
1. Is the site easy to navigate?	
2. Does the site appear professional (for example, is it free of spelling and grammatical errors and does it use graphics that are meaningful and not just "flashy")?	
3. Do the links from this site to other sites work properly?	

EVALUATE: Based on your own observation of the design of this site, are there any elements that you notice that cause you to question the quality of this site? Explain your answer.

Figure 9. The design category from the Bronstein (2007) checklist.

During the Foot-Seymour et al. (2019) study, the Bronstein (2007) checklist was revised to suit a younger audience. One of the major differences was that students were asked to make some binary "yes" or "no" decisions about each specific question before explaining their thinking. There were 17 specific questions used within Bronstein's four categories, most of which were adapted from the original list. Students were also asked to make a final decision about their overall evaluation of the website, but their final decision was turned into a continuous scale so that students could give a value between 0-100 (0 being least credible and 100 being most credible) and then explain this rating in a paragraph. We hoped that students would use a combination of these tools (the checklist, rating scale, and written paragraph) in order to formulate an opinion. The paragraph and subsequent final rating of the websites was a very important step in the credibility process, since the paragraph explaining the rating was intended to justify *why* the student gave the answer that they did. However, the problem with the 0-100 continuous scale in Foot-Seymour et al. (2019) was that both students and teachers often gave extreme ratings (Figure 10). It was hypothesized that this range was too wide, and the range was reduced in the current study.

For the current study, several of the checklist questions from these two checklists were removed and/or changed, and the continuous scale was changed to 0-10. Specific questions were changed so that the responses would result in clearer "yes" or "no" response during the lessons so that we could standardize the credibility of the websites in advance. An example of a change made was for the purpose category—the original checklist asked students, "has the author convinced you to see their point of view?" and the current checklist asked, "is the website trying to educate you with real information?," which forced students to make an overall judgment about the website purpose which forced them to decide if the information was mostly real or fake.

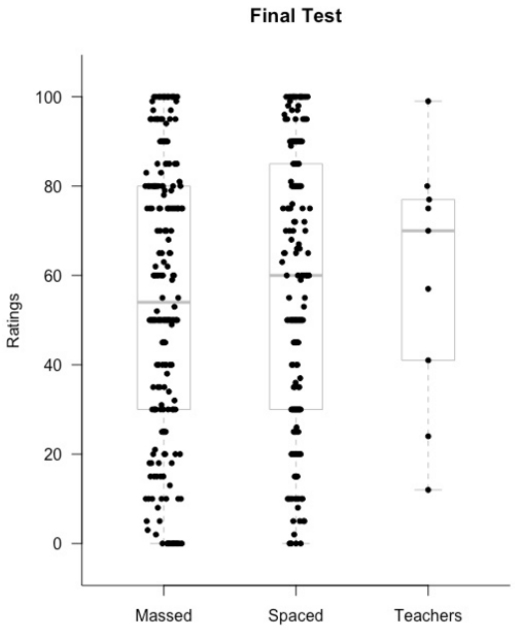


Figure 10. Dotplot of final test website ratings from Foot, 2016..

The checklists appeared to work exactly as expected—after going through the checklists, students were generally successful in matching their rating from 0-10 with their explanation. Even on the pre-test before the lessons were taught, students seemed to be pairing their answer with their rating—like this student, for example, who rated the Sea Monkey website a 5/10 during the pre-test: "I do not think that this website is that credible overall. You can learn some information from it that seems true but, on the other hand some of there information is not very believable. The credibility of this website in my opinion is in the middle." This student had not been introduced to the checklist where they learned to explain their rating. Still, they were able to successfully communicate that although they could formulate an answer (of *not credible* overall), they were still going back and forth. Other students struggled with their written answers in the pre-test, explaining that "I think that the website is pretty bad and I don't trust it. I don't know why. I wish you could just tell me what to do because I don't know what to do. I went through the website and read it but I can't tell."

After the lessons had begun and students were taught to explain ratings using evidence from the website, answers became much more substantial, and still matched their ratings. On the Brain Science website during the second lesson, one student gave the website a 3/10, correctly identifying that:

> The content was not too terrible at first, but they included a myth. It included some educational facts about the brain. The author might've been made up. According to a google search, the author isn't even a neuroscientist, what he said he was. The design looked professional, but I don't think the rest was. I feel that the main purpose of the website is to sell NeuroFlex, something that they say will increase how much you are able to use your brain power. I think that is ridiculous, considering the fact that they said you have to take 18 pills a day; 1 pill per hour. It included reviews, and it included a name of an author of a review. I looked up that name and it was spelt wrong but it was from the right place though. All of that made me think that this website is not really credible.

Analyses

A set of t-tests, one-way ANOVAs and Bayesian analyses were conducted. Bayesian analyses are being used more frequently in applied and fundamental research because they use a very different view of hypothesis testing-- one that allows researchers to incorporate background knowledge into their hypotheses (van de Schoot, Kaplan, Denissen, Asendorpf, Neyer and van Aken, 2014). Bayesian analyses use probability to determine whether “nothing is going on” (i.e., the hypothesis is null), or if something is going on, and to what degree. In order to interpret Bayesian analysis, a score of 0.33-3 indicates indeterminate evidence, 3-10 indicates moderate evidence for H1, 10-30 indicates strong evidence, 30-100 indicates very strong evidence, and 100+ indicates extreme evidence; values below 0.33 reflect evidence in favor of H0, with increasingly small values representing increasingly strong evidence. For a full description, see van de Schoot et al. (2014).

Before running the analyses, tests were conducted to ensure that assumptions were satisfied. There were violations of normality in every sample, but the results were still reported

since our sample size was large. However, a non-parametric test (Mann-Whitney U) was run on the ranks when possible to ensure the accuracy of the results. These results showed the same outcome as the t-tests on the final test measures. Levene's test for equality of variance was conducted for each test under the requirement of $p > .05$. When this assumption was violated, degrees of freedom were adjusted. Lastly, the independence of observations assumption was violated since there was nesting by classroom, but we addressed this violation by running a separate post-hoc analysis that used classrooms as independent data points instead of students. We also looked at classroom and grade effects using tests for every dependent variable (final test recall, the paragraph measures, and the ratings) to dive deeper into the data.

Due to the inclusive nature of the study, no outliers were removed. However, efforts were made to check (post-hoc) if the results would differ when certain groups of students were removed from the analyses: (1) students with self-reported effort scores of 0 or 1 out of a possible 5 at final test (n = 42); (2) students with self-reported effort scores of 0 or 1 out of a possible 5 during the learning sessions *and* at final test (lesson 1, n = 16; lesson 2, n = 18; n= 23; final test, n = 42); (3) students who had any missing data during the learning sessions, since we couldn't be sure that they were completing the full task (n = 46), and (4) one class in the spaced condition that scored significantly lower marks at final test (n = 19). Removal of these data was attempted one at a time and in combination and showed no difference in results. Therefore, the following results will include our entire final sample of n = 716. See Table 9 for a full summary of the data.

Results

Baseline

First, we examined the use of the four categories and 14 questions in the pre-test rating paragraph, in order to ensure that students did not already know the material and as a check on

the sufficiency of random assignment and stratification procedures. Since all responses were marked by hand and paraphrasing was accepted (e.g. “who made the website” earned a mark in the author category), two blind raters marked student responses (see Table 7 for examples of student response paragraphs for each website). Students were marked out of four on which categories they mentioned in the paragraph, and out of 14 on which specific questions they chose and/or remembered to use in their rating explanation. Inter-rater reliability calculated from Pearson’s *r* was .71 (website 1) and .82 (website 2) for the four categories and .81 (website 1) and .88 (website 2) for the 14 questions. The final marks were determined by taking an average between the two raters. Massed and spaced groups did not differ on how many of the four categories were used to explain their rating in a paragraph, $t(714) = -1.63$, $p = .10$, $d = -.12$, $BF_{10} = .31$. Groups also did not differ in their use of the 14 questions in a paragraph, $t(714) = 1.07$, $p = .29$, $d = .08$, $BF_{10} = .15$, or on the initial ratings, $t(714) = .90$, $p = .29$, $d = .07$, $BF_{10} = .15$. Bayes factors suggested “substantial” evidence that the groups had equal performance at baseline (Jarosz & Wiley, 2014). Therefore, we proceeded with our analyses as planned.

Table 7.

Examples of student paragraphs for each website used during the lessons.

Sea Monkey Online (Student Rating: 8/10)

Firstly, the site has incorrect spelling on the name of the University, on the site, it is spelled "MacGill University" and not McGill University. Additionally, in the resources section, all the websites are listed (not cited in APA format), and under the website is an arbitrary picture of a shrimp that looks out of place. in addition, the pictures do not look professional, considering the fact that in one that there is an infant in one image holding a used package of "Sea Monkeys". On the site, the author is not recognized, only when you click on the "The Office of Science and Society" is he recognized. Everything looks relevant and all the information looks correct, however, there is one unprofessional video, made by what seems to be a video producer that does not seem to inform the audience, and is made for entertainment purposes only. Although the website itself is not trying to sell the product to the audience, there are multiple links to purchase them.

Brain Science (Student Rating: 0/10)

The authority of this website is terrible. When you search up the address of Brain Science Inc. all you see is a deserted area with something that looks like a hotel nearby. The author isn't mentioned, except for the beginning part in which they say a Dr. Daniel Reid wrote the first page. They claim he is an American neuroscientist, but he is actually a physician from Nova Scotia. I doubt he wrote that first page. The website's content is even worse than its authority. Almost every fact written on it isn't true, save for the first page where only the most elementary elements of the workings of the brain are listed. Otherwise the site is a mess as far as content goes. The website barely even talks about the science of the brain, and instead rants on and on about some miracle pill that is clearly just a placebo. The brain information that is there is for the must part inaccurate. It's been proven that we don't use only 10% of our brain, and we don't need a vitamin to make it work right. The testimonials are also fake. John Green doesn't even live in Ohio! Plus, some of the ingredients in Neuroflex are not even real things. Overall, I think the most important thing about any site is it's info. The information in this site was not true at all, which is why I rated it at a 0. Although the design could be rated at a 3 or 4, everything else is a 1 or 0. I would not trust this website at all.

Bizarre Animals (Student Rating: 4/10)

The design was a little childish, not very professional. But maybe it was meant for kids. It was most likely for educational purposes because the author isn't trying to sell anything, and there is good educational information on it. There were links that took you to websites that are irrelevant to the website's purpose and topic, so that was a red flag for me. I was confused. It says that the author is "staff," but I don't know who that is. There is no way of contacting the author so that is a huge red flag for me. Also, in one of the videos, it shows a gulper eel stretching its jaw. It looked very, very fake. It looked photoshopped. But I don't know for sure. So overall, I think it was for educational purposes, but there were a couple red flags for me. It contained some great information, but the website itself was a bit

unprofessional. I would rate it higher because of the information, but lower for its unprofessionalism.

Glow-in-the Dark Bunnies (Student Rating: 6/10)

The design of the website is a bit hard to read as there isn't any headings but there small is subheadings. The background of the page is a bright neon green and the text is a light green this makes the text a bit hard to read and could impair understanding or comprehension. The author is stated at the top and doing a little bit of research on her I found out that she is the editor of the MIT technology review and is somewhat of an expert in the subject which makes the information valid. The purpose of the page is to educate you and doesn't try to sell you anything. In most of the subheading if there is any sort of research done the source is cited and is given credit for.

Association of Geniuses (Student Rating: 3/10)

The design of the website is very professional, as the fonts, colour schemes and format are all neat, clear, and organised. The authority of the website is truly dreadful. There is no clear identification of the author, and the photos of members are just random stock images of children, with no identification. The only method of communication is an obviously anonymous email, but other than that, pretty much nothing is well in terms of credibility. The content is truly a sight to forget. The information about Eisenstein's childhood and equations are wrong. The website stated that E=MC^2 is the gravity equation. BUT, it is the relativity equation, relating matter and energy as 2 forms of the same thing! NOT GRAVITY! Along with that, the website stated that Einstein was bad at math in his childhood, and liked to draw. This is far from the truth, as Einstein loved math and science, he just was very inquisitive in the classroom. The only redeeming quality of the content is that the links and websites are actually realistic and truthful, but is that really good enough to fix up the twisted lies this website makes up? Finally, we have our purpose. The purpose is to get the reader to donate to help struggling students. To do this, they make up lies and unrealistic information to make it look like he was a struggling student, so you could help other struggling students. The website used the purpose of money and funding to lie, and convince the reader that these students are the next Einsteins, but really, they just want your money.

Hypothesis 1: Fact learning tested via recall

We predicted that students in the spacing condition, when asked directly what the four categories were (design, authority, content and purpose), would *recall* more from the lessons than students in the massed condition. As expected, students in the spaced condition ($M = 2.58$, $SD = 1.30$) recalled significantly more of the four categories than students in the massed condition ($M = 2.29$, $SD = 1.40$), $t(714) = -2.82$, $p < .01$, $d = .21$, $BF_{10} = 4.02$ (Figure 11).

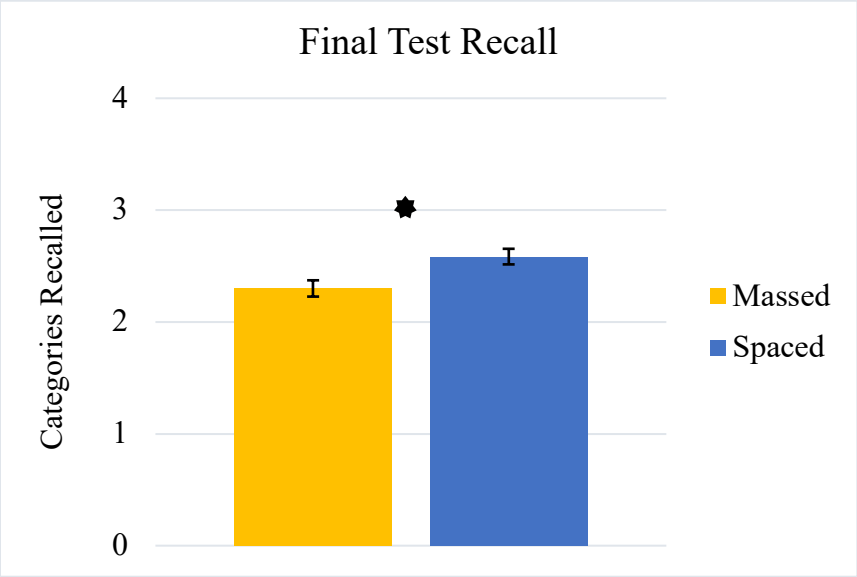


Figure 11. Final test recall (spaced, massed) when asked, "What are the four categories of website evaluation?" Error bars represent standard error.

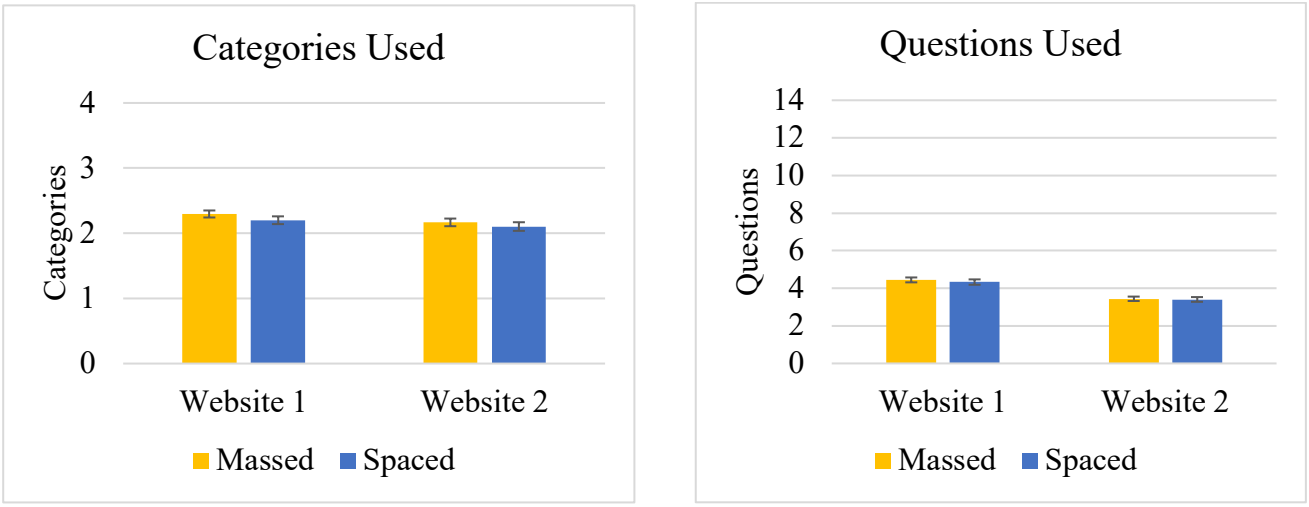
Note. A * indicates significance.

Hypothesis 2: Critical thinking tested via open-ended paragraph

We predicted that the spacing effect would benefit critical thinking: students in the spaced condition would spontaneously *use* more information in an open-ended paragraph, by giving details taught in the lessons to explain their website ratings than students in the massed condition. This was prompted by giving students two different websites at final test and simply asking for each, "Is this website credible? Please explain using evidence from the website." See Figure 12 for a visual representation of the data.

Four categories. Students in the spaced condition did not use more of the four categories to explain their rating than students in the massed condition for website 1, $t(711)= 1.2$, $p = .23$, $d = .09$, $BF_{10} = .17$, or website 2, $t(701.99) = .72$, $p = .47$, $d = .06$, $BF_{10} = .11$.

Fourteen questions. Students in the spaced condition did not use more of the 14 questions to explain their rating than students in the massed condition for website 1, $t(711) = .60$, $p = .55$, $d = .05$, $BF_{10} = .10$, or website 2, $t(699) = .20$, $p = .84$, $d = .02$, $BF_{10} = .86$.



(a) Categories used in the final test paragraph (b) Questions used in the final test paragraph

Figure 12. Categories (a) and questions (b) used in final test paragraph. Error bars represent standard error.

Hypothesis 3: Ratings during lessons

We predicted that students who were in the massed condition would be better at rating the websites *during* the daily lessons than spaced students in the weekly lessons, since that is often seen in the verbal and fact learning literature. Students were doing better if they were getting closer to the correct rating from 0-10 (shown in Figure 13). Our results indicated that there was no difference after learning occurred during lesson one, $t(668) = -.32$, $p = .75$, $d = .03$, $BF_{10} = .09$, and the spaced group ($M = 2.31$) performed better during lesson two than the massed group ($M = 2.84$), $t(678) = 3.64$, $p < .01$, $d = 0.29$, $BF_{10} = 53.23$, but there was a difference in ratings between the groups at lesson three with the massed group ($M = 1.68$) closer to the correct rating than the spaced group ($M = 2.00$), $t(628.6) = -2.59$, $p < .01$, $d = 0.20$, although the Bayes factor conflicts with the results of the t-test, suggesting indeterminate evidence, $BF_{10} = 2.27$. See raw student ratings (from 0-10) and difference from correct student ratings in Figures 13-14.

Hypothesis 4: Website ratings at final test

At final test, we predicted that students in the spaced condition would rate both websites more accurately than students in the massed condition. Although there were some differences in ratings during the lessons (Figure 13), our analyses indicated that there were no significant differences in ratings (out of 10) between the spaced and massed groups on the final test websites, for either website 1, $t(697) = -39$, $p = .70$, $d = .004$, $BF_{10} = .09$ or website 2, $t(685) = .19$, $d = .014$, $p = .85$, $BF_{10} = .09$. We also explored to see whether there was an effect of gender on the ratings, and there was no effect on website 1, $F(3, 712) = .48$, $p = .70$, $BF_{10} = .04$, or website 2, $F(3, 712) = 1.67$, $p = .18$, $BF_{10} = .25$ (Figure 15).

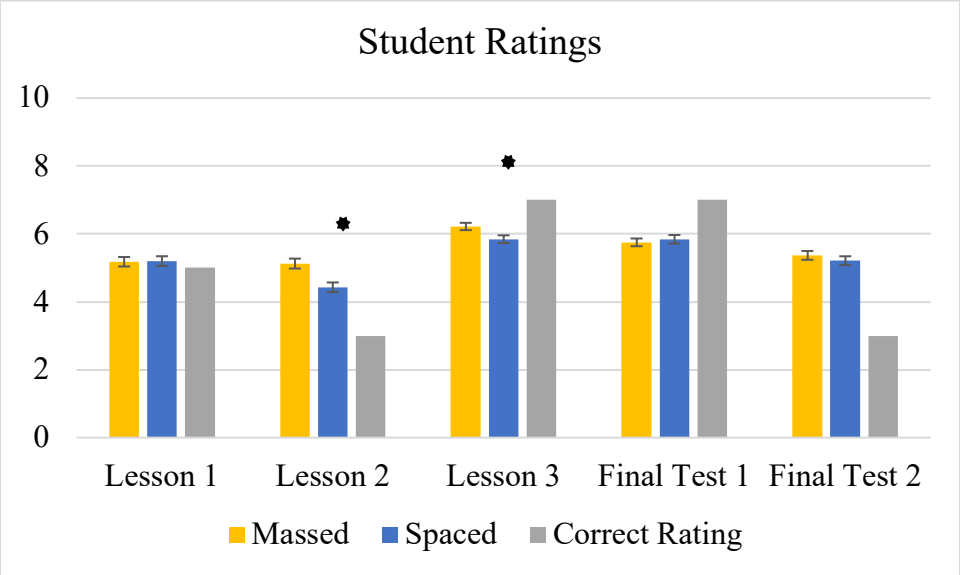


Figure 13. Student raw value ratings across the lessons.

Note. A * indicates significance at $p < .05$.

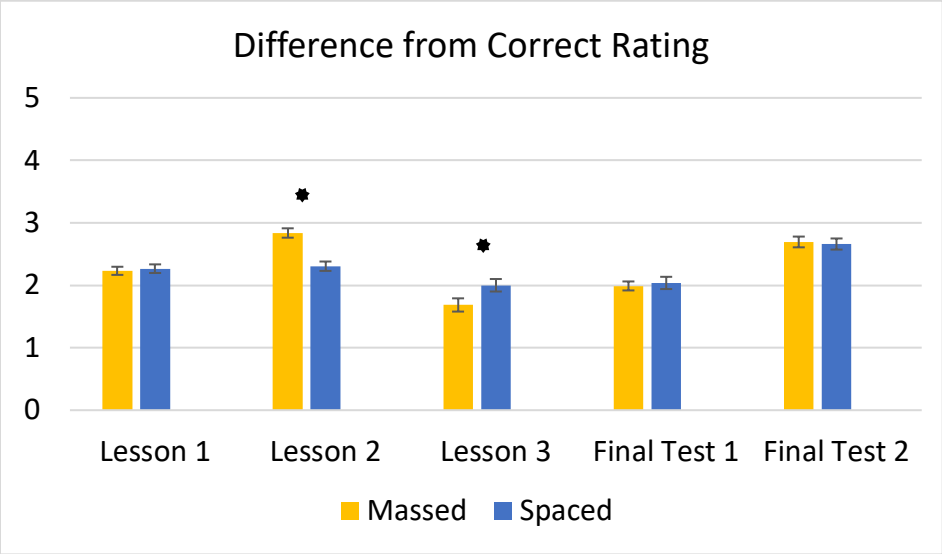


Figure 14. Student website ratings from across the lessons. Values represent differences from correct rating.

Note. A * indicates significance at $p < .05$.

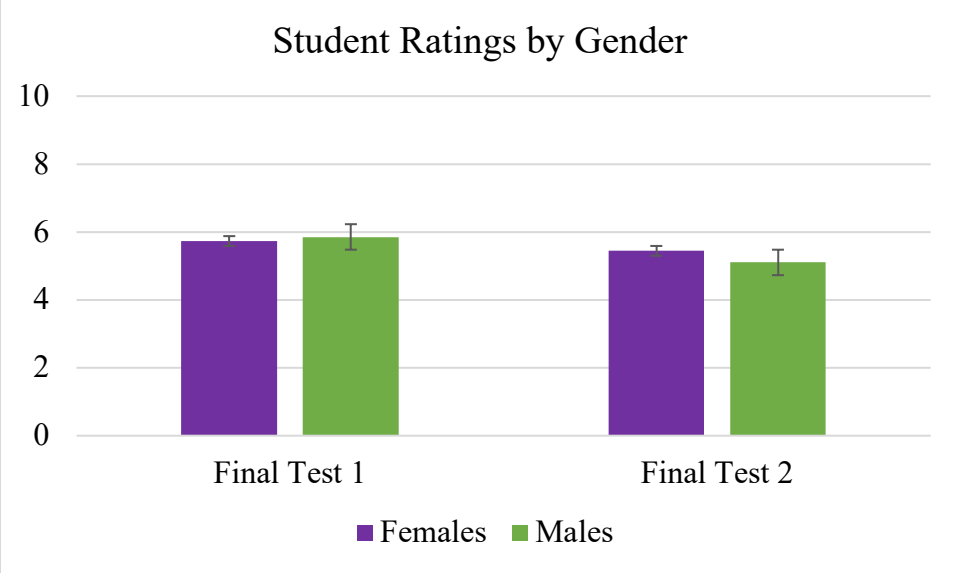


Figure 15. Student ratings by (self-identified) gender at final test. Only females and males are shown due to insufficient numbers in the "other" and "prefer not to answer" categories.

Exploratory Hypothesis

Students were asked which of the four categories they thought was the most important for determining website credibility (Table 8). Content was chosen as the most important category overall (n = 306), and those who deemed content as most important used it more often than the

other categories to support their rating at final test. However, most students used the content category to support their rating more than the other categories. Purpose was used the least.

Table 8.

Category chosen by students as most important for determining website credibility vs. what they actually used during final test to support their rating (websites 1 and 2). A high number of students described content as the most important category.

Most Important	Design Category Used (%)	Authority Category Used (%)	Content Category Used (%)	Purpose Category Used (%)
Design (n=96)				
Website 1	66.67	62.50	68.75	32.29
Website 2	52.08	48.96	75.00	31.25
Authority (n=140)				
Website 1	65.71	68.57	87.14	42.14
Website 2	64.29	64.29	85.71	37.86
Content (n=306)				
Website 1	73.11	66.23	88.85	42.30
Website 2	67.65	62.75	80.39	39.22
Purpose (n=139)				
Website 1	73.38	68.35	82.01	34.53
Website 2	68.35	58.27	79.86	41.01

Table 9.

Summary of data. Percentage accuracy scores for categories and questions in the spaced and massed conditions, at pre-test and at final test.

	Massed				Spaced			
	n	***M***	***SD***	***95% CI***	***n***	***M***	***SD***	***95% CI***
Pre-Test Paragraph								
Categories Used	367	.85	.697	.78, .92	349	.94	.723	.87, 1.02
Questions Used	367	1.43	1.21	1.30, 1.55	349	1.33	1.19	1.20, 1.46
Final Test Paragraph								
Website 1								
Categories Used	365	2.30	1.03	2.19, 2.40	348	2.20	1.11	2.09, 2.33
Questions Used	365	4.44	2.50	4.18, 4.69	348	4.43	2.67	4.06, 4.62
Website 2								
Categories Used	363	2.17	1.16	2.05, 2.28	340	2.10	1.22	1.97, 2.23
Questions Used	363	3.44	2.23	3.22, 3.68	340	3.40	2.33	3.15, 3.65
Final Test Recall								
Categories Recalled	367	2.30	1.40	2.16, 2.45	349	2.58	1.30	2.45, 2.73
Website Ratings								
Pre-Test	367	2.34	1.25	2.20, 2.48	349	2.26	1.30	2.12, 2.41
Lesson 1	346	2.23	1.40	2.06, 2.36	324	2.27	1.35	2.13, 2.44
Lesson 2	349	2.84	1.98	2.64, 3.07	331	2.31	1.82	2.12, 2.54
Lesson 3	352	1.68	1.36	1.48, 1.77	341	2.00	1.82	1.81, 2.23
Final Test 1	367	1.99	1.64	1.84, 2.21	349	2.04	1.66	1.83, 2.20
Final Test 2	363	2.70	2.10	2.39, 2.85	341	2.66	1.89	2.38, 2.80

Note: Website ratings are based on difference scores (student rating – correct rating).

Post Hoc Secondary Analyses of the Final Test

Grade Effects. Grade effects were looked at post-hoc in order to see where significance was occurring among the independent age groups. There was a significant effect of spacing for grade 6 students for categories recalled. The grade 5 spacing group reached significance for the categories used within the paragraphs for website 1, and grade 5's and 8's both saw a spacing effect on the categories used within the paragraph for website 2. See Figure 16-18 for visual representations. Table 10 contains the results of the t-tests and Bayesian analyses for each individual grade x ISI.

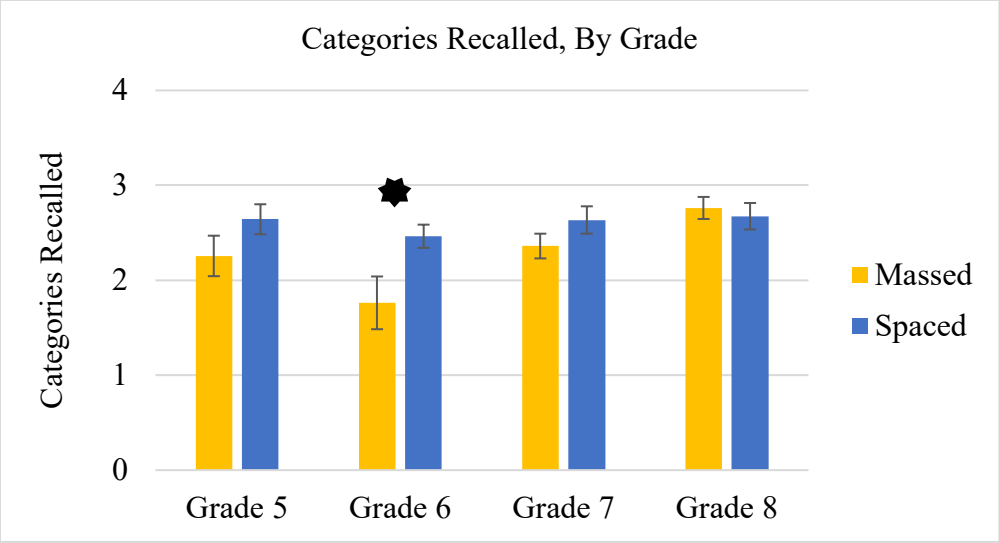
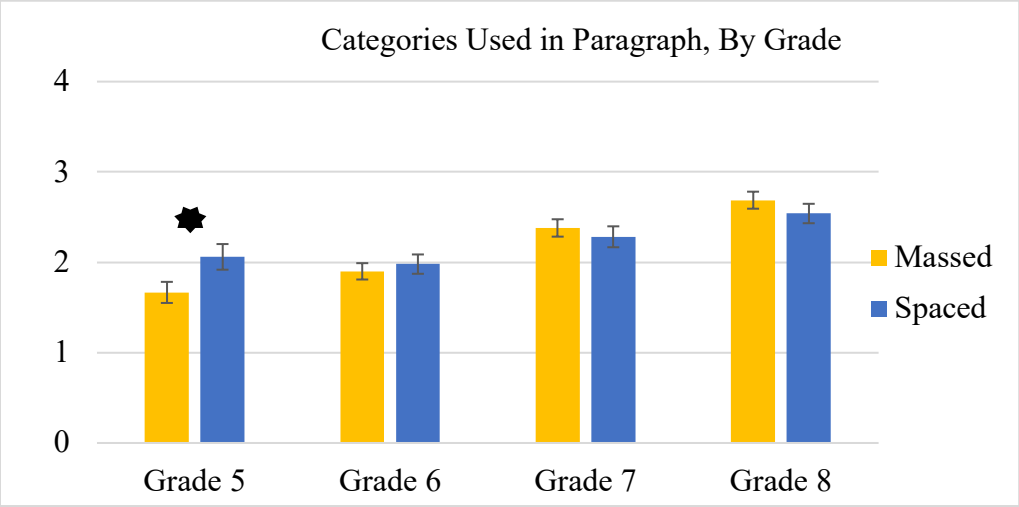
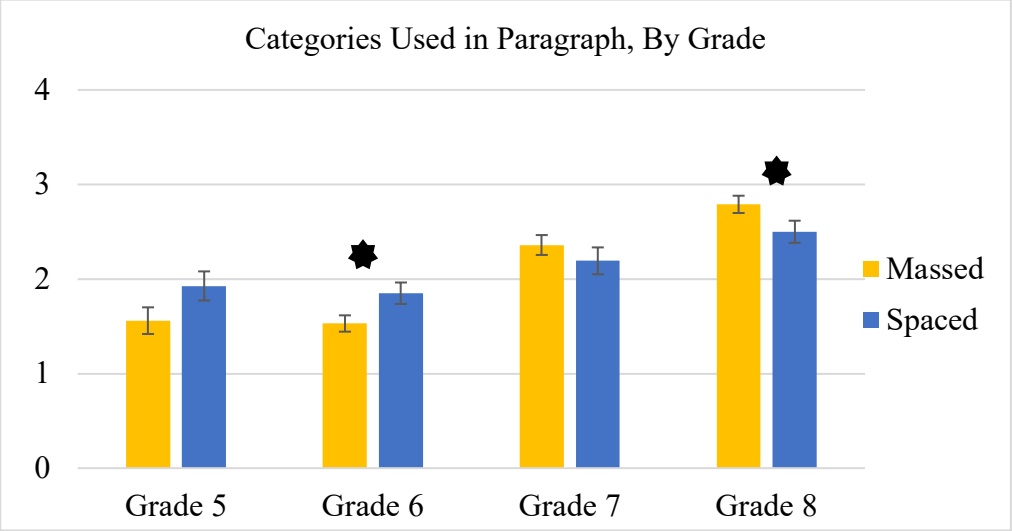


Figure 16. Final test recall (spaced, massed) when asked, "what are the four categories of website evaluation?". Spaced and massed groups have been separated by grade. Error bars represent standard error.

Note. A * indicates significance at $p < .05$.



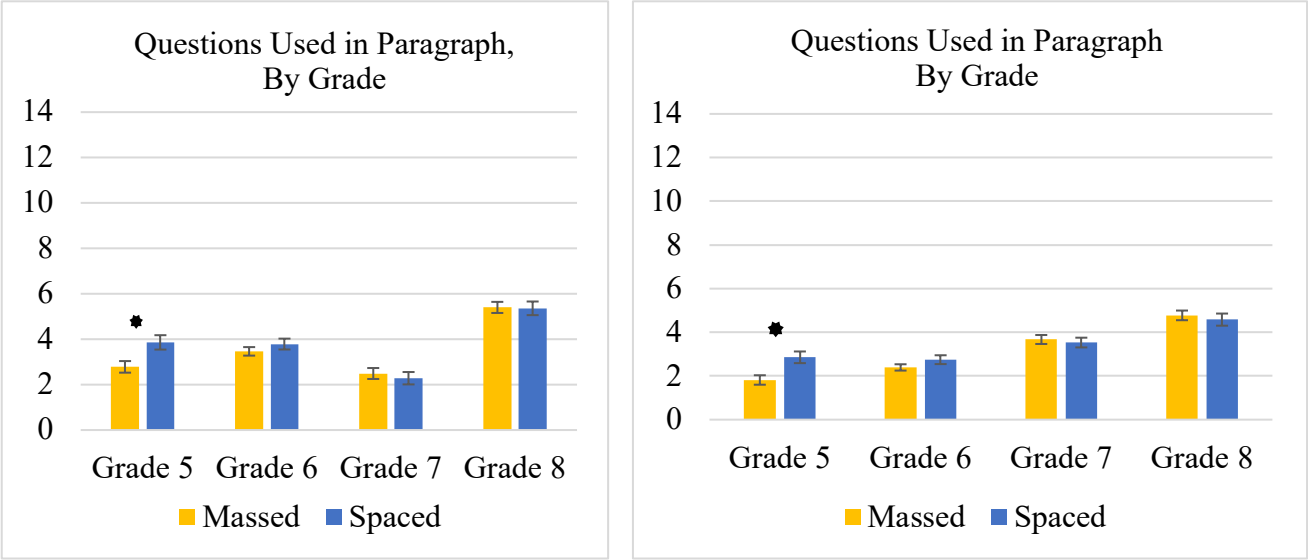
(a) Website 1



(b) Website 2

Figure 17. Use of four categories on final test (website 1) when asked, "is this website credible? Explain using evidence from the website." Spaced and massed groups have been separated by grade. Error bars represent standard error.

Note. A * indicates significance at $p < .05$.



(a) Website 1 (b) Website 2

Figure 18. Use of 14 questions on final test for both (a) website 1 and (b) website 2 when asked, "is this website credible? Explain using evidence from the website." Spaced and massed groups have been separated by grade. Error bars represent standard error.

Note. A * indicates significance at $p < .05$.

Table 10.

Statistical results by grade (spaced, massed).

Grade 5	**Massed**			**Spaced**					
	n	***M***	***SD***	***n***	***M***	***SD***	***p***	***t***	***BF*$_{10}$**
Pre-Test Paragraph									
Categories Used	39	.41	.55	56	.73	.67	.02*	-2.46	3.05^{+}
Questions Used	39	.72	.67	56	.98	.86	.11	-1.59	.66
Final Test Paragraph									
Website 1									
Categories Used	39	1.67	.73	55	2.06	1.05	.05	-1.99	1.23
Questions Used	38	2.78	1.59	55	3.86	2.35	.02*	-2.47	3.10^{+}
Website 2									
Categories Used	38	1.55	.87	54	1.93	1.13	.09	-1.71	.80
Questions Used	38	1.82	1.33	54	2.86	1.96	.06	-2.83	6.85^{+}
Final Test Recall									
Categories Recalled	39	2.26	1.33	56	2.64	1.18	.14	-1.49	.58
Website Ratings									
Pre-Test	39	2.10	1.25	56	2.39	1.40	.30	-1.04	.35
Lesson 2	33	2.09	3.24	56	1.70	2.43	.52	.65	.28
Lesson 3	33	2.24	1.66	55	1.87	1.36	.26	1.14	.40

Final Test 1	39	1.54	1.21	56	2.11	1.66	.07	-1.83	.94
Final Test 2	38	3.76	2.17	55	2.8	2.05	.03*	2.17	1.72
Grade 6	**Massed**			**Spaced**					
	n	***M***	***SD***	***n***	***M***	***SD***	***p***	***t***	***BF₁₀***
Pre-Test Paragraph									
Categories Used	101	.73	.58	124	.87	.71	.12	-1.58	.47
Questions Used	101	1.30	1.04	124	1.27	1.28	.84	.20	.14
Final Test Paragraph									
Website 1									
Categories Used	100	1.9	.91	124	1.98	1.19	.58	-.55	.17
Questions Used	100	3.47	1.87	124	3.78	2.67	.31	-1.02	.24
Website 2									
Categories Used	101	1.53	.89	120	1.85	1.24	.03*	-2.21	1.45
Questions Used	101	2.39	1.45	120	2.75	2.21	.17	-1.38	.36
Final Test Recall									
Categories Recalled	101	1.76	1.40	124	2.46	1.37	.01*	-3.76	98.28^{+}
Website Ratings									
Pre-Test	101	2.33	1.37	124	2.24	1.36	.63	.48	.16
Lesson 2	101	3.45	2.12	118	1.99	2.30	.01*	4.88	6947.27^{+}
Lesson 3	101	1.45	1.23	121	2.0	1.61	.01*	-2.79	5.55^{+}
Final Test 1	101	1.80	1.48	124	2.19	1.73	.08	-1.76	.63
Final Test 2	101	2.90	2.11	120	2.68	1.83	.40	.85	.21
Grade 7	**Massed**			**Spaced**					
	n	***M***	***SD***	***n***	***M***	***SD***	***p***	***t***	***BF₁₀***
Pre-Test Paragraph									
Categories Used	122	.86	.71	85	.94	.62	.39	-.85	.22
Questions Used	122	1.37	1.11	85	1.08	.90	.08	1.74	.64
Final Test Paragraph									
Website 1									
Categories Used	122	2.49	1.06	85	4.40	2.52	.17	1.37	.37
Questions Used	122	4.96	2.66	85	4.4	2.52	.13	1.53	.46
Website 2									
Categories Used	120	2.36	1.15	83	2.19	1.30	.34	.96	.24
Questions Used	120	3.68	2.25	83	3.54	2.0	.64	.46	.27
Final Test Recall									
Categories Recalled	122	2.36	1.44	85	2.64	1.33	.16	-1.40	.38
Website Ratings									
Pre-Test	122	2.60	1.22	85	2.18	1.29	.02*	2.38	2.15
Lesson 2	122	2.40	1.32	81	2.27	1.43	.25	1.16	.29

Lesson 3	120	1.72	1.22	82	1.93	1.62	.29	-1.05	.26
Final Test 1	122	2.11	1.68	85	1.99	1.64	.61	.50	.17
Final Test 2	120	2.71	2.12	83	2.83	2.05	.68	-.41	.17

Grade 8	**Massed**			**Spaced**					
	n	***M***	***SD***	***n***	***M***	***SD***	***p***	***t***	***BF₁₀***
Pre-Test Paragraph									
Categories Used	105	1.12	.73	84	1.18	.81	.63	-.49	.18
Questions Used	105	1.92	1.43	84	1.92	1.29	.97	.04	.16
Final Test Paragraph									
Website 1									
Categories Used	104	2.69	.96	84	2.54	.98	.29	1.07	.27
Questions Used	104	5.40	2.49	84	5.37	2.75	.93	.09	.16
Website 2									
Categories Used	104	2.79	.93	83	2.45	1.07	.04*	1.05	1.11
Questions Used	104	4.77	2.26	83	4.77	2.26	.59	.54	.183
Final Test Recall									
Categories Recalled	105	2.76	1.89	84	2.68	1.26	.64	.47	.18
Website Ratings									
Pre-Test	105	2.15	1.21	84	2.27	1.13	.48	-.71	.30
Lesson 2	104	1.74	2.35	78	2.42	1.07	.10	1.66	.57
Lesson 3	98	1.70	1.42	83	1.34	1.52	.15	1.45	.43
Final Test 1	105	2.10	1.68	84	1.82	1.57	.25	1.14	.29
Final Test 2	104	2.10	1.80	83	2.36	1.69	.31	-1.03	.26

Note: A * indicates significance when tested at $p < .05$. A $^{+}$ indicates a Bayes factor that is more than indeterminate (supports H1). A score of 0.33-3 indicates indeterminate evidence, 10-30 indicates strong evidence, 30-100 indicates very strong evidence, and 100+ indicates extreme evidence; values below 0.33 reflect evidence in favor of H0, with increasingly small values representing increasingly strong evidence.

Classroom Effects. In order to address our violation of the independence of observations assumption, we ran an additional analysis post-hoc with each of the 41 classes functioning as an independent data point instead of the previous analyses which had students as the independent data points (Table 11). There was no effect of class on the four category recall measure, $t(30.99) = -1.20$, $p = .23$, $d = .39$, $BF_{10} = .55$; on the use of the four categories for website 1, $t(39) = .53$, $p = .60$, $d = .16$, $BF_{10} = .35$, or website 2, $t(39) = .53$, $p = .59$, $d = .17$, $BF_{10} = .34$; or the use of the

questions for website 1, $t(39) = -.37$, $p = .71$, $d = .12$, $BF_{10} = .32$, or website 2, $t(39) = .37$, $p = .59$, $d = .11$, $BF_{10} = .32$.

Table 11.

Means of classrooms x all dependent variables at final test, by class. Massed classrooms have been coloured yellow and spaced groups have been coloured blue.

Class	Four Category Recall	Category Use Website 1	Category Use Website 2	Question Use Website 1	Category Use Website 2	Rating Difference Website 1	Rating Difference Website 2
1	1.57 (n= 29)	2.32 (n= 28)	1.88 (n= 29)	4.68 (n= 28)	2.62 (n= 29)	1.59 (n= 29)	3.45 (n= 29)
2	2.47 (n= 19)	2.12 (n=19)	2.26 (n= 19)	5.26 n= 19	3.63 (n= 19)	1.84 (n= 19)	3.37 (n= 19)
3	3.07 (n= 14)	2.93 (n= 14)	2.64 (n= 14)	7.07 (n= 14)	4.79 (n= 14)	1.43 (n= 14)	2.21 (n= 14)
4	3.19 (n= 16)	2.44 (n= 16)	2.63 (n= 16)	5.63 (n= 16)	3.94 (n= 16)	1.94 (n= 16)	2.0 (n= 16)
5	1.92 (n= 15)	1.33 (n= 15)	1.58 (n= 12)	3.13 (n= 15)	2.42 (n= 12)	2.20 (n= 15)	1.92 (n= 12)
6	1.95 (n= 19)	2.32 (n= 19)	2.16 (n= 19)	5.42 (n= 19)	4.53 (n= 19)	1.63 (n= 19)	2.32 (n= 19)
8	1.50 (n= 18)	2.00 (n= 18)	1.56 (n= 18)	3.39 (n= 18)	2.67 (n= 18)	1.72 (n= 18)	3.11 (n= 18)
9	1.72 (n= 18)	1.47 (n= 18)	1.28 (n= 18)	2.69 (n= 18)	2.0 (n= 18)	2.11 (n= 18)	3.83 (n= 18)
10	2.23 (n= 14)	1.08 (n= 13)	.69 (n= 13)	1.92 (n= 13)	1.31 (n= 13)	2.57 (n= 14)	3.29 (n= 14)
11	2.27 (n=15)	2.17 (n= 15)	2.0 (n= 15)	3.67 (n= 15)	3.27 (n= 15)	2.40 (n= 15)	2.0 (n= 15)
12	2.13 (n= 16)	1.88 (n= 16)	1.19 (n= 16)	2.41 (n= 16)	1.94 (n= 16)	2.25 (n= 16)	2.44 (n= 16)

13	2.92 (n= 13)	2.62 (n= 13)	2.69 (n= 13)	4.08 (n= 13)	2.92 (n= 13)	1.62 (n= 13)	3.92 (n= 13)
14	2.29 (n= 17)	2.53 (n= 17)	2.06 (n= 17)	4.79 (n= 17)	2.82 (n= 17)	1.53 (n= 17)	3.0 (n= 17)
15	.81 (n= 21)	1.60 (n= 21)	1.10 (n= 21)	2.31 (n= 21)	1.67 (n= 21)	1.86 (n= 21)	3.76 (n= 21)
16	2.60 (n= 15)	1.43 (n= 15)	.47 (n= 15)	2.37 (n= 15)	1.60 (n= 15)	2.0 (n= 15)	4.33 (n= 15)
17	2.57 (n= 14)	1.64 (n= 14)	1.5 (n= 14)	2.75 (n= 14)	2.43 (n= 14)	2.07 (n= 14)	3.36 (n= 14)
18	3.43 (n= 23)	2.28 (n= 23)	3.17 (n= 23)	6.67 (n= 23)	5.0 (n= 23)	1.17 (n= 23)	2.64 (n= 22)
19	2.32 (n= 22)	1.34 (n= 22)	1.23 (n= 22)	2.52 (n= 22)	1.63 (n= 22)	2.73 (n= 22)	3.27 (n= 22)
20	.94 (n= 17)	1.97 (n= 17)	1.35 (n= 17)	3.32 (n= 17)	2.0 (n= 17)	1.71 (n= 17)	2.06 (n= 17)
21	2.80 (n= 20)	1.80 (n= 20)	1.48 (n= 20)	2.75 (n= 20)	1.45 (n= 20)	1.70 (n= 20)	4.25 (n= 20)
22	2.72 (n= 18)	2.97 (n= 18)	2.56 (n= 18)	5.28 (n= 18)	4.0 (n= 18)	1.94 (n= 18)	3.06 (n= 18)
23	2.82 (n= 17)	3.35 (n= 17)	3.02 (n= 17)	6.65 (n= 17)	4.77 (n= 17)	2.59 (n= 17)	1.94 (n= 16)
24	2.71 (n= 7)	2.57 (n=7)	2.5 (n= 7)	4.21 (n= 7)	3.43 (n= 7)	2.86 (n= 7)	3.14 (n= 7)
25	3.17 (n= 6)	1.833 (n= 6)	3.00 (n= 6)	3.0 (n= 6)	5.17 (n= 6)	3.0 (n= 6)	2.0 (n= 6)
26	2.89 (n= 20)	3.0 (n= 19)	2.83 (n= 20)	5.24 (n=19)	4.15 (n= 20)	1.75 (n= 20)	1.4 (n= 20)
27	3.12 (n= 17)	2.09 (n= 17)	1.71 (n= 17)	3.62 (n= 17)	2.74 (n= 17)	1.94 (n= 17)	2.53 (n= 17)

28	1.92 (n= 13)	1.5 (n= 13)	1.79 (n= 12)	2.85 (n= 13)	2.5 (n= 12)	1.31 (n= 13)	3.58 (n= 12)
29	2.55 (n= 21)	2.10 (n= 21)	2.45 (n= 20)	4.17 (n= 21)	3.45 (n= 20)	1.86 (n= 21)	2.62 (n= 21)
30	2.88 (n= 25)	2.82 (n= 25)	2.88 (n= 24)	6.54 (n= 25)	5.40 (n= 24)	2.48 (n= 25)	1.92 (n= 24)
31	3.17 (n= 25)	2.83 (n= 25)	2.88 (n= 24)	6.28 (n= 25)	4.92 (n= 24)	2.84 (n= 25)	1.63 (n= 24)
32	1.88 (n= 17)	2.38 (n= 17)	2.47 (n= 17)	4.52 (n= 17)	3.97 (n= 17)	2.29 (n= 17)	2.18 (n= 17)
33	2.12 (n= 17)	1.53 (n= 17)	1.44 (n= 17)	3.18 (n= 17)	2.29 (n= 17)	1.71 (n= 17)	2.24 (n= 17)
34	3.29 (n= 17)	3.29 (n= 17)	3.5 (n= 17)	7.5 (n= 17)	6.41 (n= 17)	1.94 (n= 17)	1.88 (n= 17)
35	2.52 (n= 25)	2.58 (n= 25)	2.76 (n= 25)	5.3 (n= 25)	5.36 (n= 25)	1.72 (n= 25)	2.04 (n= 25)
36	2.53 (n= 17)	1.85 (n= 17)	2.06 (n= 17)	3.5 (n= 17)	3.12 (n= 17)	1.94 (n= 17)	1.94 (n= 17)
37	3.14 (n= 15)	2.30 (n= 15)	2.21 (n= 14)	3.87 (n= 15)	3.04 (n= 14)	2.33 (n= 15)	2.79 (n= 14)
38	2.55 (n= 22)	2.68 (n= 22)	3.11 (n= 22)	5.86 (n= 22)	6.0 (n= 22)	1.64 (n= 22)	1.91 (n= 22)
39	2.15 (n= 14)	1.39 (n= 14)	1.84 (n= 13)	2.5 (n= 14)	2.69 (n= 14)	3.07 (n= 14)	2.0 (n= 14)
40	2.60 (n= 17)	2.56 (n= 17)	2.76 (n= 15)	5.56 (n= 17)	5/57 (n= 15)	2.12 (n= 17)	1.8 (n= 15)
41	2.47 (n= 15)	2.23 (n= 15)	2.03 (n= 15)	3.43 (n= 15)	2.67 (n= 15)	2.40 (n= 15)	2.07 (n= 15)
42	2.53 (n= 16)	2.22 (n= 16)	1.97 (n= 15)	3.53 (n= 16)	2.57 (n= 15)	2.25 (n= 16)	4.07 (n= 15)

Conclusions

The main goal of this study was to see whether the robust spacing effects seen in the laboratory (particularly using fact and verbal learning) could also be seen in the classroom under real-world conditions, using curriculum-based materials involving critical thinking. There were both expected, and unexpected findings, which was not surprising since effectiveness trials by nature are intended to account for all of the external factors that could potentially decrease an intervention's effect and lessen the effect size. Our finding that supported the decades of fact learning literature was the spacing effect for the four category recall—Figure 11 demonstrates that the spaced group overall had moderate evidence for a spacing effect. When analysed separately, however, it appears that the grade six group may have been driving this effect since they had very strong evidence in favour of a spacing effect, $BF_{10} = 98.28$, where the other classes were indeterminate (Figure 16). Grade effects will be expanded upon later in this section. The decreased effect size seen in the category learning ($d = .21$) compared to other similar classroom studies which have a mean effect size for fact learning of $d = .47$ (Carpenter et al., 2009, Kapler et al., 2015) was likely due to all of the items discussed in Table 3—all of the uncontrolled side effects that come with running an effectiveness trial. There was an effect size of $d = .85$ (with the Bayes factor indicating extreme evidence for a group difference) on this exact measure in the efficacy trial (Foot-Seymour et al., 2019).

Although traditional spacing studies only contain two study sessions, we added a third session so that students could have another opportunity to practice their website evaluations. This mimics standard teaching practice and was also done in the efficacy trial (Foot-Seymour et al., 2019). This decision had some repercussions—for example, students were removed from the study if they missed a lesson, and adding another session increased the likelihood that it would

happen. On the other hand, it provided us with more of an opportunity to explore within-lesson trends, which are seen within both the spacing (e.g., Mozer et al., 2009) and desirable difficulties (e.g., Bjork et al., 2011) literature, suggesting that students who take part in the daily lessons should have done better immediately whereas students in the weekly lessons should have had more of a struggle to learn the content due to the time delay. We found some evidence of these expected trends, because by the time students reached the third lesson, the massed group performed better on the website ratings than the spaced group, although there was a discrepancy here between our traditional hypothesis testing which showed significance at $p < .05$, and the Bayes factor which was indeterminate. However, we did not anticipate finding a reverse spacing effect, with the spaced group rating the website more accurately than the massed group during lesson two, with the Bayes factor indicating a moderate effect. Our addition of a third study session makes this slightly more difficult to compare to the rest of the spacing literature.

According to our lesson plans, by the end of the lessons (either daily or weekly), students "would be able to effectively judge the credibility of websites. They would be skeptical of the websites they saw and be able to use collected evidence via the website evaluation checklist (a.k.a. the "scavenger hunt") to *explain* their credibility ratings. The evidence that they collected would be used to support a final rating of 0-10 (0-4 is not credible; 5 is neutral; 6-10 is credible), and while rating accuracy would improve throughout the lessons, they would learn that the most important aspect is the *process* that led them to their final decision" (Appendix A). We designed materials that would help students to achieve these learning goals, but were surprised that during the final test, although the four category recall demonstrated that students knew more of what to say, it appeared that students chose not to *use* some of the questions and categories when explaining their ratings in a paragraph. It may not have been an issue of whether they knew the material, but whether the websites accurately encouraged students to use that information. The validity of the materials was not in question

originally, since the study by Foot-Seymour et al. (2019) found an effect (extreme evidence for a group difference) of both question and category use. It is unknown if a confounding variable like the lack of control, grade effects, or class and teacher effects hid an effect that might have been present under different conditions. It is also possible that this was a partial failure of teaching the concepts with enough repetition, and the videos might not have explained well enough that it was necessary for students to consider *all* of the categories and questions—so instead they simply chose to talk about the most obvious aspects of the website.

Another possible reason why we failed to find a benefit in the paragraphs that there was a difference between the online video teaching and the traditional in-person teaching that resulted in a larger effect size in Foot-Seymour et. al (2019). While some studies have shown no difference in learning between online vs. in-person lessons (e.g., one of the largest studies conducted was by Russel, 1999), other research has shown that success in an online course is very much dependent upon the nature of student to student and student to teacher interaction (Piccaino, 2002). Therefore, it should not be automatically assumed that online and in-person teaching should result in the same learning outcomes (Manning-Oullette and Black, 2017).

Some important aspects to consider when designing an online course is planning intentional interactions and ensuring clarity of design (Ally, 2014, Piccaino, 2002), both of which were considered when designing lessons for the current study. Students have also been shown to enjoy an online learning environment more than traditional in-person teaching, since it can promote learning that can be less intimidating and encourage participation and meaningful interactions (Ni, 2013). This was qualitatively observed by teachers and volunteers during the current study—feedback from students that was collected after each lesson demonstrated that most enjoyed the online nature of the lessons. However, since there was no person on the other

side of the screen and the majority of the information was given passively via YouTube videos, we cannot be sure that students were fully engaged and absorbing the material unless the teacher was monitoring it on their own. This could have impacted the results of the study, since research has demonstrated that one of the biggest predictors of academic success is student engagement—particularly engagement with peers, the teacher, and the course material (Reiken, Dotson, Carter and Griffith, 2018).

Due to this possible lack of engagement, students might not have learned the 14 questions to threshold, and there may have been a reduction in the effect size for the four category learning. Without having full control over teaching the initial learning and final discussion (Appendix A), it was difficult to know how students were handling the concepts while the lessons took place. For example, the initial learning stage of each lesson was very important to determine how close the students were to achieving their learning goals. Students were broken down into small groups where they had to write down the categories as headings and then brainstorm as many specific questions as they could remember. The goal of this activity was to get students to struggle slightly to remember, which should have strengthened their memory trace. However, if students were not remembering at all, teachers and volunteers were instructed to cue them. If this activity was not carried out as planned, students' memory for the questions and categories would have been much weaker than expected.

While it is true that students may not have learned the content to threshold, it is also possible that the opposite is true, and that the materials and corresponding critical thinking measures (the final test website paragraphs) were too easy for some students. Evidence for this idea comes from the grade differences that we saw—spacing effects were seen more frequently for the younger grades and less for the older grades. Historically, spacing advantages are

strongest when individuals have the most to learn, and although spacing benefits have been seen throughout the lifespan, it is possible that using the same websites and marking schemes for grades 5-8 may have led to an increase in learning for the younger grades, but not for the older grades. The process of website evaluation itself was used in Bronstein (2007) with high school students, which is one of the reasons why this particular age group was chosen, but it may have been more effective to increase website difficulty with age. This would have presented some methodological challenges, since spacing effects would have to be explored separately if different websites were used with different age groups.

Another question raised by these results are whether domain for spacing matters. To expand on an earlier point—we know that based on previous research that the magnitude of spacing effects likely depends on type of content or the skill that is being learned (for a full review, see Wiseheart et al., 2019). In the literature, five types of learning have been summarized: intellectual skills, verbal information, cognitive strategies, motor skills, and attitudes (Gagné, 1977; 1984), but only three have been investigated in regard to the spacing effect. Verbal information consisting of facts, ideas and information, has been shown to work very well with spacing (Wiseheart et al., 2009). Motor skills, on the other hand, which are organized sets of actions (e.g., learning the piano or playing a sport) have had less consistent results (e.g., Donovan & Radeovich, 1999, Simmons, 2001, Wiseheart, D'Souza, & Chae, in preparation). Lastly, and most related to the current study, is research on intellectual skills. As previously explained, seeing whether spacing benefits intellectual skills is still an underexplored domain, especially since critical thinking and website evaluation, although falling into the intellectual skills category, could potentially cross into more of Gagné's (1977, 1984) types of learning. Website evaluation certainly includes verbal learning but based on our definition of

critical thinking, it also crosses into the cognitive strategies (since students learned a *skill*), and attitudes domains. The difficulty in generalizing results seen from this study is that we don't know with certainty where critical thinking falls on the learning spectrum, and it is not as simple to define as some of the other topics which have been shown to work well with the spacing effect. More research is needed in all five of these domains, separately and together, to investigate if there is a boundary where spacing effects are most effective.

Challenges and Limitations

The main challenge and limitation of this study was one that is present in all classroom studies—there was a lack of scientific control. Each class is composed of its own group of individuals who have different social, emotional, and academic needs. During recruitment, we requested that teachers let all students in the room participate if possible, which increased the variability of our sample, albeit in the way that we wanted since it was an effectiveness trial. Teachers also vary widely across classrooms. Although all were certified by the Ontario College of Teachers, there were teachers of varying levels of experience (one mother-daughter duo, for example). Teachers also have different teaching styles and have different personalities and approaches. We tried to include a sufficient number of teachers in each condition so that mean teacher experience and effectiveness would be the same in each experimental condition.

There are interruptions in a typical school day—during the 129 lessons that took place as part of this study, we had four snow days, fire drills, a power outage, and several interruptions of Internet service. There were several times when teachers had an interruption to their day and had to reschedule one or all of the lessons. In order to keep track of these interruptions when lessons were teacher-led, volunteers were sent to the classroom as much as possible. There were also changes in context from class to class. Some teachers opted to run the lessons in their normal

classroom, others in the computer lab or library, and each school varied in terms of space. At every stage, teachers were given the freedom to use their normal practice. All of these things were expected, and we prepared for it by collecting a large sample size, hoping that the random assignment of classrooms would handle these differences, but there may have been confounds which affected results or lessened our effect size.

Lesson plans and online materials were released to teachers with minimal additional instruction with the expectation that they would carry out the intervention as planned. Lesson plans were created to follow typical practice and could have been followed easily by any certified teacher. Most of the time, participating teachers did what they were asked to do and took interest in the research. Other times, it was clear that they had originally agreed because they saw value in the research topic but had not read the lesson plans in advance and therefore did not remember that they were responsible for some of the teaching. In order to manage this right away and add value to the lessons, the lead researcher (VF) ran the lesson on day one to model the day for teachers and to show them that since the majority of the lessons were online, there was minimal in-person teaching required. This was not originally planned and reduced time that was originally intended for recruitment and administration.

There was a difference not only in lesson preparedness but also in teacher enthusiasm for the lessons. Enthusiasm carried across from teachers to their students—classes where teachers were keen and prepared seemed to be more engaged. Students from these classes were more likely to write in comments about how much they enjoyed the lessons. This may have also affected the results, since if the teachers did not supplement the online videos with a quality introduction activity and discussion, students would not have been as engaged. There was no evidence for class effects within the 41 analysed classes (when group means were analysed), but

sample sizes for those two groups were small due to the nature of the analysis. There were certainly differences, and therefore added noise, between classes because 42 (including the class that was excluded from the analyses) classes meant 42 different teachers.

One of the other limitations of this particular study, in comparison to the previous study (Foot-Seymour et al., 2019), was the potential bias involved in creating "correct ratings" for the websites. Pre-existing websites are subject to change during the intervention, so this time the websites were designed by the lead researcher, along with a group of volunteers. This was done so that website ratings could be standardized, and all aspects could be controlled. However, some of the "correct" ratings still seemed to be subjective—an example of this was in the Glow-in-the-Dark Bunnies website where the design was a bright green background with red writing. The design was intended to be poor so that when students were asked "if the photos and/or colour choices look professional," the answer would clearly be no. However, many commented that the bright green background was fitting for a website about an animal that glows.

A number of administrative challenges came with the study. For example, websites could not be counterbalanced. Lesson plans with website links were distributed months in advance for teacher review, so that they knew which sites students were doing on which day. There were multiple times when lessons ran simultaneously, so it would have been difficult to administer one classroom with website 1 (link A), and another classroom with website 2 (link B). Teachers needed to know up front what the procedure was for teaching the lessons, so they could not be blinded to conditions: they knew whether they were in the massed or spaced condition. We could not blind them to the hypotheses, a requirement of the external research board for York Region. We made every effort to not bias student marks, and teachers were allowed to request a copy of the data, so if there were teaching partners, for example, teachers needed to know that classroom

A (massed) might have performed better *during* the lessons and classroom B (spaced) might have performed better *after* the month had passed. However, it was not concerning that teachers were not blinded to the main hypotheses, since each was responsible for their own classroom and it was unlikely that they would alter their teaching practice to cause any bias.

Recommendations for Future Research

Implementation of the curriculum is something that should be considered whenever planning long-term learning goals for the school year. If the goal of for students is to retain as much information as possible, teachers need to be aware of cognitive strategies like spacing, so that they can make small changes to their teaching practice to help students become more successful. A possible barrier to this might be that teaching resources are cumulative, and teachers often use similar materials from year to year, so asking them to change their plans entirely could be intrusive and intimidating. The benefit of using spacing is that the only adjustment that needs to be made is in the timing of long-range plans. This is not only achievable but would be beneficial to both students and teachers by saving time in the long run.

As a next step for spacing effect, researchers should be repeating effectiveness trials with different subject material and a wide scale of measurements. An intervention using fact learning has never been done with multiple subjects, using multiple dependent variables that have been tested for validity in other learning studies. Critical thinking was a major focus in the current study, which is why website evaluation was used, and it served to enhance our recruitment because it added value to teaching programs that were already in place. However, by using other curriculum-based subject material (perhaps something similar to the standardized EQAO, or Educational Quality Assessment Office, tests which create questions using the levels of Bloom's

Taxonomy), the results that we saw during the fact learning measures can be better evaluated and used to recommend spacing in a real-world classroom setting.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Zhang, D. A. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage one meta-analysis. Review of Educational Research, 78, 1102–1134. https://doi.org/10.3102/0034654308326084

Ally, M. (2014). Foundations for educational theory for online learning. In Anderson, T. Elloumi, F (Eds.), *Theory and Practice of Online Learning.* Athabasca, AB: Athabasca University.

Alston, K. (2001). Re/thinking critical thinking: The seductions of everyday life. Studies in Philosophy and Education, 20, 27–40. https://doi.org/10.1023/A:1005247128053

Angeli, C. & Valanides, N. (2009). Instructional effects on critical thinking: Performance on ill-defined issues. *Learning and Instruction. 19, 322-334.* https://doi.org/10.1016/j.learninstruc.2008.06.010

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 56-64.

Bloom B. S. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain.* New York: David McKay Co, Inc.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research, 74,* 245-248. https://doi.org/10.1080/00220671.1981.10885317

Boushey, G., & Moser, J. (2014). *The daily 5: fostering literacy independence in the elementary grades.* Portland: Stenhouse.

Boyd, N. L. (2019). Using Argumentation to Develop Critical Thinking About Social Issues in the Classroom: A Dialogic Model of Critical Thinking Education. In *Handbook of Research on Critical Thinking and Teacher Education Pedagogy* (pp. 135-149). IGI Global.

Bronstein, D. M. (2007). The Efficacy of a Web Site Evaluation Checklist as a Pedagogical Approach for Teaching Students to Critically Evaluate Internet Content (Unpublished doctoral dissertation). The Graduate School of Computer and Information Sciences Nova Southeastern University.

Bruner, J. S. (1960). On learning mathematics. *The Mathematics Teacher*, *53*(8), 610-619.

Byrnes, J. P., & Dunbar, K. N. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review*, *26*(4), 477-493. https://doi.org/10.1007/s10648-014-9284-0

Burbules, N. C., & Callister, T. A. (2000). Universities in transition: The promise and the challenge of new technologies. *Teachers College Record*, *102*(2), 271-93. https://doi.org/10.1111/0161-4681.00056

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760-771. doi:10.1002/acp.1507

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236-246. https://doi.org/10.1027/1618-3169.56.4.236

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095-1102. doi:10.1111/j.14679280.2008.02209.x

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation*, *53*, 63-147. doi:10.1016/S0079-7421(10)53003-2

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In *Memory* (pp. 317-344). Academic Press. https://doi.org/10.1016/B978-012102570-0/50011-2

DeRemer, P., & D'Agostino, P. R. (1974). Locus of distributed lag effect in free recall. *Journal of Verbal Learning and Verbal Behavior, 13*, 167-171. https://doi.org/10.1016/S0022-5371(74)80041-1

Descours, K. (2013). 21st Century Pedagogy: A Classroom Perspective on Critical Thinking (Unpublished master's thesis). York University.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect. *Journal of Applied Psychology, 84*, 795-805. https://doi.org/10.1037/0021-9010.84.5.795

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology* (H. A.

Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications. (Original work published 1885).

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron, & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York: W. H. Freeman and Company.

Ennis, R. H. (1996). Critical thinking. Upper Saddle River, NJ: Prentice-Hall.

Ennis, R.H. (2018). Critical thinking across the curriculum: A vision. *Topoi* 37 (1):165-184. https://doi.org/10.1007/s11245-016-9401-4

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations. Newark, DE: American Philosophical Association. (ERIC Document Reproduction Service No. ED315423)

Facione, P., Sánchez, C., Facione, N., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education, 44*(1), 1-25.

Foot, V. L. (2016). Judging Credibility: Can Spaced Lessons Help Students Think More Critically Online? (Unpublished Masters Thesis). York University.

Foot-Seymour, V., Foot, J., & Wiseheart, M. (2019) Judging credibility: Can spaced lessons help students think more critically online? *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3539

Fullan, M. (2013). *Great to excellent: Launching the next stage of Ontario's education agenda*. Retrieved from: http://www.edu.gov.on.ca/eng/document/reports/FullanReport_EN_07.pdf

Gagné, R. M. (1977). The conditions of learning (third ed.). New York: Holt, Rinehart, and Winston.

Gagné, R. M. (1984). Learning outcomes and their effects. *American psychologist*, *39*(4), 377-385. https://doi.org/10.1037/0003-066X.39.4.377

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*, 95-112. doi:10.3758/BF03197590

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. *Applied Cognitive Psychology*, *28*(2), 266-273. https://doi.org/10.1002/acp.2997

Greenberg, J., Pomerance, L., & Walsh, K. (2016). Learning About Learning: What Every New Teacher Needs to Know (Rep.). Retrieved March 23, 2016, from National Council on Teacher Quality website: http://www.nctq.org/dmsView/Learning_About_Learning_Report

Gustav, A. (1969). Retention of course material after varying intervals of time. *Psychological Reports, 25*, 727-730. https://doi.org/10.2466/pr0.1969.25.3.727

Halpern, D. F. (2013). *Thought and knowledge: An introduction to critical thinking*. Psychology Press.

Halpern, D. F. & Butler, H. (2019). Teaching Critical Thinking as if Our Future Depends on It, Because It Does. *Cambridge Handbook of Cognition and Education, eds Dunlosky J, Rawson K (Cambridge Univ Press, New York)*, 51-66.

Harden, R. M. (1999). What is a spiral curriculum? *Medical teacher*, *21*(2), 141-143. https://doi.org/10.1080/01421599979752

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30,* 138-149. https://doi.org/10.1086/374692

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Long-term retention benefits from the spacing effect in a simulated undergraduate classroom using simple and complex curriculum material. *Learning and Instruction.* https://doi.org/10.1016/j.learninstruc.2014.11.001

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237-284). Academic Press.

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhauser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science, 42*, 373-388. https://doi.org/10.1007/s11251-013-9285-2

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*, 12-26, 46. doi:10.3102/0013189X028002016

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports*, *6*, 40-41.

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: a case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, *28*(6), 684-706. https://doi.org/10.1080/20445911.2016.1181637

Manning-Oullette, A. & Black, K. M. (2017). Learning leadership: A qualitative study on the differences of student learning in online versus traditional courses in a leadership studies program. *Journal of Leadership Education, 16(*2), 59. https://doi.org/10.12806/V16/12/R4

Moss, V. D. (1995). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student (Doctoral dissertation, Utah State University, 1995). *Dissertation Abstracts International, 56,* 5204.

Mozer, M. C., Pashler, H., Cepeda, N. J., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Systems 22*. San Diego, CA: Neural Information Processing Systems Foundation.

New Literacies Research Team* & Internet Reading Research Group. (2006). Results summary report from the Survey of Internet Usage and Online Reading for School Districts 03-C and 05-C (Research Report No. 5). Storrs: University of Connecticut, New Literacies Research Lab.

Ni, A. Y. (2013). Comparing the effectiveness of classroom and online learning: teaching research methods. *Journal of Public Affairs Education,* 9(2), 199-215. https://doi.org/10.1080/15236803.2013.12001730

Ontario Ministry of Education. (2005). The Ontario curriculum grades 1-8: Language. [Program of Studies]. Retrieved January, 2016, from: http://www.edu.gov.on.ca/eng/curriculum/elementary/language18currb.pdf

Piccaino, A. G. (2002). Beyond student perceptions: Issues of interaction, presence, and performance in an online course. *Journal of Asynchronous learning networks, 6*(1), 21-40. https://doi.org/10.24059/olj.v6i1.1870

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*, 70-80. https://doi.org/10.1037/0022-0663.97.1.70

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning, 4,* 11-18.

Reder, L. M., & Anderson, J. R. (1982). Effects of spacing and embellishment on memory for the main points of a text. *Memory & Cognition, 10*, 97-102. https://doi.org/10.3758/BF03209210

Reed, H. B. (1924). Distributed practice in addition. *Journal of Educational Psychology, 15*, 248-249. https://doi.org/10.1037/h0070683

Regional Municipality of York, 2016. Census Profile. Retrieved January 2019 from: https://www.york.ca/wps/portal/yorkhome/yorkregion/yr/statisticsanddata/censusanddemographicdata/censusanddemographicdatalist

Reiken, C. J., Dotson, W. H., Carter, S. L. & Griffith, A. K. (2018). An evaluation of interteaching in an asynchronous online graduate-level behaviour analysis course. *Teaching of Psychology, 45*(3), 264-269. https://doi.org/10.1177/0098628318779275

Rohrer, D. (2009). Research commentary: The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40*, 4-17.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*, 481-498. https://doi.org/10.1007/s11251-007-9015-8

Rovee-Collier, C., Evancio, S., & Earley, L. A. (1995). The time window hypothesis: Spacing effects. *Infant Behavior and Development*, *18*(1), 69-78. https://doi.org/10.1016/0163-6383(95)90008-X

Russell, T. L., (1999). *The No Significant Difference Phenomenon: as reported in 255 research reports, summaries and papers.* North Carolina: North Carolina State University.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology, 19*, 107-122. https://doi.org/10.1002/acp.1066

Siegel, H. (1988). Educating reason: Rationality, critical thinking, and education. New York, NY: Routledge.

Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and translational gastroenterology*, *5*(1), e45. https://doi.org/10.1038/ctg.2013.13

Simmons, A. L. (2012). Distributed practice and procedural memory consolidation in musicians' skill learning. *Journal of Research in Music Education*, *59*(4), 357-368. https://doi.org/10.1177/0022429411424798

Simone, P. M., Bell, M. C., & Cepeda, N. J. (2013). Diminished but not forgotten: Effects of aging on magnitude of spacing effect benefits. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(5), 674-680. https://doi.org/10.1093/geronb/gbs096

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology, 25*, 763-767. doi:10.1002/acp.1747

Taraban, R., Rynearson, K., & Stalcup, K. A. (2001). Time as a variable in learning on the world-wide web. Behaviour Research Methods, Instruments & Computers, 33, 217–225. https://doi.org/10.3758/BF03195368

Thios, S.J. & D'Agostino, P.R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behaviour*, *15,* 529-536. doi: 10.1016/0022-5371(76)90047-5

Toplak, M. E., West, R. F., & Stanovich, K. E. (2012). Education for rational thought, In J. R. Kirby and M. J. Lawson (Eds.), Enhancing the quality of learning: Dispositions, instruction, and learning processes (pp. 51-92), NY: Cambridge University Press.

Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology, 51*, 123-138. https://doi.org/10.1016/0022-0965(91)90079-8

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Ozsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology, 22*, 685-695. https://doi.org/10.1002/acp.1388

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child development*, *83*(4), 1137-1144. https://doi.org/10.1111/j.1467-8624.2012.01781.x

Wiseheart, M., D'Souza, A. A., & Chae, J. (in preparation). Spacing effects during piano learning.

Wiseheart, M., Küpper-Tezel, C., Weston, T., Kim, A. S. N., Kapler, I. V., & Foot-Seymour, V. (2019). Enhancing the quality of student learning using distributed practice. *Cambridge Handbook of Cognition and Education, eds Dunlosky J, Rawson K (Cambridge Univ Press, New York)*, 550-584.

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*(1), 41-44. https://doi.org/10.3758/BF03329756

Zhang, S., Duke, N. K., & Jiménez, L. M. (2011). The WWWDOT approach to improving students' critical evaluation of websites. *The Reading Teacher*, *65*(2), 150-158. https://doi.org/10.1002/TRTR.01016

Appendix A

To Our Wonderful Participating Teachers,

I hope that you and your students enjoy these lessons! They have been designed for easy implementation—the majority of the teaching materials are online. Each student will need his or her own computer and a pair of headphones. You can use your own room or take your students to the library or computer lab.

Throughout these online lessons, your students will be learning to judge the credibility of websites. They will become skeptical of the websites they see and will learn to use collected evidence via the online website evaluation checklist (a.k.a. the "scavenger hunt") to explain their credibility ratings. The evidence that they collect will be used to support a final rating of 0-10 (0-4 is not credible; 5 is neutral; 6-10 is credible), and while rating accuracy will improve throughout the lessons, the most important aspect is the *process* that leads them to their final decision. By *searching* for the author, *questioning* the site's purpose, *exploring* the content and *evaluating* the site's design, your students are building skills which will make them better critical consumers of the internet.

Your job is to facilitate the delivery of these lessons within the timeframe discussed: Each lesson will require one language block (3 x approximately 90 minutes), followed-up by an additional block one-month later. The reason that we have set the timing for you, is because spacing out learning has been shown time and time again to be effective for retention. For the first time, we get to work together to see if it works with all of the unpredictability that comes with a "regular" classroom. As long as you teach the lessons in the timing that we discussed and focus on the learning objectives, you can do everything else that you would normally do otherwise. That being said, here are a few specific scientific controls that we DO need to have: Please don't reveal our hypothesis re: spacing to students and please do not formally review the lesson's categories or questions within 30 days of the last lesson—it may be impossible to avoid having students talk about it, but please don't encourage these potential discussions. These controls are necessary so that we can properly assess their memory and I appreciate your assistance in this regard. Once the lessons are done, if you need the information for your own assessment purposes, you can ask your principal to e-mail me and I will send you all of the data. Otherwise, I'll give you a separate assessment once the lessons are completely finished.

If you know of any other teachers who may want to participate, please ask them to contact me. The more, the merrier! Once all of the lessons are done, I will make the materials open access so that you and your colleagues can use them in the future.

Thank you very much for participating! Please read the lesson plans carefully and review all of the materials before running the lessons. If you have any questions, please check the FAQ on the last page and/or feel free to contact me.

Vanessa Foot-Seymour, MA, OCT

JUDGING THE CREDIBILITY OF ONLINE SOURCES: OVERVIEW

LEARNING OBJECTIVE(S)

By the end of these lessons, students will be able to effectively judge the credibility of websites. They will be skeptical of the websites they see and will be able to use collected evidence via the website evaluation checklist (a.k.a. the "scavenger hunt") to *explain* their credibility ratings. The evidence that they collect will be used to support a final rating of 0-10 (0-4 is not credible; 5 is neutral; 6-10 is credible), and while rating accuracy will improve throughout the lessons, they will learn that the most important aspect is the *process* that leads them to their final decision.

MATERIALS

Computers with internet access, headphones, markers, chart paper, teacher projector & computer

	WEBSITE(S)	TASKS
BEFORE		*Teachers, please review lesson plans and materials and ensure all participating students have parental consent.*
DAY ONE	**Sea Monkey** www.seamonkeyonline.net	Introduction (In Class) Lesson One (Online) Discussion (In Class)
DAY TWO	**Brain Science** www.brain-science.ca	Review (In Class) Lesson Two (Online) Discussion (In Class)
DAY THREE	**Bizarre Animals** www.bizarre-animals.ca	Review (In Class) Lesson Three (Online) Discussion (In Class) *Teachers, for the next month, please do not do any refresher activities.*
FINAL TEST (ONE MONTH LATER) WILL CONTACT	**Research Science** www.researchscience.net **Association of Geniuses** www.associationofgeniuses.org	Final Test (Online) Discussion (In Class)

Lesson One

Time Required	Website	Tasks	Materials
Literacy block (90-100 mins)	**Sea Monkey** www.seamonkeyonline.net	1. Introduction (In Class) 2. Lesson One (Online) 3. Discussion (In Class)	• Computers with internet access and headphones for each student • Teacher projector and computer • Chart paper and markers

1. Introduction (5 minutes)

Tell students that they'll be learning to judge the credibility of websites. These lessons have been designed for them and are entirely online! Students will log in to their computers. Write this URL on the board for students to copy into their browser *(www.credibilitylesson1.weebly.com).* You may need to explain to the students what a URL is and remind them of how to type it in without doing a Google search. The URL will contain the link to the lessons and the link to the website if they close it by mistake. You can ask students to help a buddy if they are having trouble typing in the URL. The survey auto-saves so they will never lose work if they close it or need to go back.

2. Lesson One: Online (60 minutes)

Students do the lesson on their own computer. The lesson will lead students through a diagnostic assessment (assessment *for* learning). Students will be asked not to talk during this task. This should be challenging, and students can write things like, "I don't know how to do this" if they are struggling.

Then, students continue to learn about the four categories of website design: design, authority, content and purpose, and how they can be used to identify whether a website is credible or not. They will be taught how to use the scavenger hunt to dive deeper into the website. **Students can whisper if they want to but should be encouraged to write down everything that is exciting enough to share.**

Most of this lesson should run on its' own. Please just circulate and help with technological assistance. Stick to your standard practice— do what you would normally do! As students finish the lesson, ask them to continue exploring the website. Alternatively, you can ask students to get a book or do other unfinished work (please don't give them a game or other online preferred activity to prevent students from rushing to finish).

3. Discussion

Log off of the computers and come together as a class. Say this to students: *when you first encountered this website, you were probably in a neutral position (50/50). You've never seen it before, so you probably didn't know what to think. However, as you explored the website, you saw things (maybe some red flags) that pushed you towards thinking it was credible or not credible. What were some things that you found?*

Choose some students to come to the front of the class to share collected evidence with everyone on the projector. Ask them to explain why the evidence pushed them towards thinking it was credible or not. Make a T-Chart on the board or on chart paper to keep track of the responses (one side will say "CREDIBLE", and the other side will say "NOT CREDIBLE"). Encourage them to focus on things they learned during the lesson.

Teacher Note: You don't need to keep the chart paper after. Please don't post it in the room for review. If you remember, please take a picture and text/email it to me!

Lesson Two

Time Required	Website	Tasks	Materials
Literacy block (90-100 mins)	**Brain Science** www.brain-science.ca	1. Review (In Class) 2. Lesson Two (Online) 3. Discussion (In Class)	• Computers with internet access and headphones for each student • Teacher projector and computer • Chart paper and markers

1. Introduction (15 minutes)

Ask students, *what does credibility mean?*
Answer: how **believable** and **trustworthy** something is.

Ask them how they can decide whether they can believe or trust a website? Brainstorm as a class. *The answer is:* the four categories (design, authority, content and purpose). Then, break students into groups of 4-6 and give each group a piece of chart paper. Ask them to write down the categories as headings and list as many specific questions as they can remember. The goal of this activity is to get students to struggle slightly to remember. This should strengthen their memory trace. You can go around and give clues if students are stuck! The answer key is attached. Encourage students to try to organize the questions in the appropriate categories, but since some overlap (for example, the links question was sorted under design, but it could also be a content question), it doesn't really matter.

Teacher Note: This is a brainstorming session. You don't need to keep the chart paper after. Please don't post it in the room for review.

2. Lesson Two: Online (45 minutes)

Students will log in to their computers. They will go to *www.credibilitylesson2.weebly.com* (please write this on the board), and the browser will contain everything they need for the day. You can ask students to help a buddy if they are having trouble typing in the URL. The survey auto-saves so they will never lose work if they close it or need to go back. **Students can whisper if they want to but should be encouraged to write down everything that is exciting enough to share.**

Most of this lesson should run on its' own. Please just circulate and help with technological assistance. Stick to your standard practice— do what you would normally do! As students finish the lesson, ask them to continue exploring the website. Alternatively, you can ask students to get a book or do other unfinished work (please don't give them a game or other online preferred activity to prevent students from rushing to finish).

3. Discussion

Log off of the computers and come together as a class. Say this to students: *when you first encountered this website, you were probably in a neutral position (50/50). You've never seen it before, so you probably didn't know what to think. However, as you explored the website, you saw things (maybe some red flags) that pushed you towards thinking it was credible or not credible. What were some things that you found?*

Choose some students to come to the front of the class to share collected evidence with everyone on the projector. Ask them to explain why the evidence pushed them towards thinking it was credible or not. Make a T-Chart on the board or on chart paper to keep track of the responses (one side will say "CREDIBLE", and the other side will say "NOT CREDIBLE"). Encourage them to focus on things they learned during the lesson

Lesson Three

Time Required	Website	Tasks	Materials
Literacy block (90-100 mins)	**Bizarre Animals** www.bizarre- animals.ca	1. Review (In Class) 2. Lesson Two (Online) 3. Discussion (In Class)	• Computers with internet access and headphones for each student • Teacher projector and computer • Chart paper and markers

1. Introduction (15 minutes)

Repeat introduction from Lesson Two.

2. Lesson Three: Online (45 minutes)

This lesson is the same as Lesson Two, except that students will go to *www.credibilitylesson3.weebly.com*.

3. Discussion

Repeat discussion from Lesson Two.

Final Test (Approximately One Month After Lesson 3: Will Contact with Reminder)

Time Required	Websites	Tasks	Materials
Literacy block (90-100 mins)	**Research Science** www.researchscience.net **Association of Geniuses** www.associationofgeniuses.org	1. Final Test (Online) 2. Discussion (In Class)	• Computers with internet access and headphones for each student

1. Introduction (2 minutes)

Tell students that they are going to be doing a website credibility test and evaluating two websites. Tell them to remember *as best as they can*. There is no pressure: if they don't remember something, they can leave it blank. Once they submit the final test, they may not come back to it later.

This will be where we see whether the spaced or massed group remembered more from the lessons. We expect the average recalled categories to be around 2/4, and the average specific questions to be approximately 5/17. Please do not provide any hints and treat this like a normal test.

2. Final Test: Online (60 minutes)

Students will log in to their computers. They will go to *www.credibilityfinaltest.weebly.com* (please write this on the board), and the browser will contain everything they need for the day. Students cannot chat during this final test.

Just like the previous lessons, most of this lesson should run on its' own. Please just circulate and help with technological assistance. Stick to your standard practice here—do what you would normally do! As students finish the lesson, ask them to continue exploring the websites. If they find anything new, this time they **cannot** go back. Alternatively, you can ask students to get a book or do other unfinished work (please don't give them a game or other online preferred activity to prevent students from rushing to finish).

3. Discussion: In Class

Log off of the computers and come together as a class. Ask students what they learned from the lessons. Have a group discussion about the final websites and reveal the truth about the final test websites. If there is extra time, students can share what they came across on the final two sites.

Frequently Asked Questions

What if students do not get consent to participate? Even though these lessons reflect standard teaching practice, it is still a research study, so we need parental consent for participation. If a student does not get permission to participate, please confirm whether or not they can still be physically present for the lessons. Most of the time, parents will let them stay in the room. In the case that parents do **not** let them stay in the room (this is unlikely), please make alternative arrangements for them.

What if I cannot run the lessons in the timing that we planned? Please try your best! I am trying to gather more evidence on whether lesson timing affects retention. If it does, together we might be able to make recommendations for professional development courses on the spacing effect. That being said, things happen!! If you are interrupted at any point (fire drill, snow day, internet goes down, etc.), follow up with the rest of the lesson ASAP and send me an e-mail to let me know. Try to keep the timing as close to what we planned as possible.

I have some kids who leave for language. Should they still go? You (and the SERT teacher) can make that decision. These lessons so that they can be used across a wide range of ages and capabilities. If your students stay, please let me know if there is any reason why I should expect different results from them compared to the rest of your class.

You are asking me to teach critical thinking. What is it? Our definition of critical thinking is simple. We are asking students to use reasonable, reflective doubt to decide what to believe or what to do. To be a good critical thinker, you have to be willing to deal with being unsure. You may notice that students have an issue with the uncertainties with the websites, but please keep encouraging them to just do their best at making a decision. The goal is not in deciding how credible they are—we care about the **quality** of their justification. The websites are purposefully ambiguous so that each student can use different reasoning to explain their perspective.

Can students use their iPhones or iPads? No. The screen is too small, and students need a keypad for the online survey. Please avoid using these devices.

What if a website is blocked, or if the technology doesn't work? I have tested all the URLs on the YRDSB wifi, but in the case that something happens, there is an "emergency kit" folder on the Google Drive. This has the PowerPoint, the videos, and offline checklists that you will be able to use. However, since this is a media study, if the students can't get on the websites (if the Internet is down), you won't be able to continue. Please handle the situation however you would normally (i.e., giving some DPA while you wait for the internet to come back up, moving around the daily schedule if possible).

Can I use any of this for my assessment? If you want the assessment that I collected for the research, please ask your principal to contact me and I'll release all of the data to them. However, once your class has completed all of the lessons, I will be sending a separate assessment for you to use.

Why are you talking about credibility and not having students identify real/fake sites? We have found that students want to use the terms "real/fake" because they're easier to conceptualize (it's very black and white!). However, we know in reality that judging website credibility is not so easy. There is a spectrum— website creators love to give opinions online, but that doesn't necessarily mean that the site itself is fake. At the end of the day, it's about how much we trust and believe what we encounter online. *Do you think that the travel blogger is an expert, and will you follow his/her to-do list when you go to Paris? What about the Pinterest recipe for angel food cake… do you think that it will work out?* We would rather have students be more skeptical than less skeptical. In a perfect world, students should target the red flags and assume everything is fake until being convinced otherwise. That's what these websites will do. Remember that whatever students decide, it's the process of decision making that matters.

Answer Key (Categories and Questions)

Design

Do the photos and colour choices look professional?

Is the website nicely organized and easy to navigate?

Are there any obvious spelling errors or typos?

Is the layout consistent from page to page?

Authority

Is the author/creator of the website clearly identified?

Is the author of the website an expert in their field?

Is there a way to contact the author by phone, mail or e-mail?

Content

Does the website say when it was created?

Does the website say when it was last updated?

Can you confirm that the information is correct by doing a Google search?

Are the links relevant to the subject? In other words, do the links take you somewhere that makes sense if you click on them?

Purpose

Is the website trying to educate you with real information?

Is the author trying to sell you something?

Do you think the author has intentionally left out any important information that could help you decide if it's credible or not?

Appendix B

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of W1_4cat is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.365	Retain the null hypothesis.
2	The distribution of W1_14q is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.583	Retain the null hypothesis.
3	The distribution of W2_4cat is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.494	Retain the null hypothesis.
4	The distribution of W2_14q is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.702	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.