Judging Credibility:

Can Spaced Lessons Help Students Think More Critically Online?

Vanessa Lauren Foot

A THESIS SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTERS OF ARTS

August 2016

© Vanessa Foot, 2016

Abstract

Despite its prevalence in the psychological literature, the spacing effect has not yet been fully explored in real-world classroom settings using curriculum-based material. The current study investigated whether laboratory effects of spacing can also be seen in the classroom, and if the spacing effect is still robust when extending from fact learning to critical thinking. Students were taught direct instruction in critical thinking where they judged the credibility of online sources as part of either a three-day consecutive or one per week set of lessons. Thirty-five days after the final lesson, students were tested in order to see how much of the material they retained and could apply to evaluating a new website. Results demonstrated that there were significant effects of spacing on the final test after 35 days. Students in the spacing condition were better able to explain their website ratings and remembered more of the facts from the lessons than students in the massed group. However, the website ratings did not differ significantly between the two groups at final test.

Keywords: spacing effect, classroom, critical thinking, credibility, higher-level thinking

Acknowledgements

Thank you to everyone involved in this project. First of all, thank you to my team at York. My supervisor, Melody Wiseheart, and my lab mates and mentors: Tina Weston, Justeena Zaki-Azat , Annalise D'Sousa, and Katie Matthews. Thank you to my lead volunteer Stephanie Kaczer, who was my partner in crime for a majority of the lessons, and to my other volunteers: Zita Lau, Michael Seymour, Julia Martini, Lysianne Buie, Zunaira Amin, Anthony Fallico, and Shira Springer, many of whom spent time in the classroom and marked countless student papers.

Thank you to the principals who welcomed us into their schools. These individuals took the time to get to know us and what we were doing, and some even took the time to help us recruit additional teachers: Earl Liverance, Tim McFadden, Tim Dunn, Steve Young, Brian Donnelly, and James Flynn. Also, thank you to the 22 teachers in 20 classrooms (one left for maternity leave, and the other was on an occasional contract) and 558 students without whom this study would not have been possible. We built relationships with these teachers and their students and were sad when the time with each class came to an end.

Last but certainly not least, thank you to my incredible friends and family for all of their love and support—particularly my mom, June Foot, who is a retired teacher. She was able to use her expertise to help to plan the timing and content of each lesson, and as such she spent countless hours familiarizing herself with the spacing literature. She was also able to use her teaching connections to help with recruitment (more than half of the principals who took part in the study were ex-colleagues of hers). In addition, when I couldn't attend a testing session she went in my absence, and helped to mark papers after the final test. My mom is the reason that I became a teacher and I am incredibly thankful that we could share these new experiences together.

Table of Contents

List of Figures

List of Tables

Introduction

The Spacing Effect

The spacing effect is a memory advantage that occurs when information is reviewed or re-learned after a temporal gap, as opposed to a more condensed time frame (or in a massed fashion). Distribution of learning episodes is often seen in the real world when students begin to study several days leading up to a test or exam, instead of chunking the information and reviewing it over the entire term. According to spacing research, individuals who have spaced out their learning over a longer period of time are better able to retain material than those who cram it into a shorter time frame, given equal amounts of study time for each learning session (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). In traditional spacing studies, to-be-learned information is studied in two study sessions separated by a manipulated period of time known as the inter-study interval (ISI). After another period of time known as the retention interval (RI), a final test is administered to assess level of retention (*Figure 1*).

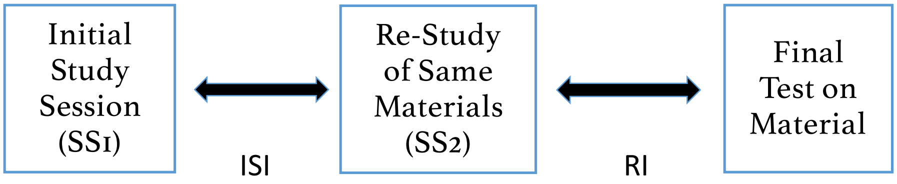


Figure 1. Visual representation of spacing design

There are several theories that attempt to explain why spacing works the way that it does. Glenberg (1979; see also Estes, 1955) believed that each item is stored in memory along with the specific context that it was learned in, and that context changes over time. Glenberg's encoding variability theory supposes that the greater the number of unique contexts that are associated with each item, the greater number of memory traces that can be drawn upon in order for the information to be retrieved.

Alternatively, the study-phase retrieval theory (Thios and D'Agostino, 1976) suggests that learning of an item will be greater if the first presentation can be retrieved from memory and updated. 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, because their relearning is spaced out, it will be more effortful to reconstruct the memory trace, and thus the memory trace will become stronger. Recent theories of the spacing effect combine encoding (i.e., context) variability and study-phase retrieval accounts (Delaney, Verkoeijen, & Spirgel, 2010).

Spacing Effect in the Classroom

According to Bahrick and Hall (2005), the spacing effect is one of the oldest and best-documented phenomena in the history of learning and memory research. However, it is important to highlight the difference between the hundreds of spacing studies that have been run in the laboratory involving factual and verbal material and those that involve higher-order thinking processes and ask subjects to think critically. The spacing effect literature has provided a solid base for fact learning studies (for a review of 839 effect sizes, see Cepeda et al., 2006), but there is little evidence of its application in real classroom settings (Dempster, 1988). In the classroom, students are often asked to memorize facts, which can become a common practice, leaving less time for higher levels of thinking. These facts are not always applied to real-world scenarios that ask students to think for themselves in problem solving situations. As such, when they are asked to explain, evaluate, analyze, and consider alternative perspectives— they are not always confident in doing so. It is not obvious whether or not the same memory effects seen within the fact and verbal literature will be seen in the classroom using higher-order thinking, given the relative scarcity of critical thinking and memory research. For the remainder of this paper, the terms "higher-order thinking" and "critical thinking" will be used interchangeably.

Since spacing could potentially improve classroom learning, it is important for memory researchers to investigate whether spacing is robust when applied in more ecologically valid settings with real curriculum-based material. There are several spacing studies that have done this. Smith and Rothkopf (1984) taught university students video lecture material from a statistics course. They looked to see whether the distribution of lessons over four days was more effective than a single day presentation. Overall, no spacing effect was found on the problem solving part of the testing measure, but a significant effect was found on the fact related questions. Yazdani and Zebrowski (2006) investigated whether the scheduling of plane geometry homework (defined as either massed daily drilling after each covered topic or spaced homework over an extended period of time) would result in a higher level of achievement. The study found significant findings that supported a spaced instructional design, but all seven study sessions had unequally spaced intervals, which meant the time in between the lessons was not constant. There was also an unequal amount of review at each lesson, as amount of relearning was also allowed to vary. Bird (2010) conducted a study where he tested the ability of English language learners to detect and correct verb morphology. Although this study did require complex thinking and students seemed to benefit significantly from a spaced schedule, this is the only study that used five study sessions, making it unique and difficult to compare to the rest of the literature.

Another classroom study was conducted by Sobel, Cepeda, and Kapler (2011), who taught 39 uncommon English dictionary words to fifth-graders. Teachers used traditional methods (slides, oral practice, paper and pencil tests), teaching students the content again either immediately, or one week later. Five weeks after the second tutorial, students were given a final vocabulary test in order to test their retention. Students who participated in the spaced condition accurately recalled more definitions (20.8%) than students who participated in the massed condition (7.5%).

The most recent study on the spacing effect in the classroom was conducted by Kapler, Weston, and Wiseheart (2015), where researchers invited participants to attend a simulated university lecture. In these lectures, students learned both factual and higher order material and reviewed it in either a massed or spaced fashion. Results demonstrated that reviewing the material with more time in between the review sessions led to better long-term retention for both facts and higher-order thinking tasks.

The study by Kapler et al. (2015) has been the nearest to identify spacing effects using critical thinking material in the classroom because it used higher-order thinking curriculum materials, relevant inter-study intervals (1 day, 8 days), a long retention interval (35 days), and time-efficient review methodology. However, the primary limitation of this study was that it took place in a mock classroom, and as such there was more experimental control than would be expected in the real-world.

The results of classroom spacing research suggest some benefits when using the spacing effect, but there is not enough evidence to make recommendations for its widespread usage. Perhaps the reason for limited spacing studies in the classroom is because of the number of extraneous variables that researchers would undoubtedly face. Each classroom, even if students are the same age, can differ depending on the school, teacher, social dynamics of the classroom, socio-economic status, distractions (such as fire drills and lockdown drills), time constraints, and students' previous knowledge of the subject due to additional learning outside of class time.

There are any number of circumstances that could arise and cause issues—for example, in the current study, snow days caused scheduling shifts for three of the sessions, school assemblies and holidays had to be carefully accounted for, and some homeroom teachers were absent on certain days, which led to the students losing focus.

Definition of Critical Thinking

Teaching critical thinking is a lofty goal for educators, since it can be a challenge to break the concept down in a way that makes it easily accessible for teachers to use in the classroom. Bloom's taxonomy of educational objectives (Bloom, 1956) is a commonly used resource that attempts to break down the components of critical thinking in a way that is both simple and accessible. Bloom's taxonomy is a hierarchical framework that is intended 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 when developing critical thinking measures because there is little evidence that thinking is hierarchical in nature—Ennis (1996) suggests that learning facts and complex thinking are completely distinct processes.

A qualitative study by Descours (2013) found that teachers in Canada have varied definitions of critical thinking. Teachers who were surveyed agreed that thinking critically 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.

Though there are many definitions of critical thinking, it is a concept that dates back to the work of philosopher John Dewey in the early 20th century. In his book, *How We Think*, he presented the original notion of what it means to think critically—he believed that questioning the world was one of the most fundamental aspects of being an intelligent human being (Dewey, 1909). A good critical thinker, according to Dewey, carefully listens to the beliefs and opinions of others and takes time to thoroughly investigate all possible aspects before deciding whether or not they agree with their perspective. The necessity for good critical thinkers has not changed since then, and many scholars have discussed the importance of teaching it in schools.

There seems to be a disconnect when it comes to teaching critical thinking, caused in part by how difficult it is to obtain a clear working definition that can be applied to practical learning situations. There have been theorists in education, psychology, and philosophy who have tried to break down critical thinking into smaller components in order to provide a solution to this, but there is a wide range of contrasting theories and opinions on the matter (Brodin, 2007; Ennis, 1987; Glaser, 1941; Halpern, 2003; Kuhn, 1999; Lewis, 1929; McPeck, 1981; Paul, 1985; Siegel, 1988). Although these theorists are considered experts in the field, their theories have some fundamental differences. For example, Ennis (1987) suggests that critical thinking is a general skill that involves “reasonable, reflective thinking that is focused on what to believe or do” (p. 10). Critical thinking, according to Ennis, contains a set of skills and dispositions and should be taught explicitly and then infused in order to work it into everyday life and create an implicit understanding. On the other hand, McPeck (1981) defines critical thinking as a “reflective skepticism that is linked with specific areas of expertise and knowledge.” In contrast with Ennis, McPeck claims that critical thinking is always tied to a certain discipline and cannot be dissociated from its context. Other descriptions of critical thinking have been provided, with some theorists describing it as an attitude, a logical process, a purposeful reflection, a developmental process, or even synonymous with intelligence (McPeck, 1990; Niu, Behar-Horenstein & Garvan, 2013).

Due to the discrepancies between theorists' descriptions of critical thinking, the American Psychological Association (Falcione, 1990) brought together 46 leading scholars in order to formulate a consensus from experts and an agreed upon cohesive definition of critical thinking. The report identifies the most pertinent skills and dispositions involved in critical thinking. Their agreed upon definition is the following:

> 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.

Although this definition has been critiqued for being too broad (Alston, 2001), it provides somewhat of an explanation as to what critical thinking truly entails: purposeful, self-regulatory judgment.

Teaching Media Literacy: Judging the Credibility of Online Sources

If using spacing in the classroom is able to make students better critical thinkers, it should work in conjunction with any subject. In other words, students would not take a critical thinking course—the critical thinking content would ideally be woven into their standard curriculum. Abrami et al. (2008) conducted a meta-analysis on the effects of instructional interventions on students' critical thinking skills. They looked at 177 studies, and found that instruction of critical thinking was most effective when students were taught critical thinking instruction and subject content in approximately equal parts. This finding promotes the idea that teachers should be teaching critical thinking skills this way so that students are able to put them into context and use them before transferring them to other disciplines. In addition, students should be given practical and relevant examples of when they might use their developing critical thinking skills. In line with this approach are the findings from Falcione (1990), Halpern (1998), and Paul (1992).

In order to meet the goal of giving students a content specific and practical way to use their critical thinking skills, it was decided that students would judge the credibility of websites. Judging website credibility is a relevant topic to K-12 education, because the nature of education has changed since the advent of widespread Internet usage. Paul (1992) suggested that the world is in ever-accelerating change; information is multiplying as it swiftly becomes obsolete and out of date. In the past decade, technology has advanced to become an even more prominent figure in students' lives. A Pearson (2014) study polled over 2,300 students in the United States aged 8-18, and determined that only one percent of students did not use any digital technology for school purposes. The remaining 99 percent used desktop computers, laptops, netbooks, tablets, smart phones, and e-readers. Out of these students, 70% used their devices for conducting research on the Internet. This poll demonstrates that in this modern age of technology, the majority of students have virtually unlimited access to information on the Internet both at school and at home. As convenient as this access may seem, the ability to locate information can be problematic if students are not able to make informed decisions about whether that information is trustworthy or not. As such, there is a constant and growing need for students to obtain a more critical eye towards website content instead of simply accepting the thoughts and opinions that they are exposed to.

The Ministry of Education in Ontario (2005) discusses the importance of training students in critical literacy, and advocates that the "impact and influence of mass media and popular culture and the messages they convey, both overt and implied, can have a significant impact on students' lives. For this reason, critical thinking assumes a special significance" (p. 13).

Teaching Using the Spacing Effect

The school year in Ontario starts at the beginning of September and extends until the end of June. Every certified teacher is asked to follow the Ontario Curriculum (Ministry of Education, 2005), where topics are first divided by subject and then sectioned further to cover overall and specific expectations of their students. Recommendations for how learning should occur are noted in the curriculum, but teachers are generally given the power to decide how and when to implement each of the expectations. Teachers have flexibility within these parameters because each school comes with its own set of challenges, and combinations of learners can be vastly different from one classroom to the next. As long as expectations are met, the Ministry of Education does not tell teachers how to teach. However, by not pressing for certain standards of practice, students may be missing out on learning through some of the most important evidence-based teaching strategies, backed by psychological principles. The National Council on Teacher Quality (NCTQ, 2016) published a list of the six core strategies that every teacher should be implementing in their classroom:

1. Pairing graphics with words
2. Linking abstract concepts with concrete representations
3. Posing probing questions

4. Repeatedly alternating solved and unsolved problems (interleaving problem types; interleaving is a form of spacing)
5. Distributing practice (i.e., spacing)
6. Assessing to boost retention

To be fair, many teachers use these strategies on a regular basis. The curriculum requires that teachers use a variety of learning and assessment tools in order to provide their students with constant, ongoing feedback (Ministry of Education, 2005). Posing probing questions are also highly encouraged in the classroom and are the most commonly discussed in teacher training textbooks (NCTQ, 2016).

Although some of these practices are seen in teacher training materials, there is no consistency among them that encourages distributed practice, also known as the spacing effect (NCTQ, 2016). As previously discussed, the spacing effect occurs when information is taught to students over an extended period of time as opposed to a more condensed time frame. One of the major advantages of spacing is the fact that the optimal gap between learning sessions depends on how long an individual needs to remember the information (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). This means that there may be a mathematical model that can optimize distribution of learning, such as the Multiscale Context Model that was developed for this purpose (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009). If this mathematical formula can work in the classroom, using spacing could have incredible potential for improving student retention. In the present education system, students are asked to remember specific content from the time of initial learning until tests and assignments are completed, and then hold that information in memory until their end of unit, cumulative tests. However, educators will need to decide what the goal is in order to determine how they should space their lessons. Students who want to remember information over longer periods of time may need to make short-term sacrifices in order to reach their long-term goals. Perhaps the ideal situation would be that students retain the information over the summer and leading into their next school year. This is because student success is dependent on whether he or she is able to remember and build on a sturdy, foundational knowledge base in all subjects.

Current Study

This study explores whether spacing effects are robust enough to transfer to real-world classrooms. Students aged 9- to 12-years-old were explicitly instructed in critical thinking skills, in order to judge the credibility of online sources. The word judgment in and of itself means that students are making inferences by evaluating possibilities in terms of specific goals and evidence (Baron, 2000). The workshop took three blocks of time that were scheduled on either three consecutive days of the week (massed), or three weeks in a row (spaced, one day per week). In each condition, students explored several different websites, were guided on how to make decisions about credibility, and were taught how to locate evidence from the websites that helped them defend their decisions.

We investigated whether the spacing effect will help students think more critically while evaluating online sources. The definition of a sound credibility judgment, since the Ontario Curriculum does not directly specify what it is, came from Ennis (1987), who defined it as the ability to think critically and make a decision about a source by asking questions about points of view, conflicts of interest, scientific information, methodologies, and assumptions. This definition closely represents the ideals of both Dewey (1909) and the APA's Delphi consensus panel (Facione, 1990). Furthermore, Ennis' definition speaks to how the education community imagines critical thinking as an educational goal. Ennis provides a detailed list of pertinent dispositions and abilities, which have been implemented in validated educational assessment instruments (Ennis & Milman, 2005a, 2005b; Ennis, Milman, & Tomko, 2005; Ennis & Weir, 1985).

Hypotheses

We predict that, at final test:

1. Students in the spaced condition will use more information from the lessons to explain their website ratings than students in the massed condition.

   a. Students in the spaced condition will use more of the *four categories* when explaining their rating than the massed condition.

   b. Students in the spaced condition will use more of the *17 questions* when explaining their rating than the massed condition.

2. Students in the spaced condition will have closer ratings to teacher consensus than those in the massed condition.

3. Students in the spaced condition will remember more information from the lessons.

   a. Students in the spaced condition will recall more of the *four categories* than students in the massed condition.

   b. Students in the spaced condition will recognize more of the *17 questions* than students in the massed condition.

Method

Participants

For this study, students aged 9- to 12-years-old (grades 4-6) were recruited from the York Region District School Board. This population was chosen because the Ontario Curriculum (Ministry of Education, 2005) requires that at this point in their education, students must begin to “differentiate between fact and opinion; evaluate the credibility of sources, and recognize bias”

(p. 89), but they have not yet had enough exposure to become proficient. Students were recruited from six schools across York Region. Participating schools were chosen based on principal and teacher interest.

A total of 558 students within 20 classrooms participated in the lessons. Of the 558 students, four students did not receive parental consent (resulting in a 99.3% consent rate). Three of these students were given verbal permission to participate in the lessons, but their work was not used for research purposes. One student who did not obtain consent was withdrawn from the lessons and placed in another classroom for the morning, per parent request. A total of 170 students were excluded from data analysis due to lack of consent, missing a lesson, or being identified on an individualized program that might affect results (e.g., one student was 90% blind and required scribing; others took part in the lessons and opted to type their responses on the computer). After attrition, the final sample consisted of 388 students (178 spaced, 210 massed).

The study was approved by the York University Human Participants Review Committee. At the beginning of the lessons, students were given ID numbers and reminded that their participation was confidential. No names were written on any of the testing materials. Although lessons aligned with the Ontario curriculum, tests given during the study were not used as part of student grades, to avoid biasing course marks. Students were made aware that their grades would not be used for their report cards, which may have also influenced their desire to perform well. However, all 22 teachers reiterated that students were responsible for knowing these concepts, and informed them that they could be tested on them later as part of their coursework.

Design

A between-subjects design was used, where classes were randomly assigned to one level of the independent variable (spaced or massed), stratified as evenly as possible to ensure that classrooms in each condition were equal in ability. For example, if there were teaching partners in one school (often referred to as “team teachers” because students are in the same grade, so materials are shared), these classrooms were separated and assigned to participate in each of the two conditions. Efforts were made to ensure that there was a mixture of grades in each condition. Results of student individualized education plans (I.E.P.) were shared by the teachers when necessary in order to see if a learning disability would hinder student performance. These students were excluded from the research if teachers believed that they would be at a disadvantage in the study. However, accommodations were made in the classroom (extra help from volunteers, or online access if students needed to type instead of write by hand) so that they could still participate in lessons. At every stage of the lessons, students were taught as fairly and equitably as possible, regardless of condition.

Students in each condition (massed and spaced) were given an identical set of lessons. The massed condition received three days of training in a row, and spaced condition received their workshop with one lesson per week. The classes were booked on each day of the week (Monday-Friday, depending on teacher and researcher availability), with day of week balanced across conditions to prevent a confound from day of week effects.

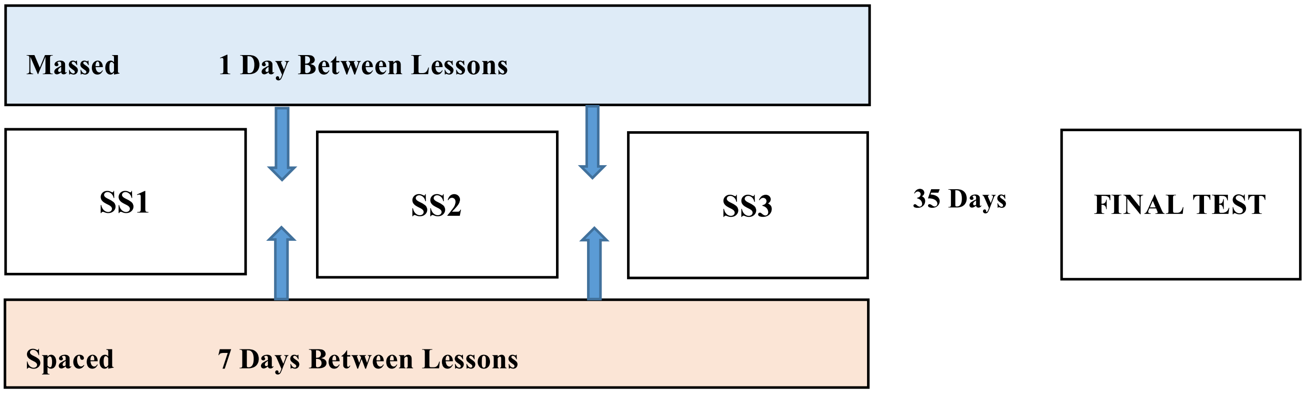


Figure 2. Visual representation of lesson design.

The Website Evaluation Checklist: Four Categories and 17 Specific Questions

When designing a lesson plan, it is often suggested that teachers follow a backward planning design, meaning that all content should be taught with the final assessment in mind. As a result, every part of designing lessons for this study was structured in a way that stimulated learning leading up to the final test. At the end of the credibility workshop, students needed to know the following pieces of information:

1) Determining website credibility is not as easy as identifying whether sites are real or fake (which would result in an extreme rating of 0% or 100% for credibility). Instead, the evidence on the website helps students to make an educated decision that will most likely fall somewhere in between 0 and 100.
2) Each of the four categories contains specific questions (17 overall), and students can look at the site and use their answers to the questions as evidence to explain their rating.
3) This collected evidence can be sorted into four main categories: design (how does the website look?), authority (who wrote the website?), content (do you trust the information on the pages?), and purpose (why was the website created?).

Based on these learning goals, a checklist was designed to help students learn how to generate a comprehensive rating of the website.

Our checklist which included the four categories and 17 questions *(Appendix A*) was based off of a checklist by Bronstein (2007), which she created with the assistance of a Delphi panel of experts. Bronstein explored the validity and reliability of the checklist as a pedagogical tool, and she established the four main categories of website evaluation: design, authority, purpose, and content. The checklist was created in a way that would lead students to their final decision, guiding them along the way. She recommended that checklists used for critical thinking purposes should involve continuous scales rather than only yes/no answers, because critical evaluation is an ambiguous process that involves many different options for premises and different forms of reasoning that are equally legitimate. The goal is to gain deeper insight into students' thought processes.

Existing checklists usually differ in three main aspects: the categories used (which Bronstein refers to as criteria), the response method, and the final assessment. For a summary of these checklists, refer to *Table 1*.

The reason that Bronstein's (2007) website evaluation checklist was not used directly for this study is because her questions were designed for older (high-school) students. We adapted our questions within each category to suit a younger audience. Given the age of the students in this study, yes/no options were used in conjunction with a continuous rating scale and narrative (short answer section) with the rationalization that students would use a combination of these tools in order to formulate an opinion. The checklist was not assessed or used in data analyses.

Title/Author	Criteria for Evaluation (Categories)	Response Method	Method of Final Assessment
Critical Evaluation of a Web Site Shrock (2003)	Technical and visual aspects, content, authority (10-19 questions per section)	Checklist (Yes/No) Narrative response	Narrative asks students to look at their responses on the checklist and explain themselves.
Website Evaluation 2Learn.ca (2004)	12 questions are followed by a fill-in box requiring a narrative response	Typed narrative response	Final Question: Can I use this Website?
Web Page Evaluation Worksheet Everhart (1996)	Currency, content/information, authority, navigation, experience, multimedia, treatment, access, misc. (3-8 questions per category)	Each category has a maximum point value. Students assign zero to the maximum number of points based on questions	Scores are totaled and 90-100 is excellent, 80-89 is good, 70-79 is average, 60-69 is borderline acceptable, and below 60 is unacceptable.
Guide to Effective Instruction Ministry of Education (2008)	Authority, audience, design	Narrative response	Rubric that highlights Ministry expectations
Bronstein (2007)	Authority, purpose, content, design	Narrative	Look at question, write-in decision (look/decide/evaluate) Final decision out of four (excellent/acceptable/questionable/ unacceptable
Zhang & Duke (2011)	Authority, accuracy, design, currency, usefulness (intended purpose)	Narrative Checklist (not available online)	Narrative and ranking scale

Table 1: Comparison of Evaluation Checklists (adapted from Bronstein, 2007)

Lessons

Each lesson was designed and taught by an Ontario certified teacher (Vanessa Foot), who also designed and carried out all aspects of the lessons. The same lessons were used in each condition, where students were taught a three-day workshop on judging the credibility of online sources. The lessons were designed to include a combination of group work, partner work, and individual tasks in order to keep the students stimulated and engaged.

During each lesson, students explored a new website and were asked to make a decision as to how credible it was, via a rating scale. The websites were all found online and were chosen due to their ambiguous nature. Two of the websites (SS1, SS2) were hoax websites, and two of the websites had real content, but each website had strength in at least one category and weakness in another. Students were not asked to identify whether the website was simply real or fake, to avoid binary thinking.

Lesson one consisted of an introduction, which included the pre-test (SS1) and initial learning, and lessons two (SS2) and three (SS3) allowed students to practice their evaluation skills using the credibility checklist. No time limits were imposed, but all students completed the task within the allotted time frame (an average of 1.5 hours, depending on school timetables).

After the last study session was completed, the students were assessed after 35 days using the same worksheet that they practiced during the initial study session, but with a new website. Thirty-five days was chosen as the retention interval for two reasons: because that is typically the longest that any particular unit will be where students have to retain the information taught and apply it to a cumulative test, and because it is known that for a 35-day retention interval, 7 days is likely to be the optimal inter-study interval (Cepeda et al., 2008).

Each lesson and assignment followed the standards suggested in the Ontario Curriculum (Ministry of Education, 2005), and lessons were designed so that they fit standard educational practice and could easily be replicated by any teacher. The teacher's presence in the classroom was not intrusive because the research supplemented curriculum-based learning that students were already responsible for.

Lesson 1 (SS1), Introduction to Credibility: *Dog Island (www.thedogisland.com).* Lesson one consisted of a brief introduction to the topic of credibility. Students were taught the definition of credibility – "*how trustworthy and believable something is*" – and brainstormed a few examples as a class of how the topic of credibility connects to their everyday lives. All students determined that credibility was important for doing research for school.

After the full group introduction, students were given the pre-test (*Appendix C*) that they had to complete individually, and they were asked to make a decision between 0 and 100 about how trustworthy a website was. They were asked to justify their answer in narrative form (either point form or full sentences) using evidence from the website.

After the pre-test was completed, students had an anonymous vote on their decision (students were asked to close their eyes and raise their hand if they voted 0-25, 26-50, 51-75, 76-100). Students were then shown how the credibility scale worked, which was described as follows: *"when you first open a website, it is important to be objective [a definition was provided]. You will always start at a neutral spot, which is at 50/100 on the credibility scale. As you browse through the website, you need to look at the four categories and ask yourself the 17 specific questions. Each answer to these questions can be used as evidence. Each piece of evidence that you find will either move you up the scale and take you closer to 100 ("green flag"), meaning that you are closer to believing and trusting the website, or down the scale and closer to zero ("red flag") meaning that you are not believing and trusting the website. Some pieces of evidence may move you more than others."*

Students were then asked to find a partner, where one kept the website open and the other signed onto an online quiz. This activity was intended to give students initial exposure to the website checklist in a fun and interactive way. In pairs, students answered the checklist questions while looking at the website. The site that held the online quiz was created by the researchers, and it introduced students to the four categories and 17 questions within these categories (*Appendix B*). It was suggested that students look at these guiding questions with the rating scale, in order to make an educated guess about credibility as opposed to guessing like they had done at the beginning of the lesson. After the class discussion, students voted anonymously again to see whether their ratings about the website had changed after using the guiding checklist with their partner. It seemed that many students who originally rated the website an extreme of 0 brought their ratings up after learning to use the rating scale. Before the end of the lesson, students were asked to share some of the evidence they found with their partners about the website.

Lessons 2 (SS2) and 3 (SS3): *Tree Octopus (www.zapatopi.net/treeoctopus*), and *Sci-News (www.trexnews.weebly.com).* Lessons two and three consisted of relearning the four categories and 17 questions that students could use in order to make an educated decision about website credibility. Students were assigned to small groups where they attempted to recall the four categories and the 17 questions. Several younger groups struggled to recall the categories and questions, so in order for them to relearn successfully, a full classroom discussion was held and the questions and categories were written on the board instead.

After relearning, students were introduced to a new website, and they followed the checklist (*Appendices A, A1* and *A2*). Each of the four categories was rated separately first and then students evaluated the website as a whole and were asked to explain their answer in point form using evidence from the website. The short answer section prompted students to explain their answer using at least one piece of evidence from each category (design, authority, content, purpose).

Each lesson ended with a discussion where students were given an opportunity to come together as a group and share some of the “red flags” or “green flags” that moved them up or down the rating scale. Most students were given an opportunity to come to the front of the class and show a piece of evidence that they used in making their final decision.

Final Test: *Mike the Headless Chicken (www.miketheheadlesschicken.org, www.finaltestwebsite.weebly.com*). After the 35-day retention interval had passed, a researcher visited each class to administer the final test (*Appendix D*), which followed the same paragraph structure as in each lesson, and also the final questionnaire (*Appendix E*), which was used to collect basic memory data. The final test asked students to recall the four categories of website design, followed by recognition of the 17 categories, where students had to determine which 17 of 34 questions were used in the website credibility workshop. The majority of the distractor questions that were used came from Shrock's (2003) Critical Evaluation of a Website checklist. Many of these questions had not been used during the lessons because they touched on technical aspects of accessing a site (how fast the website loads, ads, etc.), which we decided not to use due to lack of control.

Teacher Baseline Ratings

Credibility ratings out of 100 were taken for each website and the final test. Student ratings were scored according to their relationship to the teacher consensus rating. Consensus scores were taken via a Qualtrics survey to Ontario certified teachers. All participating teachers were briefly taught the intervention either online or in person, and were asked to give ratings as part of a 30-minute online survey. Ratings were adjusted by subtracting the teacher score from student scores, in an effort to characterize the discrepancy between student ratings and a teacher/normative rating. Negative scores indicate a score that is lower than teacher consensus, and a positive score indicates a score that is higher than teacher consensus. A rating of 40 was considered less correct than a rating of 75 when the teacher consensus rating was 70 (difference score of -30 vs. +5, with an ideal difference scores of zero). *Figure 3* shows a histogram of teacher ratings for each of the four websites: SS1, SS2, SS3 and the final test.

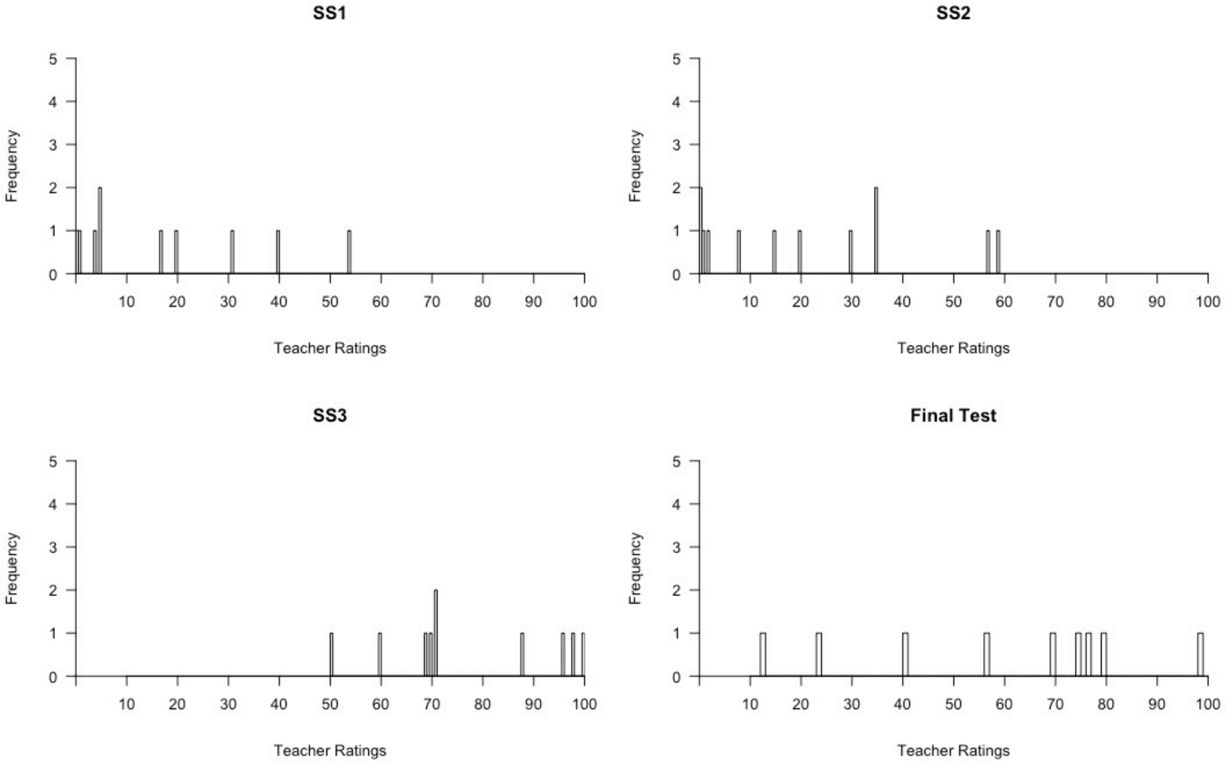


Figure 3. Histogram of teacher ratings for the four websites.

Results and Discussion

Primary Analyses

Sixteen separate t-tests were conducted, looking at each dependent variable against the spacing effect. Before running the analyses, tests were conducted to ensure that the assumptions underlying t-tests were satisfied. In order to test for homogeneity of variance, Levene's test for equality of variance was conducted for each test under the requirement of $p > .05$. The results are summarized in *Table 2*. Owing to this violated assumption, these tests were conducted assuming that the equality of variance assumption had not been met, and the degrees of freedom were adjusted. In order to test for the normality assumption, the values of skewness and kurtosis were investigated, and the Shapiro-Wilk test was conducted. After running these tests, there was no evidence of normality (*p* = .001) in any of the samples when tested at *p* > .05. In spite of this, the decision was made not to alter scores prior to running this data set due to the fact that populations in this study were not expected to be normally distributed. Ratings were often extreme in the samples due to bias. For example, websites that were deemed fake were often given low extreme scores and websites that were deemed real were given high extreme scores. Even though students were encouraged to avoid binary thinking by using the rating scale, many students still gave extreme ratings and variance was large in each group. Given the large sample size in this study and the robustness of the independent samples t-test, we continued with the analyses as planned. However, we also ran a non-parametric test on the ranks (Mann-Whitney U) in order to be sure of the accuracy of our results despite the assumption violations. These results can be found in Appendix F. In addition, because of the extreme ratings that were often given, chi-square tests were conducted as secondary post-hoc analyses, to look at trends that did not include extreme scores of 0, 100, or undecided ratings of 50. Lastly, the independence of observations assumption has also been violated in the current sample, since there was nesting by classroom. This was addressed by running a separate analysis with students constituting repeated measures, looking at classrooms as independent data points.

	Massed			**Spaced**			
	n	***M***	***SD***	***n***	***M***	***SD***	***p***
SS1	210	34.7	29.8	177	30.7	28	.175
Pre (17.7)**		7.0			13.0		
SS2							
Design	210	48.7	31.1	177	44.2	27.7	.137
Design (49.4)**		-0.7			-5.2		
Authority	208	30.9	27.0	176	23.5	22	.003*
Authority (21)**		9.9			2.5		
Content	209	36.1	30.7	176	28.0	25.4	.005*
Content (29.5)**		6.6			-1.5		
Purpose	209	36.3	30.9	173	31.1	28.4	.089
Purpose (37.5)**		-1.2			-6.4		
Overall	209	31.7	29.2	176	28.7	26.3	.299
Overall (21.8)**		9.9			6.9		
SS3							
Design	210	60.2	27.2	177	58.3	25	.471
Design (69.5)**		-9.3			-11.2		
Authority	210	56.5	32.7	177	52.1	30.6	.176
Authority (70.8)**		-14.3			-18.7		
Content	209	59.5	29.5	177	50.4	28.0	.002*
Content (80.1)**		-20.6			-29.7		
Purpose	209	59.7	30.5	175	53.5	28.4	.039*
Purpose (66.5)**		-6.8			-13.1		
Overall	210	68.6	27.3	177	58.6	27.5	.001*
Overall (77.3)**		-8.7			-18.7		
Final Test	210	55.5	30.6	177	56.5	30.8	.748
Final Test (59.4)**		-3.9			-2.9		
Free Recall Four Categories	210	0.7	.86	177	1.2	1.06	.001*
Free Recall 17 Questions	210	3.2	1.87	177	4.1	2.06	.001*
Cued Recall Four Categories %	186	30	1.15	176	57.5	1.26	.001*
Cued Recognition 17 Questions Signal %	186	62.6	.31	177	68.4	.27	.057^

* A *p*-value of <.05 indicates significance

** These scores were adjusted to a teacher consensus baseline (teacher consensus is in brackets)

^Marginally significant

Table 2. Results of independent samples t-tests conducted on baseline adjusted scores.

Hypothesis 1: ***Students in the spaced condition will use more information from the lessons to defend their website ratings than students in the massed condition.*** At final test, students were asked to rate the website and defend their rating in a paragraph. Students were marked to see how many of the categories and questions that they used. A correlation between the two raters was calculated in order to determine interrater reliability. The interrater reliability was found to be .91 for the four categories, and .90 for the seventeen questions. Students received a mark out of four if they mentioned one of the four categories (paraphrasing was accepted). For example, if students said, "I saw who created the website," students got a mark in the authority category. As an extension of that, students were also marked out of 17 to see which specific questions they chose to answer during final test. For example, students may have said, "I saw who created the website [adds their name] and I looked them up and it tells me who they are. I think they are experts on the topic and I believe what they are saying." This response would have received three marks for authority—one for mentioning the author's name, one for adding additional details about them, and one for deciding if they're an expert. Results showed that students in the spaced group used significantly more of the four categories than the massed group, $t(337.79) = 4.66$, $p = .001$, and the spaced group also used more of the 17 questions than the massed group, $t(385) = 4.78$, $p =.001$, to defend their ratings.

Hypothesis 2: ***Students in the spaced condition will be have closer ratings to teacher consensus than the massed condition at final test.*** At final test, there was no significant difference of ratings between groups, $t(385) = .32$, $p = .748$, (see *Figure 4* for a histogram of student scores at final test, and *Figure 5* for a visual representation of student raw scores vs. teacher raw scores).

Post Hoc Secondary Analyses of the Final Test

An independent samples t-test was conducted on the final test raw scores, which eliminated extreme ratings (0 or 100) or a middle “undecided” score of 50. There was a non-significant difference between groups when 0, 50, and 100 scores were removed, *t*(271.97) = .27, *p* = .79. A chi-square was conducted in order to compare the 0, 50, and 100 values to all other scores, and revealed a non-significant difference, $\chi^2(1, N = 388) = 2.51, p = .11$.

Classroom effects were also explored, and when classroom averages were used for the t-test (thereby removing the violation of independence issue from use of a nested design), the results were non-significant, *t*(16.92) = .11, p = .91. When dot plots were explored at this stage, it was revealed there was an outlier classroom. This classroom was removed and a t-test was conducted on the remaining classrooms. There was a non-significant difference between groups after the outlier was removed and classroom averages were used, *t*(16.17) = .72, *p* = .48. Welch tests were used for both of the classroom t-tests, and the adjusted degrees of freedom are reported because of a violation of the equality of variance assumption.

Grade effects were also explored post-hoc, and there were no significant differences between the teacher adjusted baseline ratings between the grade four classrooms, *t*(31.75) = 1.15, *p* = .26, grade five classrooms, *t*(88.30) = .17, *p* = .86, or grade six classes, *t*(177.81) = .68, *p* = .5, when comparing the spaced and massed conditions. Welch tests were also used for the grade-separated t-tests, and the adjusted degrees of freedom are reported because of a violation of the equality of variance assumption.

Hypothesis 3: ***Students in the spaced condition will remember more information from the lessons when cued.*** When students were cued and asked to recall the four questions, students in the spaced group remembered significantly more than in the massed group, *t*(360) = 8.13, *p* =

.001. When prompted to recall the 17 questions, there was a marginally significant difference in percent recalled, *t*(358.36) = 1.91, *p* = .057.

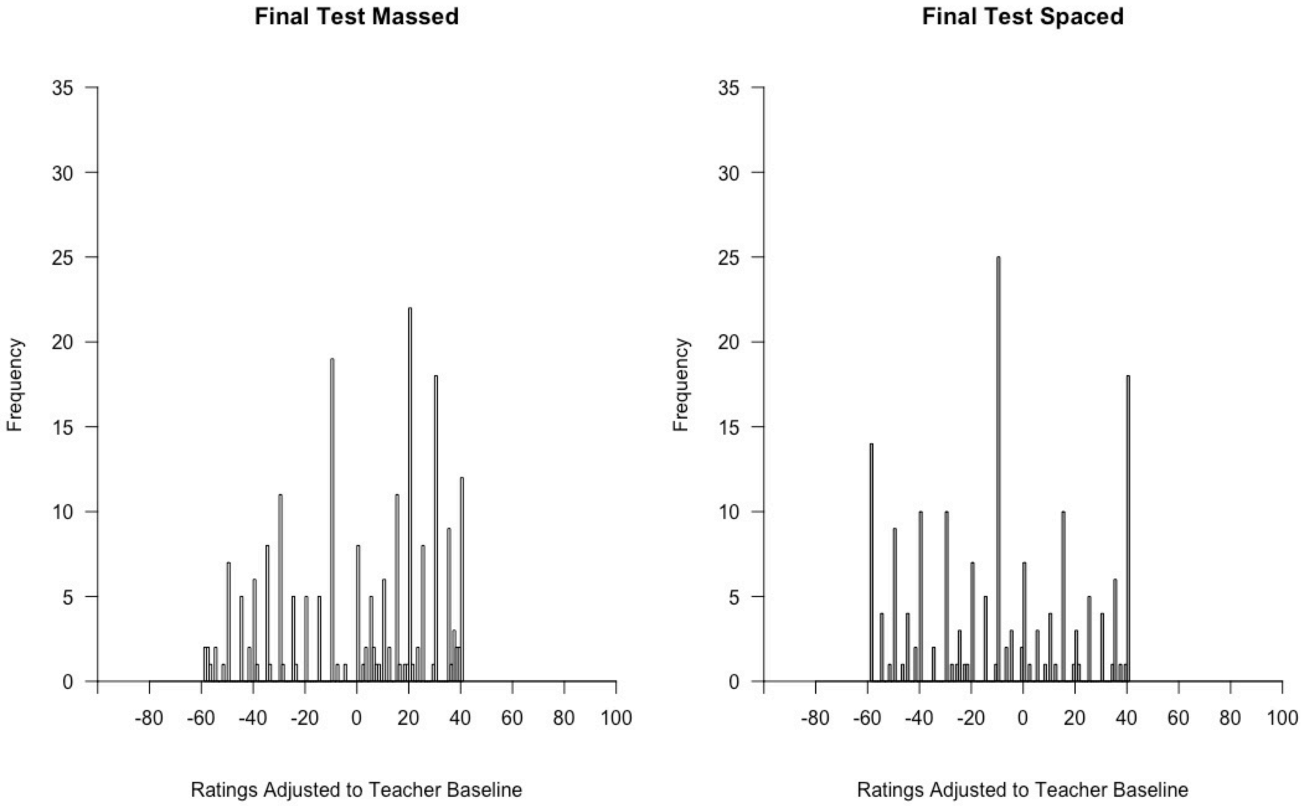


Figure 4. Histogram of scores from the final test. Scores have been adjusted to their respective teacher baselines (*Table 2*).

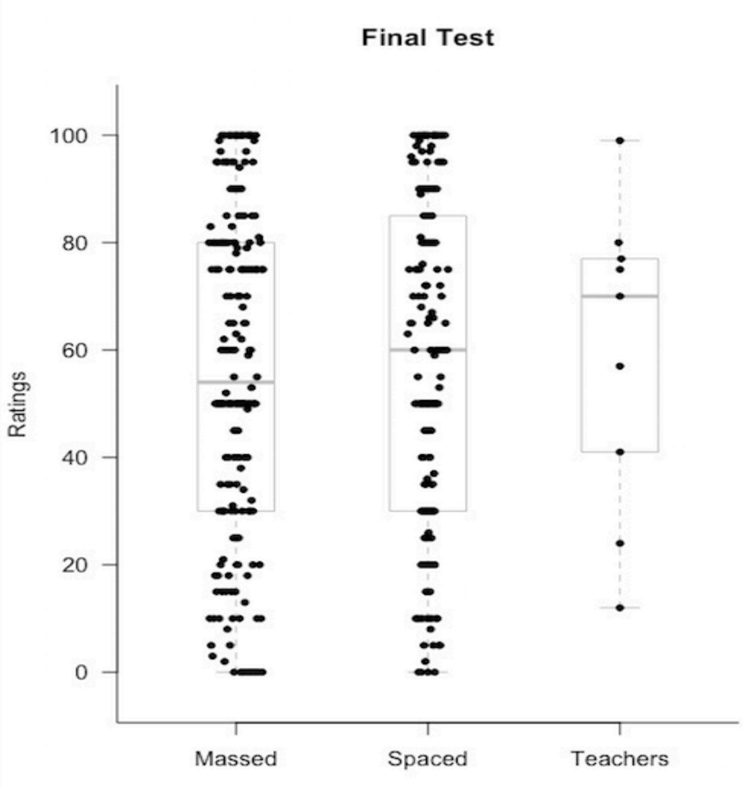


Figure 5. Box plot with overlaid dot plot of student and teacher distribution of raw scores at final test.

Student Hypotheses

A class discussion was led by the researcher after the final test. Students were debriefed, which also served as a teaching opportunity for students to see how the scientific method worked in practice. Before going through the hypotheses, students were told an overview of the methods section and were asked to form their own hypotheses. One class that was part of the massed condition took the time to write their answers down. Their answers are shown in *Table* 3.

Most students who were asked assumed that there would be a spacing advantage, and their suggestions coincided with existing theories of spacing. Many of the students spoke about "letting the information sink in," and described the forgetting process which is central to how spacing works. One of the students who did not write down their answer talked about how their class would have "paid more attention" during spaced sessions because during lessons two and three, the content seemed easier because they had "just done it the day before." This coincides with the study-phase retrieval theory of spacing, and could explain why there was a significant effect of spacing in all of the fact learning aspects of the study.

Student 1: *I think it will be easier to remember information over 3 weeks than everyday for 3 days because you have more time to let it sink in and remember it. Also we did a project last year that we worked on once a week for 10 weeks and I think it was much better than if we did it every day.*

Student 2: *I think it will be easier to remember information over a course of 3 weeks, rather than everyday for 3 days. I think this because over a course of three weeks the information will have more time to process rather than rushing through three days to only forget a month later.*

Student 3: *I think it will be better to remember over 3 weeks and not three days because over the three days you learn it all at once but if you learn over weeks we will be refreshed every week and [we will] be used to remembering the credibility information rather than 3 days which is when we learn it all at once and we are not used to remembering the whole credibility lesson and had some trouble doing the final test!*

Student 4: *I think it would be better over a course of 3 weeks rather than 3 days because its more time for it to sink in and think about than 3 days where you put new info in and leave it as short term [memory] rather than a long term [memory].*

Student 5: *I think it would be better if we did it every 3 weeks not every 3 days because in 3 weeks you would have to remember all the stuff in a week and if you did 3 days you would remember quickly but when you take the test in a month you would forget all the stuff that they told you in 3 days. Therefore I think it is better to take the 3 weeks class.*

Student 6 and 7 (responded together): *We think that we will remember information better after 3 days in a row. We think this because after each day if we forget something we can relearn it the next day. Then because our minds are refreshed [and] we can learn something new and learn better.*

Student 8: *I think you will learn better over 3 weeks. Reason is because then when you have to wait 1 month, you will be used to remembering over long periods of time, and you will remember more.*

Table 3. Student hypotheses when asked the question, “[after a month passes], do you think you will remember information better after 3 days in a row or once a week for 3 weeks?” Spelling errors were corrected when typing student answers.

General Discussion

Connection with Fact Learning

The current study explored spacing in a classroom setting, using curriculum-based materials and relevant inter-study intervals and retention intervals. Since the spacing effect has been shown to be robust with fact learning material (Cepeda et al., 2006), it was important that students in the spacing condition show a memory effect on the recall and recognition test (*Appendix E*).

The students who were in the spaced condition remembered more of the four questions (M=2.3) than the massed group (M=1.2). The spaced group also remembered more of the questions after a signal score was taken (hits-false alarms). The spaced group accurately recalled an average of 68.4%, and the massed group accurately recalled an average of 62.6%. These results help to support the idea that that the spacing effect can hold in a more ecologically valid setting than the laboratory. The study-phase retrieval theory (Thios and D'Agostino, 1976; Rawson and Dunlosky, 2012) may help to explain why there was an effect of spacing on these measures.

At the beginning of SS2 and SS3, students were asked to recall the four categories and 17 questions of website evaluation through a group discussion. Students in the massed condition seemed to have a much easier time remembering the information since they had been exposed to the information the day before, during SS2. The spaced group, on the other hand, often struggled to remember the categories during the second lesson (students needed lots of prompting and reminders), but got noticeably better during the discussion at the beginning of third lesson. This may have been because after students in the spaced condition failed to remember the information at the beginning of the second lesson, they evaluated their strategies and adapted their methods so that they would be more successful next time around. The discussions at the beginning of the lessons were vital because the relearning helped students to build a fundamental knowledge base, which students used through the remainder of the lessons and then on the final test after 35 days.

Also in line with traditional fact learning studies were the results of the website ratings during SS1, SS2, and SS3. In an online fact learning study that taught participants random trivia facts, researchers found that performance increases immediately after initial learning and then forgetting begins to occur after approximately 7 days (Cepeda et al., 2008). As expected, as the gap becomes longer, more forgetting occurs. A classroom study by Carpenter, Pashler and Cepeda (2009) tested student retention of history facts, and found that students who were given immediate recall performed better after one week than students who were given a test after 16 weeks. In light of this research evidence, in the current study we did not have a hypothesis about how students would fare during the lessons since they were given a refresher at the beginning of each study session. This was done in order to mimic standard teaching practice where students receive a review of the content before having to do an assignment or task. Despite the fact that previous studies did not include refreshers at the beginning of the study session, in the current study we found similar results to that of traditional studies even when this relearning occurred.

Data were collected during the lessons, and are summarized in *Table 2*. There was no significant difference between the two groups when rating the overall website during SS2, *t*(383)=3.59, p=.299 (*Figure 6*). However, during SS3, there was a significant difference between the two groups when rating the overall website, *t*(385)=3.59, p=.001—the massed group was significantly closer than the spaced group to teacher consensus (*Figure 7*). This suggests that, as in traditional spacing studies that used factual material, massed learning produces a significant advantage over spacing when recall is immediate. Students who took part in the massed condition had an advantage over the spaced group by the third study session. *Figures 8* and *9* show a distribution of the raw scores of both students and teachers during SS2 and SS3.

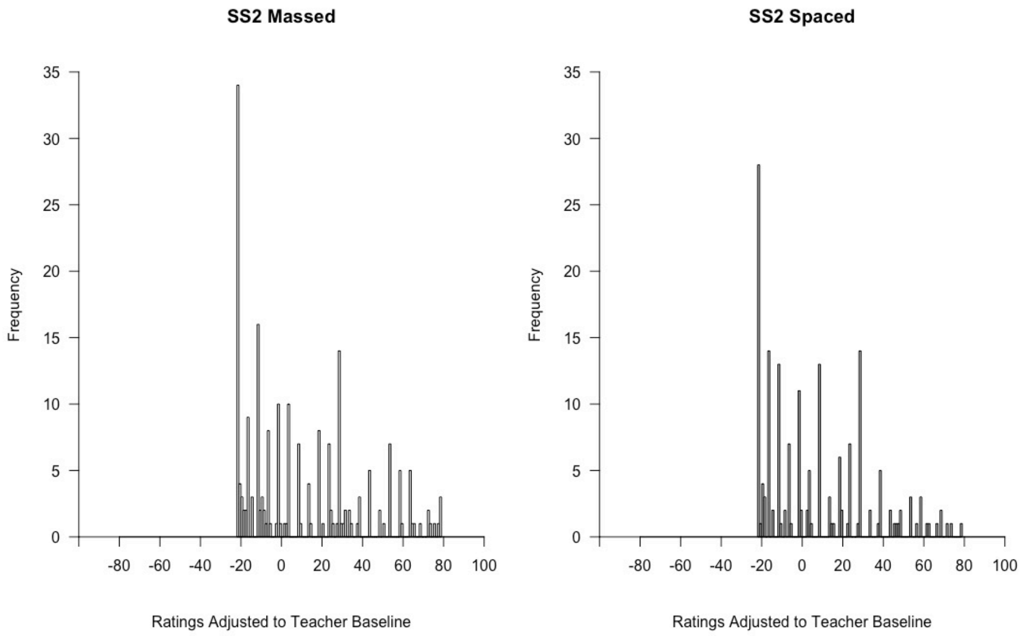


Figure 6. Histogram of SS2 Ratings adjusted to teacher baseline.

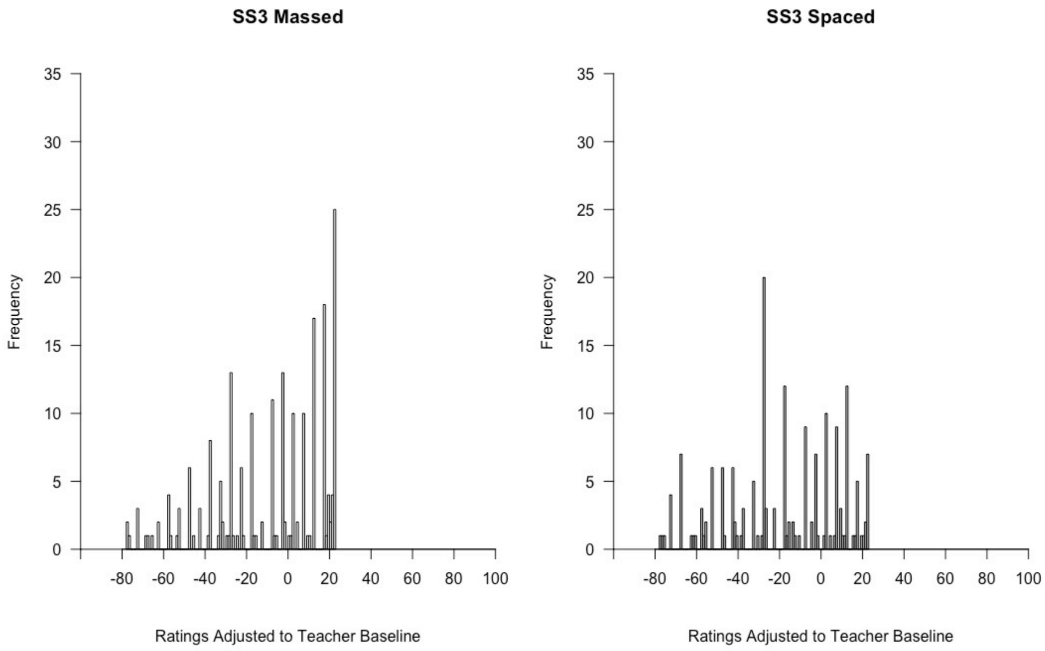


Figure 7. Histogram of SS3 ratings adjusted to teacher baseline.

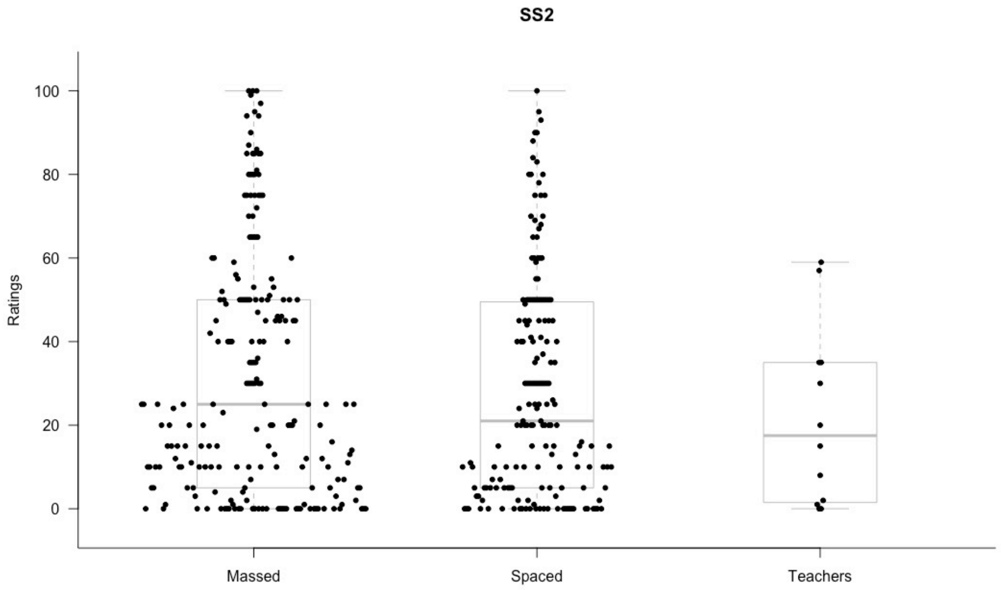


Figure 8. Box plot with overlaid dot plot of student and teacher raw scores in SS2.

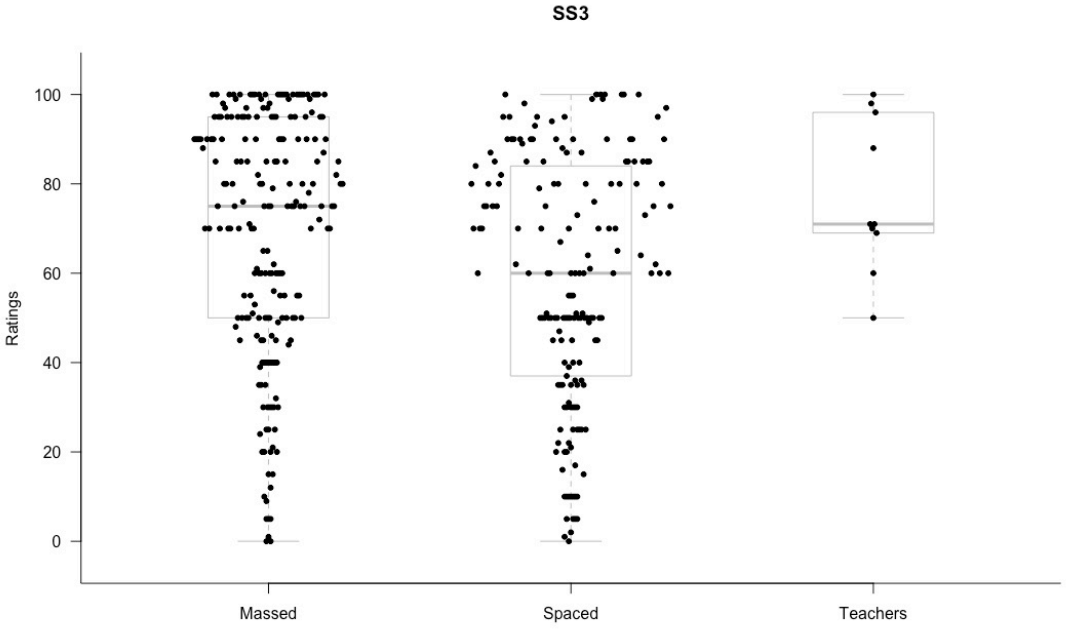


Figure 9. Box plot with overlaid dot plot of student and teacher raw scores SS3.

Connection with Critical Thinking

All of the lessons were associated with critical thinking because students were encouraged to judge everything they were exposed to online. Immediately after students were shown the first website and learned that it was a hoax, their guards were up for the remainder of the lessons. Students became extremely critical of what they saw online, searching for flaws in all aspects of the websites. Even though they may not always have been able to pinpoint the problems with what they encountered online, senses were tuned and critical thinking was taking place (Gilbert, 2014).

In an effort to teach students how to identify key features of the websites, we taught them exactly what to look at and how to discuss what they saw with the purpose of coming to a specific outcome (completion of the rating). This is closely linked to Ennis' (1987) perception of a good critical thinker—someone who uses reasonable, reflective thinking when deciding what to believe or do, and also Dewey (1909), who asked critical thinkers to be objective and consider their options while taking time to investigate before deciding whether or not they agree. In order to do this successfully, students needed the foundational materials that they could use to explain their perspective. Although students also had to remember the questions and the categories, they also had to manipulate and apply these facts to their specific situation. There was a goal (evaluating the website and giving it a rating) and students used what they learned to meet this goal.

The fact that students in the spaced condition used more of the categories and questions to explain their rating suggests that they were better able to communicate their perspective and explain why they chose the option (rating) that they chose. See *Figure 10* for an example of a student answer from SS2 (where students were still learning skills) and from the final test.

SS2: Tree Octopus Rating 30/100 Please explain your rating using information from the website. I gave 30% because the design was organized well and because it didn't say much about the author and his own webpage about himself is all comics and says nothing about him being an expert. It says when it was made and it was last updated in 2015 so the content seems reasonable. It's a good purpose to save the endangered animals but the things they want you to do are crazy! For example, "participate in Tree Octopus marches. You can demonstrate their plight during the march by having your friends dress up as tree octopuses while you attack them in a lumberjack costume."
Post: Mike the Headless Chicken Rating 15/100 Please explain your rating using information from the website. For authority it doesn't really show who made the website but it shows who designed the website. The purpose of the website well I don't know all the website is saying it what he did and how long he survived. For the links they work so that is good for the website. It also says that "mike is going for the gold" but that looks fake because he is dead now. It says that the designer is aha consulting so that shows that the design is real or probably real because I looked them up and they are a real thing. In the website it doesn't look like there is any spelling mistakes so that is good. The author looks like he is trying to persuade me that mike the headless chicken is real and he is also trying to persuade me to buy something like shirts and shows pictures and prices. To me I don't like the content and the pictures because I don't trust them and the fact that the author is showing me pictures of the headless chicken that all look fake. My last thing to me is that this website doesn't look credible.

Table 4. Examples of website evaluations from SS2 (practice) and at post.

Challenges and Limitations

One of the biggest challenges and limitations of this study was choosing the websites. It was difficult to find sites that were ambiguous so that ratings were not always obvious. The websites were selected so that they could fall on a continuum on the rating scale—most sites were good at one thing but failed in another (for example, the Tree Octopus SS2 design was beautiful, but the content was terrible). This was done in an effort to teach students to be objective and not give extreme ratings. However, because of the challenging nature of the websites, there was high variability in the sample. Websites at SS1, SS2, and SS3 were more clearly credible/not credible, whereas the final test website was debatable. As a result, student ratings did not differ from teacher consensus between the two groups at final test, and there was large variability in the samples of both students and teachers (*Figure 5*). Teachers rated the final website an average of 59.4 out of 100. The median teacher score was 61.5. Of the students, the spaced group rated the website an average of 56.5, with a median of 50 and mode of 50, and the massed group rated it an average of 55.5, with a median of 50 and mode of 50. The website was about a headless chicken who was able to survive for 85 days because its brain stem was left intact—an unbelievable, but true story. Students who remembered to check outside sources were able to find the story of Mike the Headless Chicken online, but others used their prior knowledge to say that living without a head was impossible. Both answers were fair, but if students had remembered their training, they would have known to consider the four categories and not be biased from only one piece of information (their prior knowledge). As a result, their ratings were often skewed to 0 or 100 once participants, both students and teachers alike, had the confidence to make a decision.

The credibility scale which would produce ratings from 0-100 was originally chosen to explain that determining credibility was a fluid process. Additionally, students are often asked to use number lines in class when completing mathematical tasks, so the linear scale during the lessons would have been familiar to them. However, the ratings may not have been an accurate depiction of website credibility for several reasons. First of all, students may have been biased in their ratings. Although students were not told to rate the sites based on whether they were real or fake, some gave extreme ratings based on the idea that fake websites should be near zero, whereas real websites should be near 100. Even if the websites were real or fake, there were always redeeming qualities which students may not have considered because of their biases. Secondly, there may have been individual differences in understanding instructions. Lastly, students may have had individual differences in how they chose to rate the websites. For example, some students may have chosen to stay between 25-75 (as to avoid extreme ratings), some chose to be non-committal and stay at 50, and some may have just been prone to giving extreme ratings of 0-5 and 95-100 due to their confidence in the website.

Another limitation was in the materials themselves. The narrative section given to students during the pre-test (SS1) and final test had space for students to write their answers down in a paragraph. However, it became clear when marking the tests that students may not have had room on the paper to discuss all of the 17 questions if they wanted to. Perhaps there should have been more space available so that students could have been encouraged to write more. A high number of the students only used one or two sentences to support their rating on the final test.

Another more general limitation is the reality of conducting classroom research. Each classroom varies on so many levels that are difficult to control, and student performance can depend on many factors. There are also constant interruptions in the classroom. During this study we had a fire drill, a lockdown procedure, three snow days that caused minor shifts in the sessions, and teachers were absent on certain days, which caused a break in student routine.

Conclusion and Future Directions

This study attempted to address concerns that there is not yet enough research on the spacing effect in the classroom (Dempster, 1988), and that many of the classroom studies that have been run before have not used educationally relevant inter-study intervals and retention intervals (Cepeda et al., 2006). In addition, this study looked at whether effects of spacing can extend to include critical thinking skills.

One of the primary goals of any education system is to help students remember the information they are taught at school so that they can apply it to their everyday lives. This may require students to hold on to specific pieces of information for extended periods of time, only to see it out of context and have to manipulate it to fit a certain situation. As such, if spacing studies in the classroom prove to enhance retention of these critical thinking skills, it may help to reduce forgetting, which can sometimes cause barriers to student learning.

If spacing in the classroom proves to be as useful as researchers think it should be (e.g., it is listed twice in NCTQ [2016] recommendations), it is a relatively easy adjustment to make in classrooms. Teachers would not have to directly adjust their lesson plans, but rather the timing of how they are delivered to students. This could be done at the beginning of the year when creating long-range plans, and would only require slight alterations in order to be successful.

A possible next step would be to explore whether spacing is effective on a wider scale, using different teachers and subjects. In this study, lessons were carried out by one Ontario certified teacher to help control for teacher effects. In the real world, however, teachers are vastly different. Researchers need to investigate whether results that are seen still occur when different teachers are using the same lesson plans, as would happen in the typical classroom. In addition, a variety of curriculum-based subject material should be used (inquiry math would be an excellent choice). If those factors can be explored, perhaps recommendations can be made to start using spacing in the classroom. I hope that the present study will encourage future investigations on the spacing effect and real-world classroom learning.

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. doi:10.3102/0034654308326084

Bahrick, H.P. & Hall, L.K. (2005). The importance of retrieval failures to long-term retention: A cognitive explanation of the spacing effect. Journal of Memory and Language, 52, 566-577. doi:10.1016/j.jml.2005.01.012

Baron, J. (2000). Thinking and Deciding. New York: Cambridge University Press.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. Applied Psycholinguistics, 31, 635–650. doi:10.1017/S0142716410000172

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.

Brodin, E. (2007). *Critical thinking in scholarship: Meanings, conditions and development*. Lund University, Department of Education: Media –Tryck, Sociologen.

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.

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., 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

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

Dewey, J. (1991). How We Think. New York: Prometheus Books. (Original work published 1910).

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

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., & Millman, J. (2005a). Cornell critical thinking test (5^{th} Edition). Pacific Grove, CA: Critical Thinking Books & Software.

Ennis, R.H. and Millman, J. (2005b). Cornell critical thinking test, level Z (5^{th} edition). Pacific Grove, CA: Midwest Publications.

Ennis, Robert H., Millman, Jason, & Tomko, Thomas N. (2005). Cornell critical thinking tests: Administration manual (Fifth Edition). Seaside, CA: The Critical Thinking Company.

Ennis, R.H. and Weir, E. (1985). The Ennis-Weir critical thinking essay test. Pacific Grove, CA: Midwest Publications.

Everhart, N. (1996). Web page evaluation worksheet. Retrieved February 28, 2016, from terpconnect.umd.edu/~cpikas/MLSfiles/650/webevaluation.doc

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)

Gilbert, M. A. (2014). Arguing with people. Peterborough, Ont.: Broadview Press.

Glaser, E. M. (1941). An experiment in the development of critical thinking. New York, NY: Teachers College, Columbia University.

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

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

Halpern, D.F. (1998). Teaching critical thinking for transfer across domains. American Psychologist, 53(4), 449-455.

Halpern, D. F. (2003). *Thought and Knowledge: An Introduction to Critical Thinking* (4th Edition). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

Harris Poll. (2014, May 9). Pearson Student Mobile Device Survey 2014: Grades 4 through 12 (Rep.). Retrieved March 25, 2016, from Pearson Education website:

http://www.pearsoned.com/wp-content/uploads/Pearson-K12-Student-Mobile-Device-Survey-050914-PUBLIC-Report.pdf

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction*, *36*, 38-45.

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

McPeck, J. E. (1981). *Critical thinking and education*. New York: John Wiley.

Niu, L., Behar-Horenstein, L.S., & Garvan, C.W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. Educational Research Review, 9, 114-128.

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

Paul, R. W. (1985). The critical thinking movement: A historical perspective. National Forum: Phi Kappa Phi Journal, 42, 2–3.

Paul, R. W. (1992). Critical thinking: What, why, and how? *New Directions for Community Colleges, 1992*(77), 3–24.

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

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

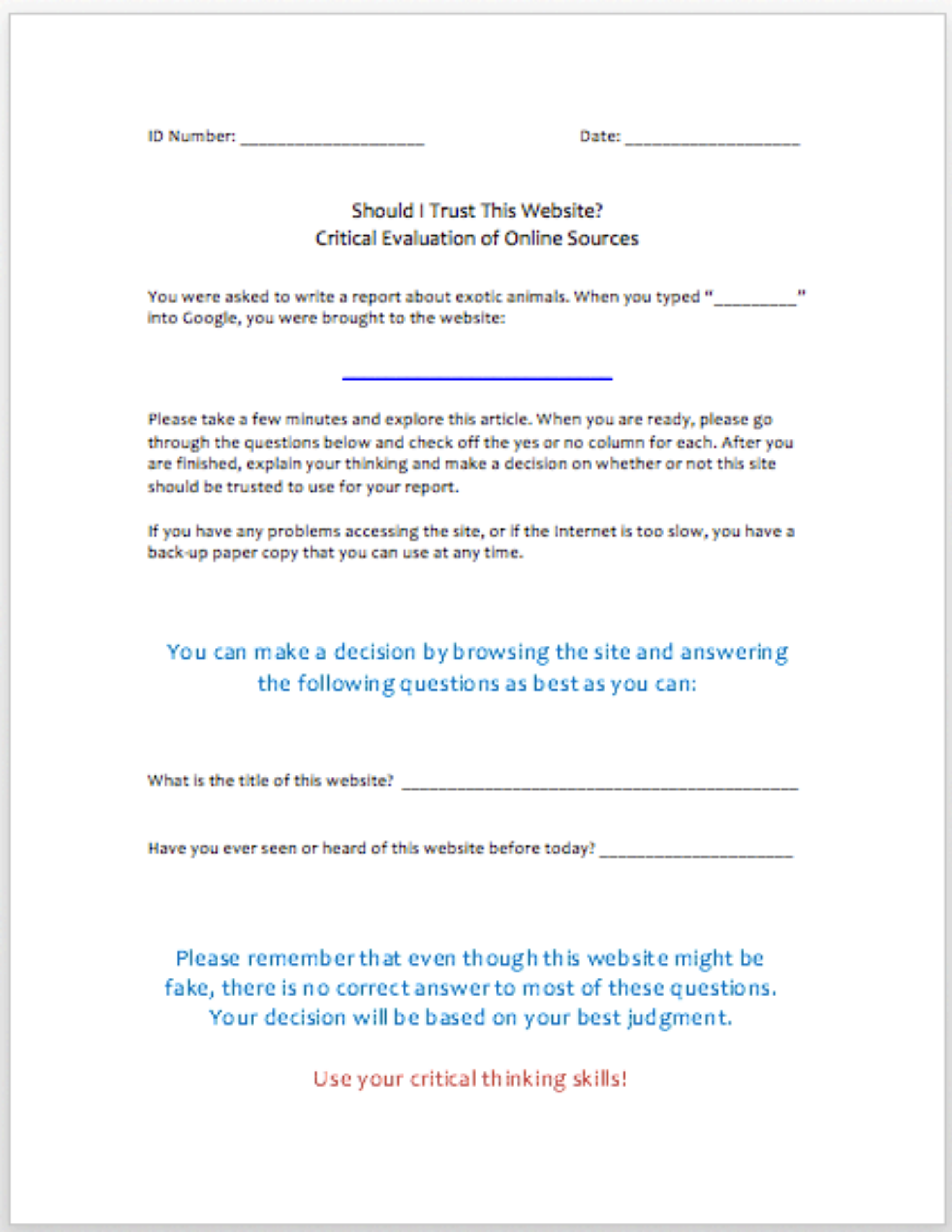
Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1, 341-358*.

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, 5, 529-536.*

Yazdani, M. A., & Zebrowski, E. (2006). Spaced reinforcement: An effective approach to enhance the achievement in plane geometry. *Journal of Mathematical Sciences and Mathematics Education, 1*, 37-43.

2learn.ca Education Society of Alberta. (2004). 2learn.ca netcheck: Web site evaluation form for grades 7-9 students. Retrieved February 28, 2016, from http://www.2learn.ca/evaluating/div3netscheck2.html

Appendix A



ID Number: ____________________ Date: ___________________

Should I Trust This Website?
Critical Evaluation of Online Sources

You were asked to write a report about exotic animals. When you typed "_________" into Google, you were brought to the website:

____

Please take a few minutes and explore this article. When you are ready, please go through the questions below and check off the yes or no column for each. After you are finished, explain your thinking and make a decision on whether or not this site should be trusted to use for your report.

If you have any problems accessing the site, or if the Internet is too slow, you have a back-up paper copy that you can use at any time.

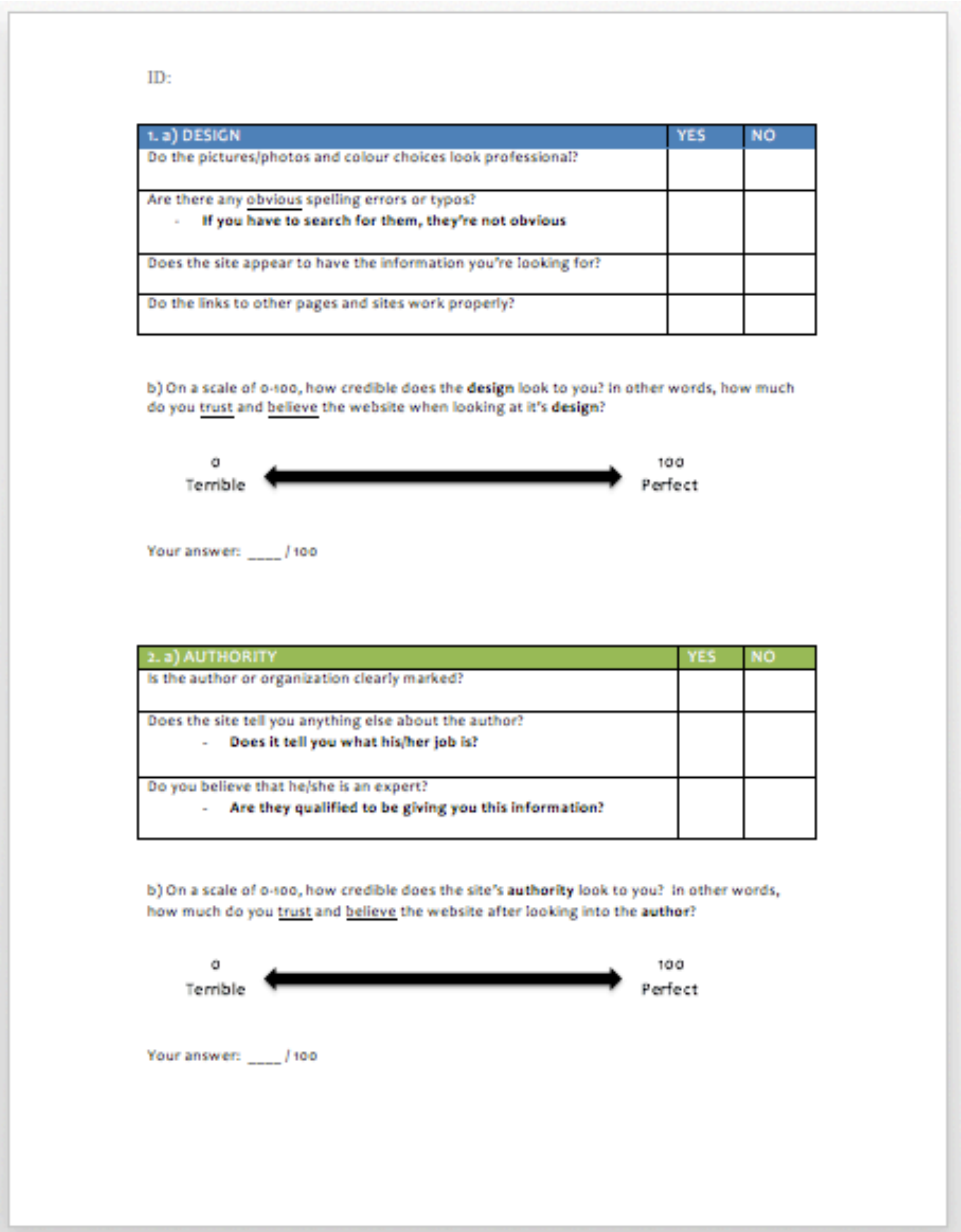
You can make a decision by browsing the site and answering the following questions as best as you can:

What is the title of this website? ____

Have you ever seen or heard of this website before today? _____________________

Please remember that even though this website might be fake, there is no correct answer to most of these questions. Your decision will be based on your best judgment.

Use your critical thinking skills!



ID:

1. a) DESIGN	YES	NO
Do the pictures/photos and colour choices look professional?		
Are there any obvious spelling errors or typos? - **If you have to search for them, they're not obvious**		
Does the site appear to have the information you're looking for?		
Do the links to other pages and sites work properly?		

b) On a scale of 0-100, how credible does the **design** look to you? In other words, how much do you trust and believe the website when looking at it's **design**?

0 Terrible ⟷ 100 Perfect

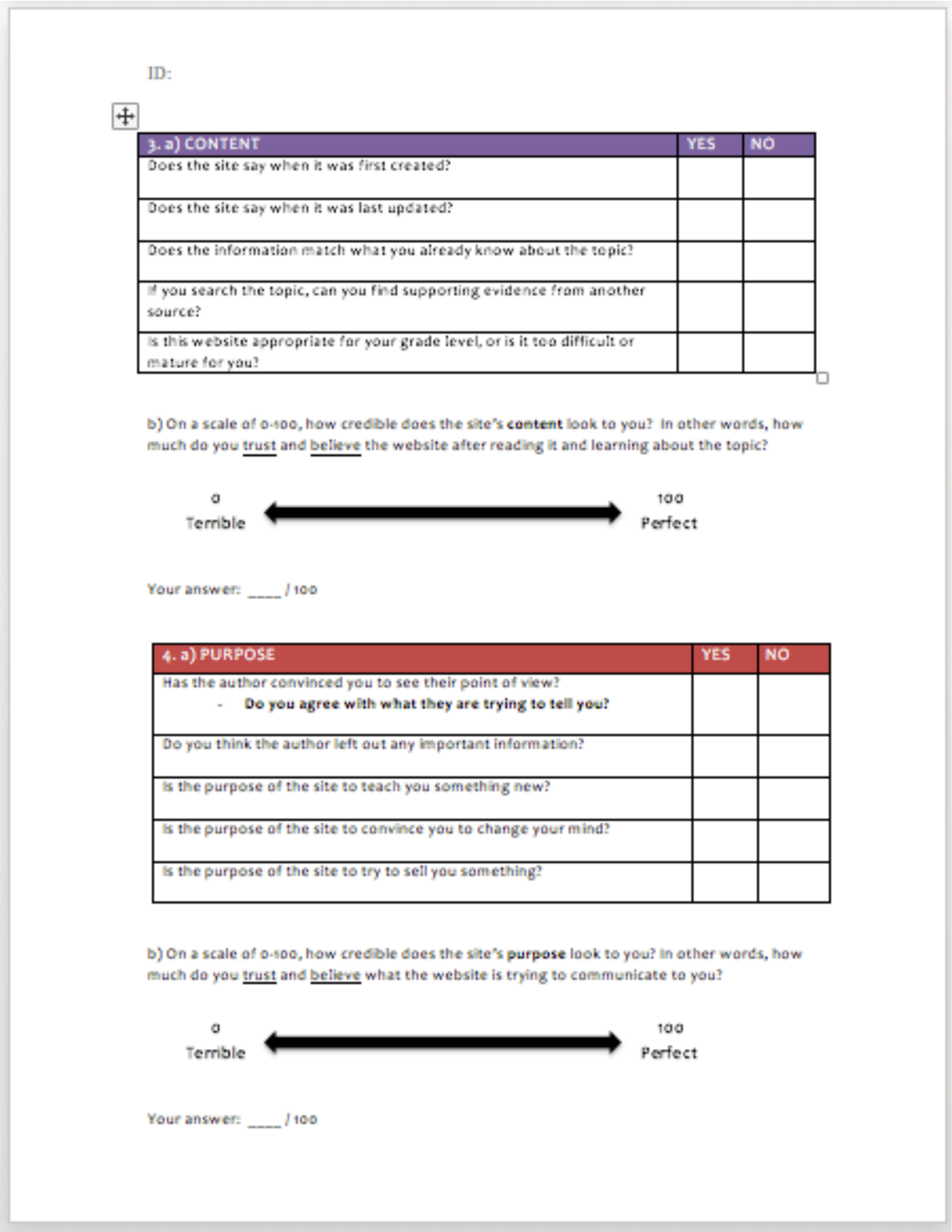
Your answer: ____ / 100

2. a) AUTHORITY	YES	NO
Is the author or organization clearly marked?		
Does the site tell you anything else about the author? - **Does it tell you what his/her job is?**		
Do you believe that he/she is an expert? - **Are they qualified to be giving you this information?**		

b) On a scale of 0-100, how credible does the site's **authority** look to you? In other words, how much do you trust and believe the website after looking into the **author**?

0 Terrible ⟷ 100 Perfect

Your answer: ____ / 100



ID:

3. a) CONTENT	YES	NO
Does the site say when it was first created?		
Does the site say when it was last updated?		
Does the information match what you already know about the topic?		
If you search the topic, can you find supporting evidence from another source?		
Is this website appropriate for your grade level, or is it too difficult or mature for you?		

b) On a scale of 0-100, how credible does the site's **content** look to you? In other words, how much do you trust and believe the website after reading it and learning about the topic?

0 Terrible ⟷ 100 Perfect

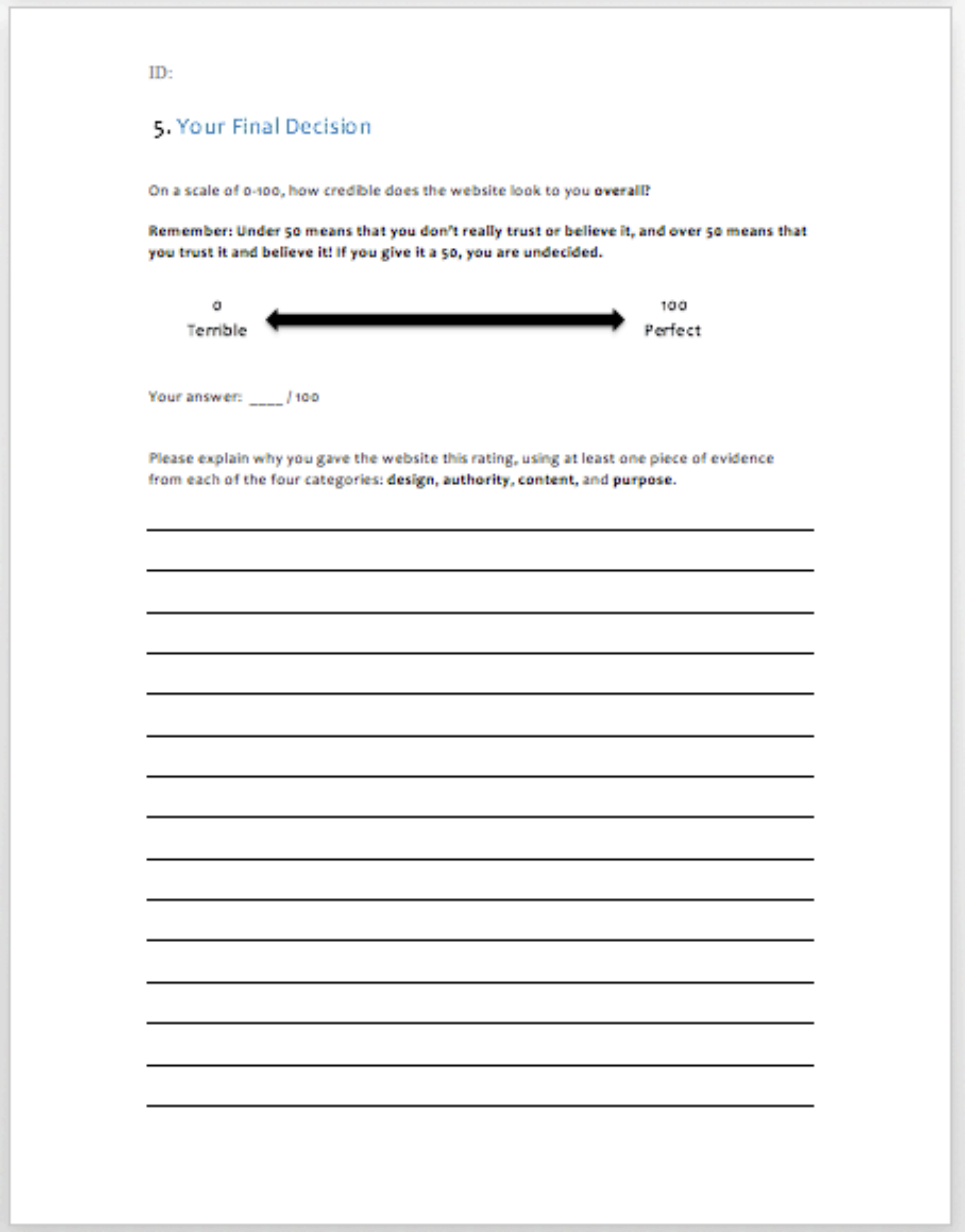
Your answer: ____ / 100

4. a) PURPOSE	YES	NO
Has the author convinced you to see their point of view? - **Do you agree with what they are trying to tell you?**		
Do you think the author left out any important information?		
Is the purpose of the site to teach you something new?		
Is the purpose of the site to convince you to change your mind?		
Is the purpose of the site to try to sell you something?		

b) On a scale of 0-100, how credible does the site's **purpose** look to you? In other words, how much do you trust and believe what the website is trying to communicate to you?

0 Terrible ⟷ 100 Perfect

Your answer: ____ / 100



ID:

5. Your Final Decision

On a scale of 0-100, how credible does the website look to you **overall?**

Remember: Under 50 means that you don't really trust or believe it, and over 50 means that you trust it and believe it! If you give it a 50, you are undecided.

0 Terrible ⟷ 100 Perfect

Your answer: ____ / 100

Please explain why you gave the website this rating, using at least one piece of evidence from each of the four categories: **design**, **authority**, **content**, and **purpose**.

____

____

____

____

____

____

____

____

____

____

____

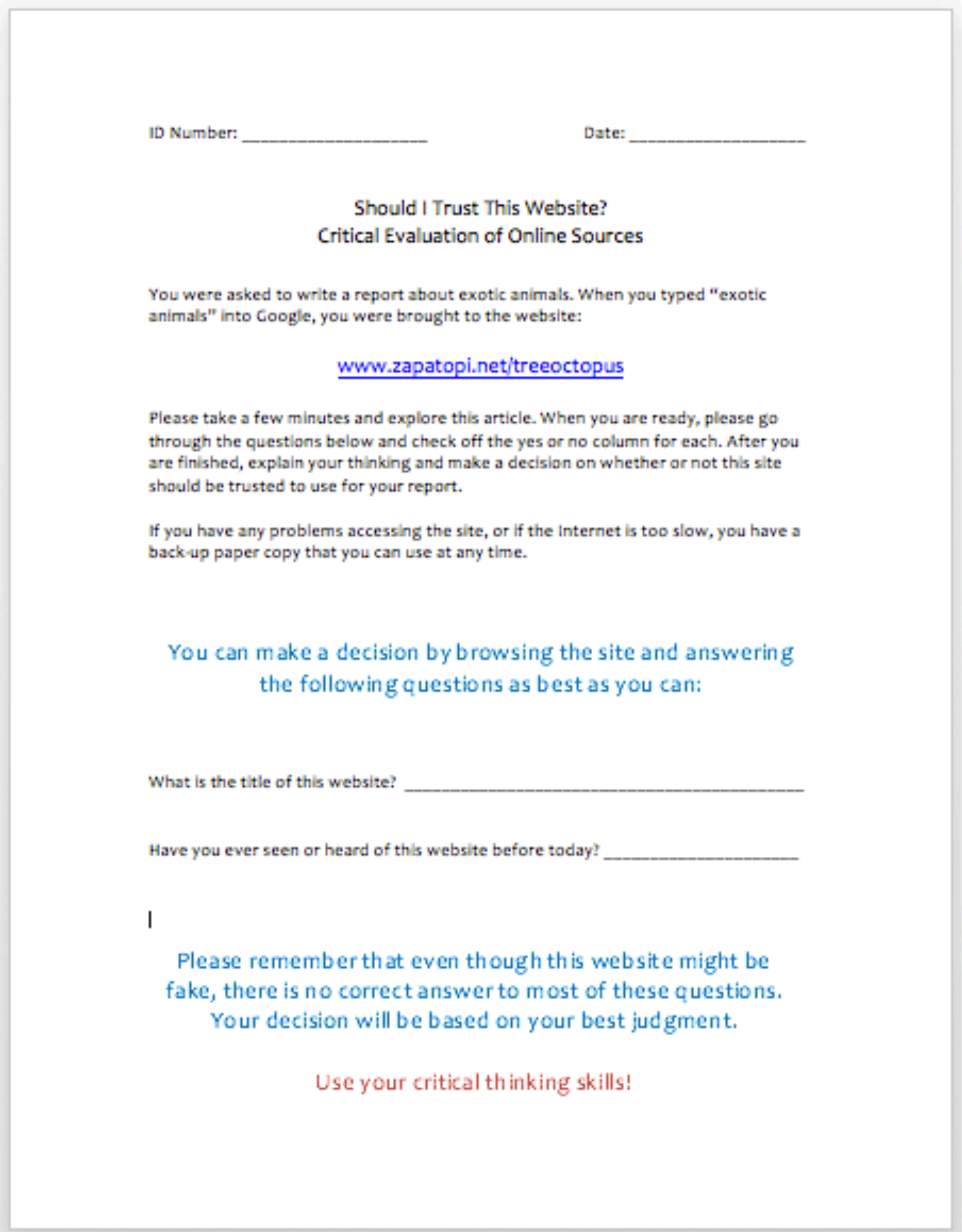
____

____

____

____

Appendix A1



ID Number: ____________________ Date: ___________________

Should I Trust This Website?
Critical Evaluation of Online Sources

You were asked to write a report about exotic animals. When you typed "exotic animals" into Google, you were brought to the website:

www.zapatopi.net/treeoctopus

Please take a few minutes and explore this article. When you are ready, please go through the questions below and check off the yes or no column for each. After you are finished, explain your thinking and make a decision on whether or not this site should be trusted to use for your report.

If you have any problems accessing the site, or if the Internet is too slow, you have a back-up paper copy that you can use at any time.

You can make a decision by browsing the site and answering the following questions as best as you can:

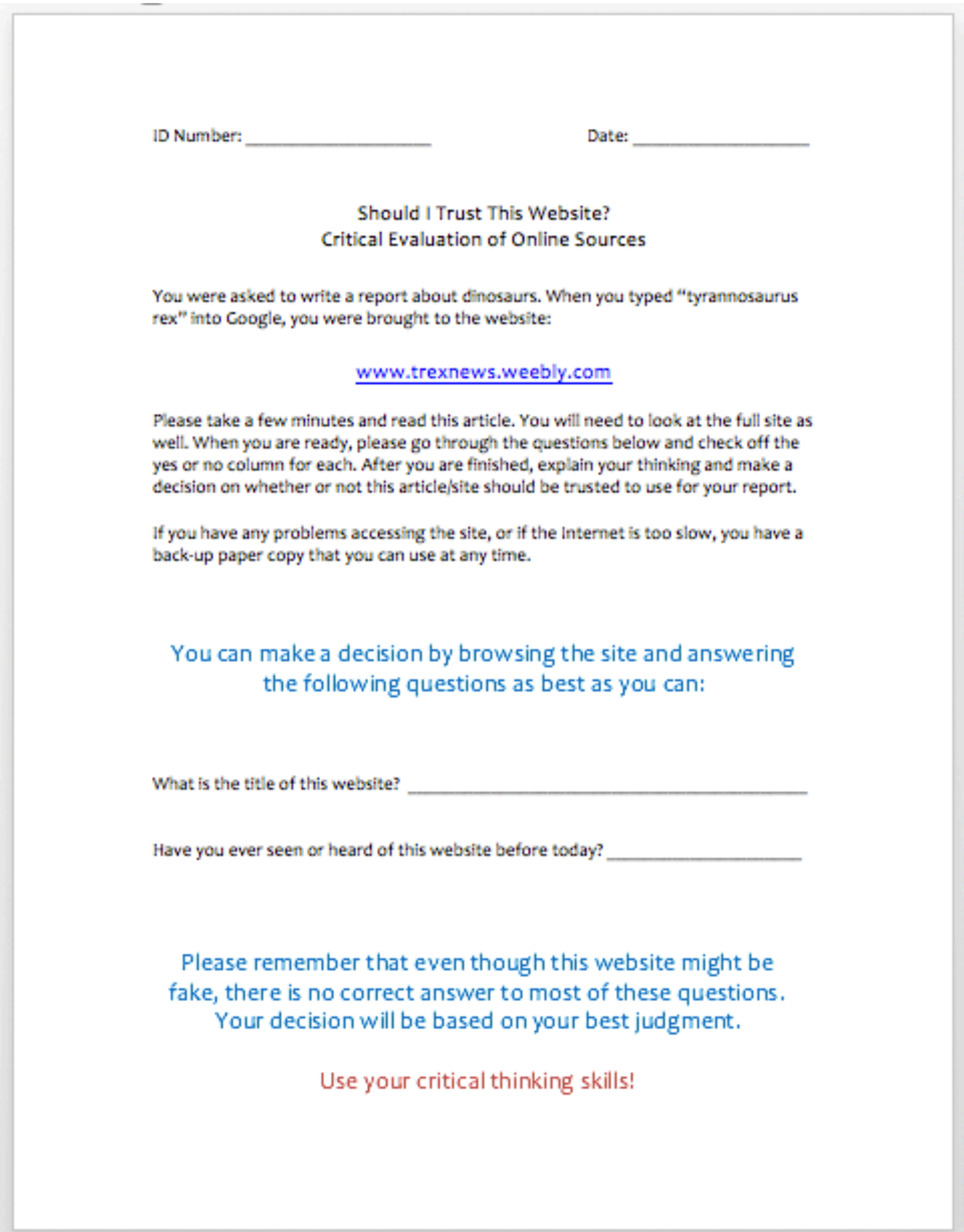
What is the title of this website? ____

Have you ever seen or heard of this website before today? _____________________

Please remember that even though this website might be fake, there is no correct answer to most of these questions. Your decision will be based on your best judgment.

Use your critical thinking skills!

Appendix A2



ID Number: ____________________ Date: ____________________

Should I Trust This Website?
Critical Evaluation of Online Sources

You were asked to write a report about dinosaurs. When you typed "tyrannosaurus rex" into Google, you were brought to the website:

www.trexnews.weebly.com

Please take a few minutes and read this article. You will need to look at the full site as well. When you are ready, please go through the questions below and check off the yes or no column for each. After you are finished, explain your thinking and make a decision on whether or not this article/site should be trusted to use for your report.

If you have any problems accessing the site, or if the internet is too slow, you have a back-up paper copy that you can use at any time.

You can make a decision by browsing the site and answering the following questions as best as you can:

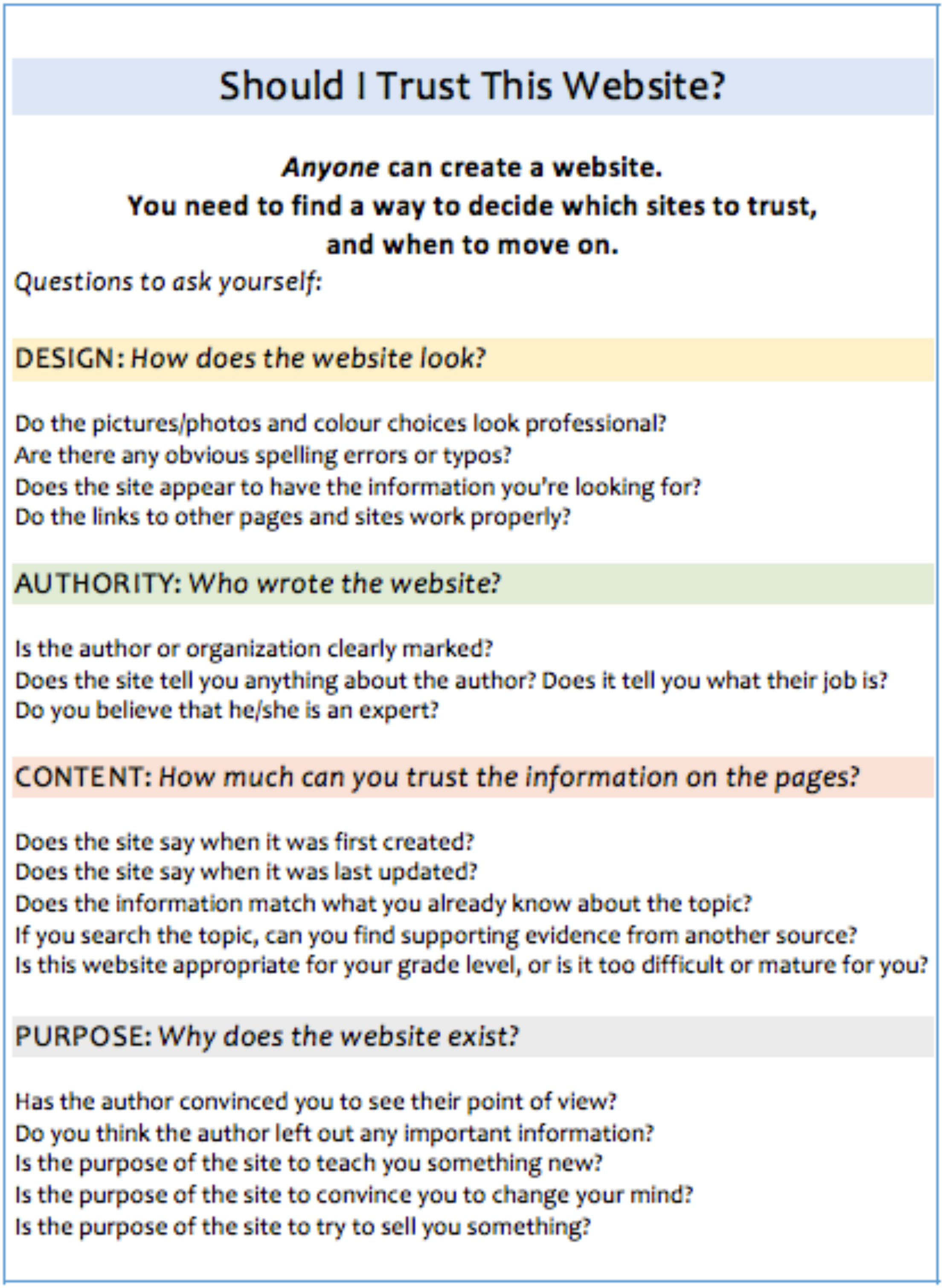
What is the title of this website? ____

Have you ever seen or heard of this website before today? ____________________

Please remember that even though this website might be fake, there is no correct answer to most of these questions. Your decision will be based on your best judgment.

Use your critical thinking skills!

Appendix B



Should I Trust This Website?

Anyone **can create a website.**
You need to find a way to decide which sites to trust,
and when to move on.

Questions to ask yourself:

DESIGN: *How does the website look?*

Do the pictures/photos and colour choices look professional?
Are there any obvious spelling errors or typos?
Does the site appear to have the information you're looking for?
Do the links to other pages and sites work properly?

AUTHORITY: *Who wrote the website?*

Is the author or organization clearly marked?
Does the site tell you anything about the author? Does it tell you what their job is?
Do you believe that he/she is an expert?

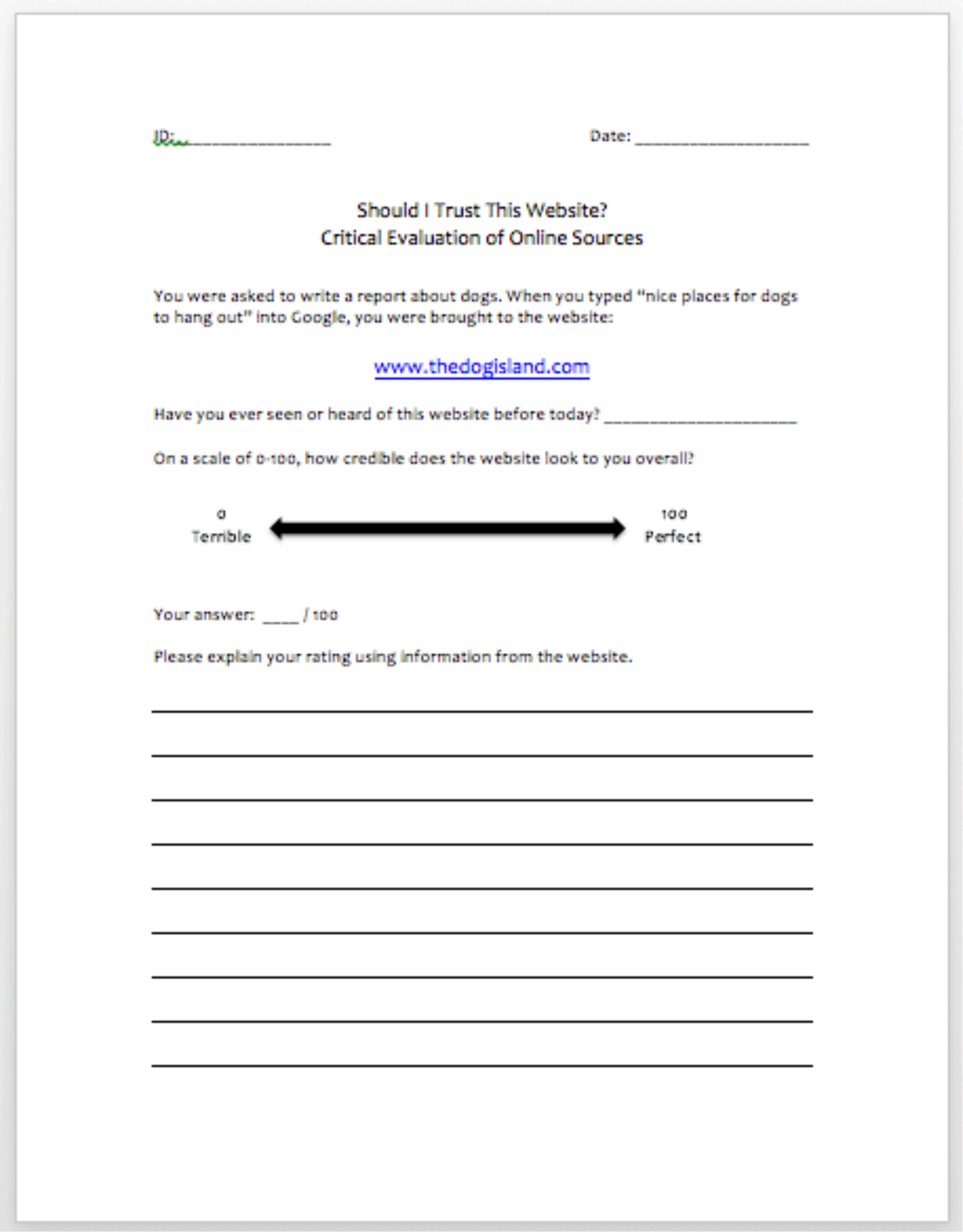
CONTENT: *How much can you trust the information on the pages?*

Does the site say when it was first created?
Does the site say when it was last updated?
Does the information match what you already know about the topic?
If you search the topic, can you find supporting evidence from another source?
Is this website appropriate for your grade level, or is it too difficult or mature for you?

PURPOSE: *Why does the website exist?*

Has the author convinced you to see their point of view?
Do you think the author left out any important information?
Is the purpose of the site to teach you something new?
Is the purpose of the site to convince you to change your mind?
Is the purpose of the site to try to sell you something?

Appendix C



ID:_______________ Date: ___________________

Should I Trust This Website?
Critical Evaluation of Online Sources

You were asked to write a report about dogs. When you typed "nice places for dogs to hang out" into Google, you were brought to the website:

www.thedogisland.com

Have you ever seen or heard of this website before today? _____________________

On a scale of 0-100, how credible does the website look to you overall?

0 Terrible ⟷ 100 Perfect

Your answer: ____ / 100

Please explain your rating using information from the website.

____

____

____

____

____

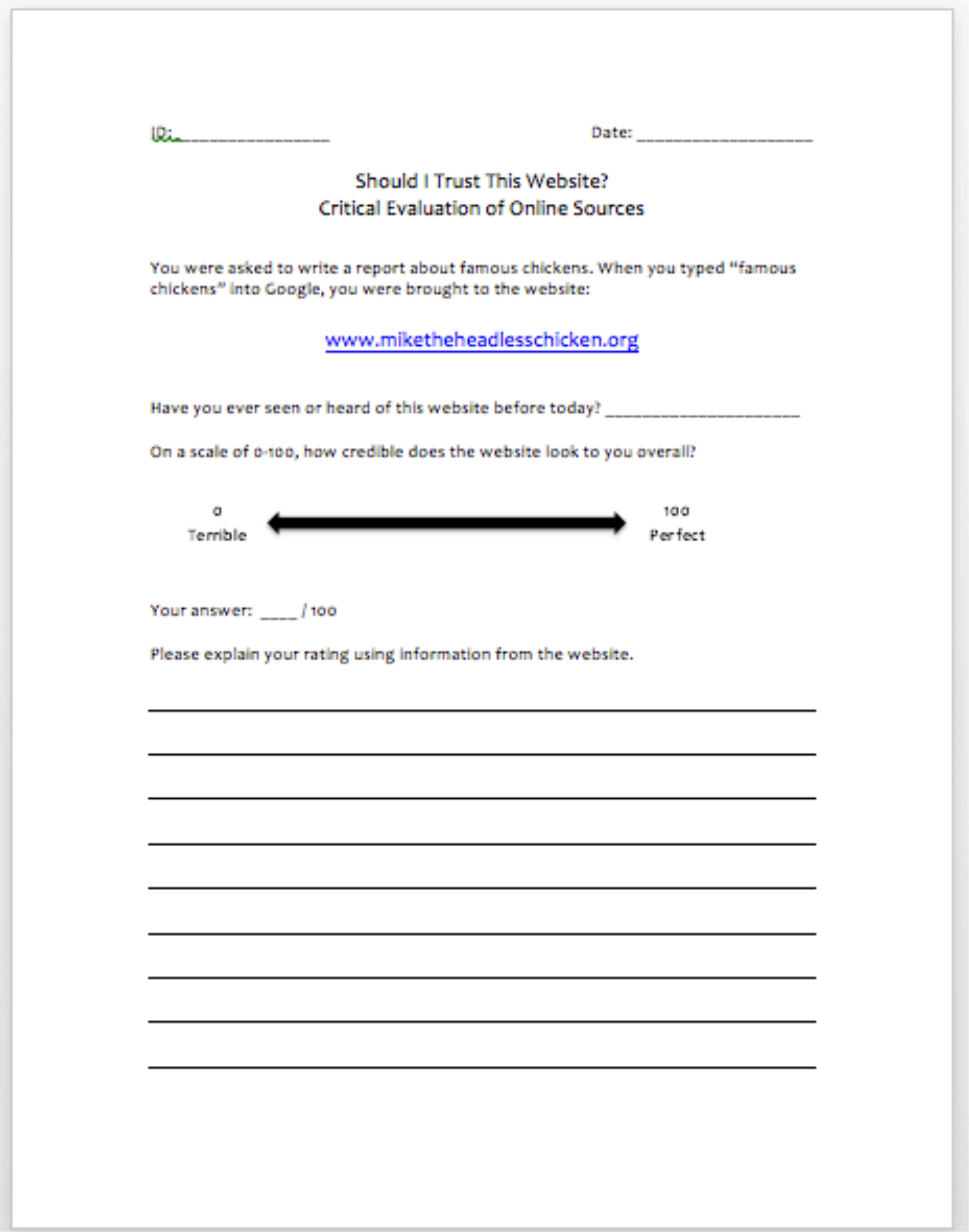
____

____

____

____

Appendix D



ID:__________________ Date: ____________________

Should I Trust This Website?
Critical Evaluation of Online Sources

You were asked to write a report about famous chickens. When you typed "famous chickens" into Google, you were brought to the website:

www.miketheheadlesschicken.org

Have you ever seen or heard of this website before today? ______________________

On a scale of 0-100, how credible does the website look to you overall?

0
Terrible
100
Perfect

Your answer: ____ / 100

Please explain your rating using information from the website.

____

____

____

____

____

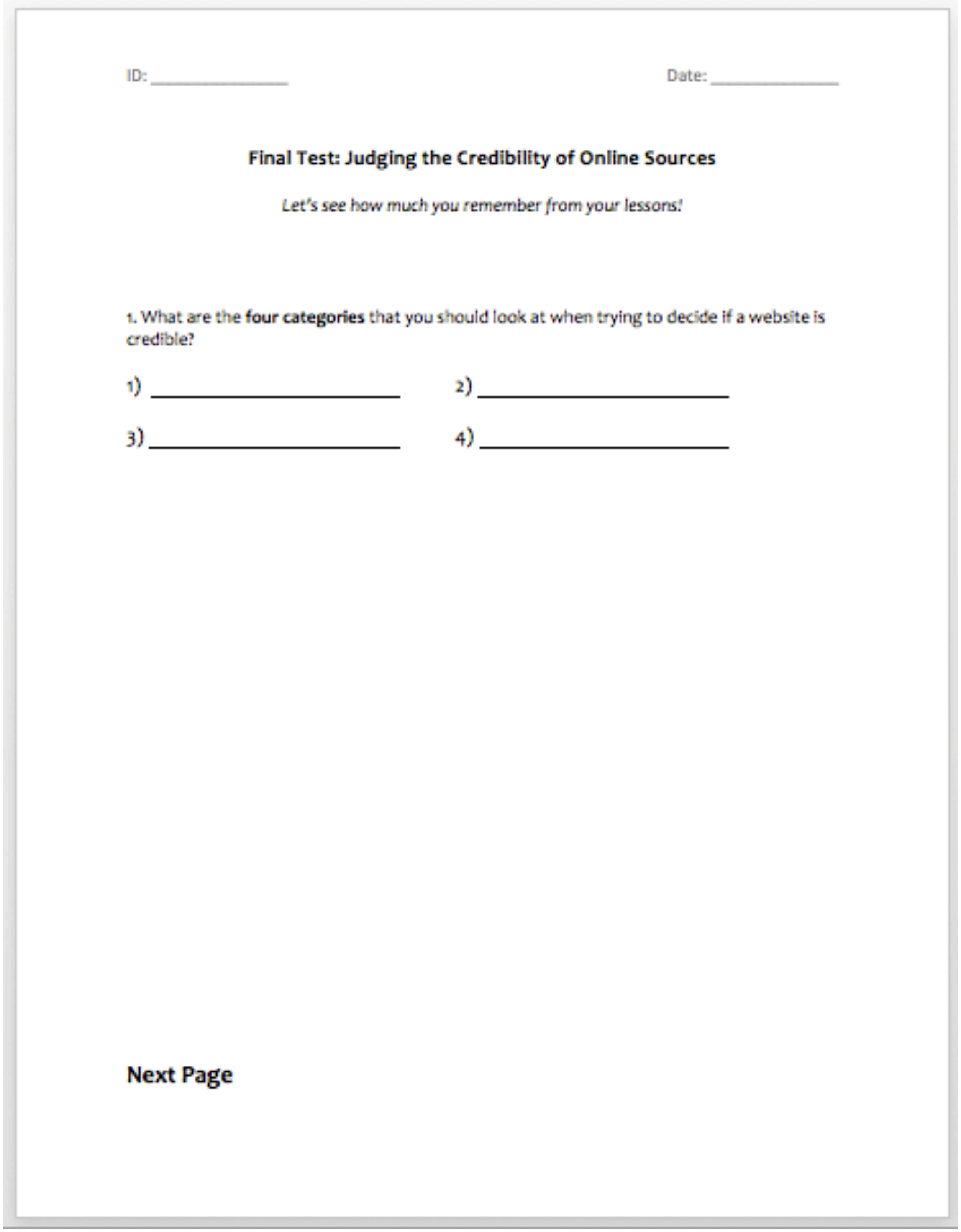
____

____

____

____

Appendix E



ID: ______________   Date: ______________

Final Test: Judging the Credibility of Online Sources

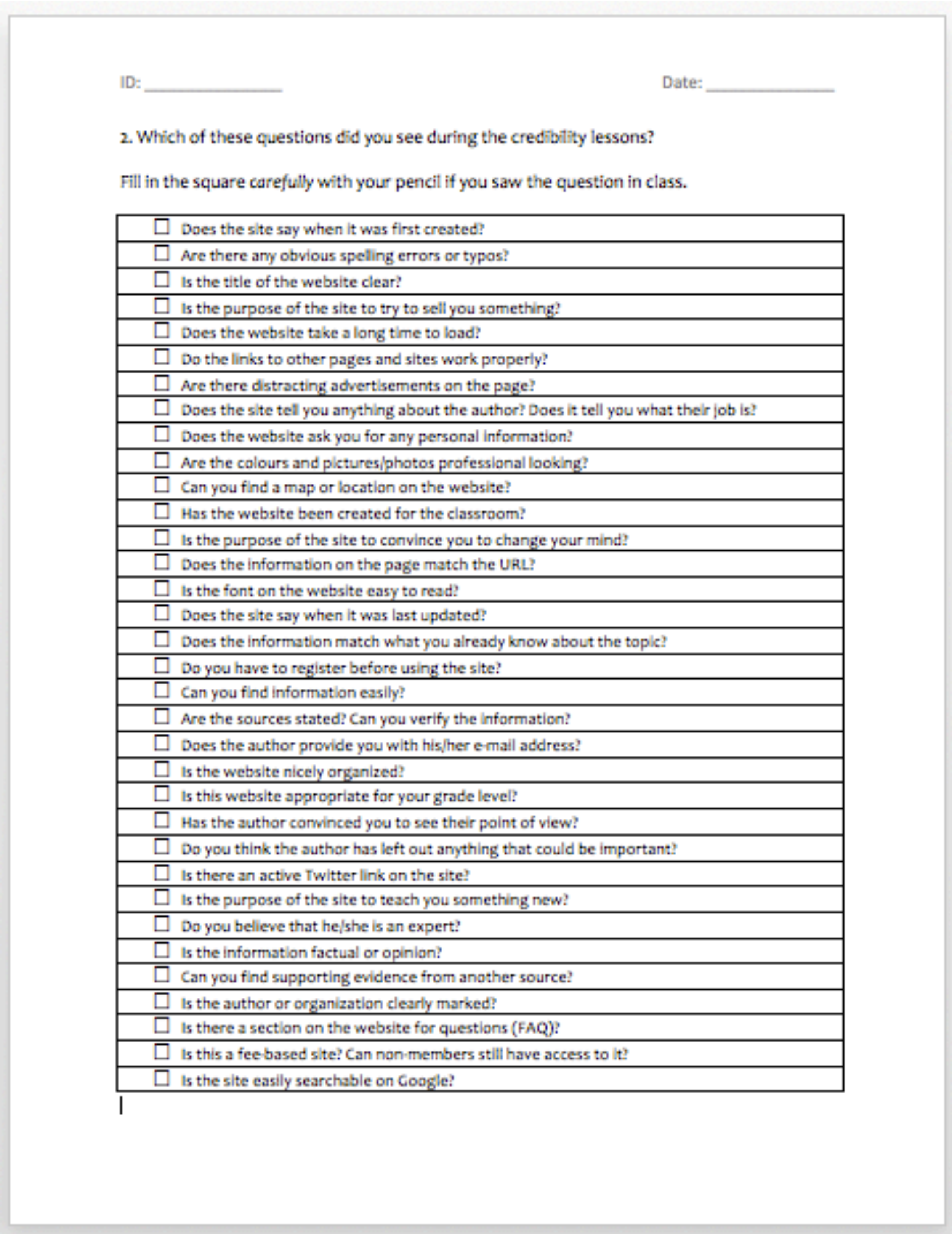
Let's see how much you remember from your lessons!

1. What are the **four categories** that you should look at when trying to decide if a website is credible?

1) ________________________   2) ________________________

3) ________________________   4) ________________________

Next Page



ID: ________________ Date: _______________

2. Which of these questions did you see during the credibility lessons?

Fill in the square *carefully* with your pencil if you saw the question in class.

□ Does the site say when it was first created?

□ Are there any obvious spelling errors or typos?

□ Is the title of the website clear?

□ Is the purpose of the site to try to sell you something?

□ Does the website take a long time to load?

□ Do the links to other pages and sites work properly?

□ Are there distracting advertisements on the page?

□ Does the site tell you anything about the author? Does it tell you what their job is?

□ Does the website ask you for any personal information?

□ Are the colours and pictures/photos professional looking?

□ Can you find a map or location on the website?

□ Has the website been created for the classroom?

□ Is the purpose of the site to convince you to change your mind?

□ Does the information on the page match the URL?

□ Is the font on the website easy to read?

□ Does the site say when it was last updated?

□ Does the information match what you already know about the topic?

□ Do you have to register before using the site?

□ Can you find information easily?

□ Are the sources stated? Can you verify the information?

□ Does the author provide you with his/her e-mail address?

□ Is the website nicely organized?

□ Is this website appropriate for your grade level?

□ Has the author convinced you to see their point of view?

□ Do you think the author has left out anything that could be important?

□ Is there an active Twitter link on the site?

□ Is the purpose of the site to teach you something new?

□ Do you believe that he/she is an expert?

□ Is the information factual or opinion?

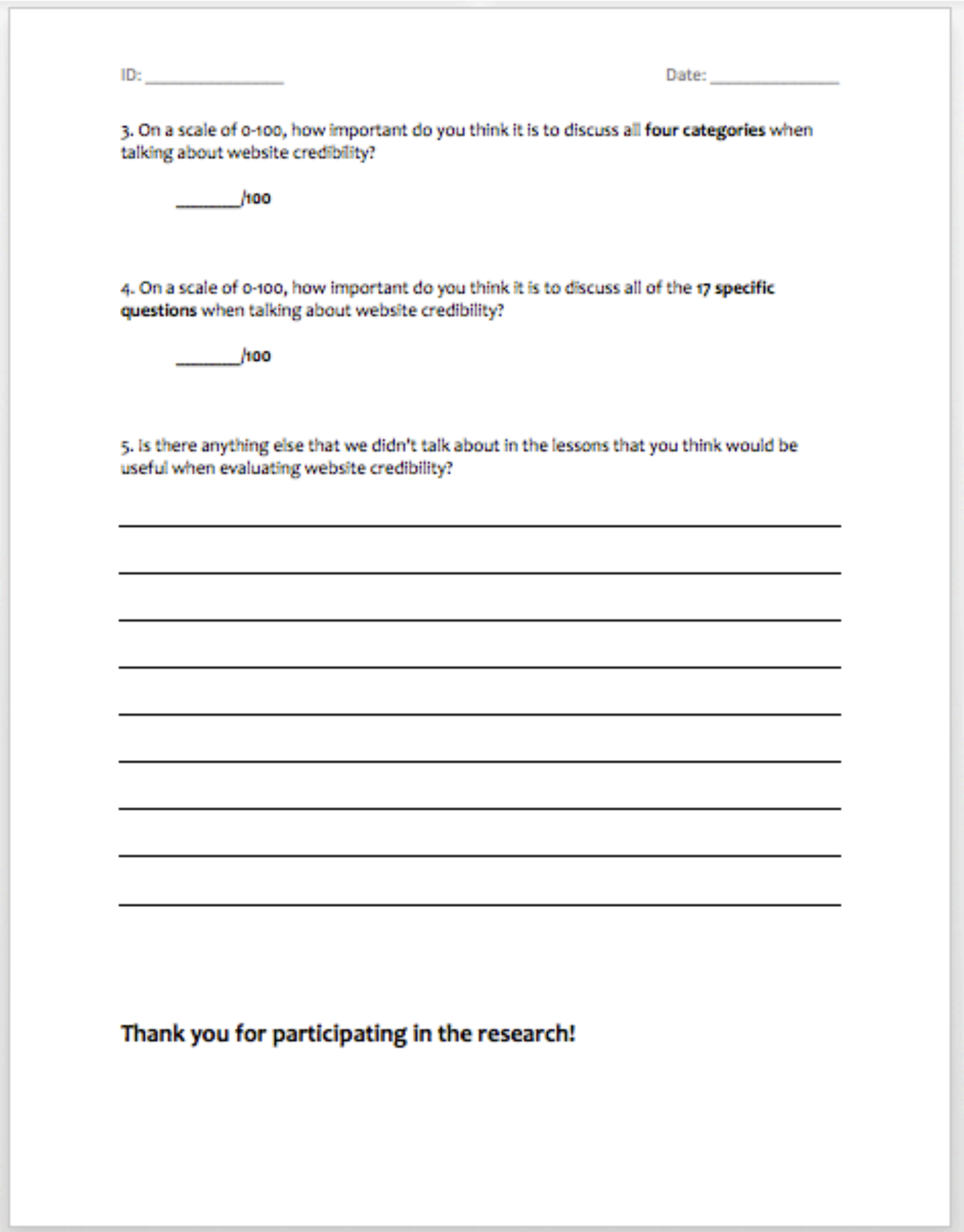
□ Can you find supporting evidence from another source?

□ Is the author or organization clearly marked?

□ Is there a section on the website for questions (FAQ)?

□ Is this a fee-based site? Can non-members still have access to it?

□ Is the site easily searchable on Google?



ID: ________________ Date: ________________

3. On a scale of 0-100, how important do you think it is to discuss all **four categories** when talking about website credibility?

________/100

4. On a scale of 0-100, how important do you think it is to discuss all of the **17 specific questions** when talking about website credibility?

________/100

5. Is there anything else that we didn't talk about in the lessons that you think would be useful when evaluating website credibility?

____

____

____

____

____

____

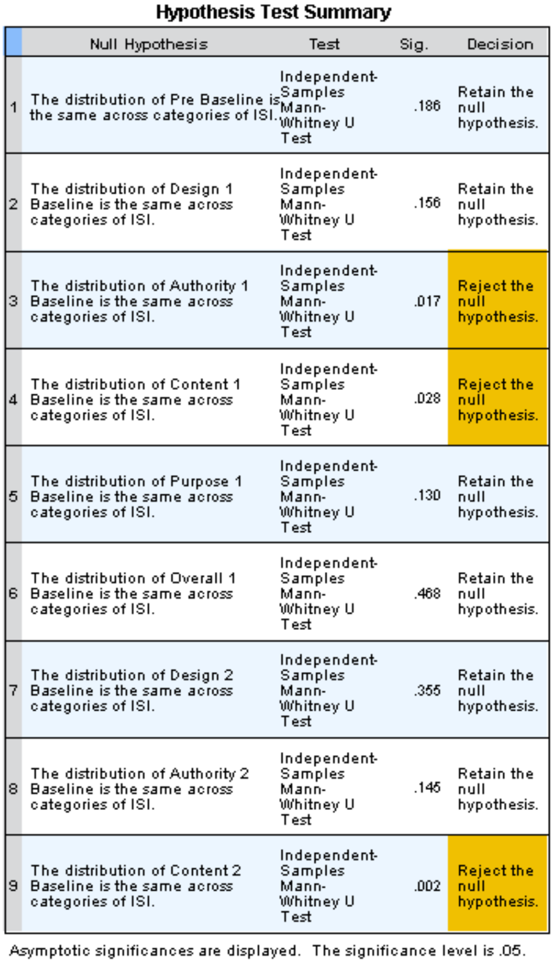
____

____

____

Thank you for participating in the research!

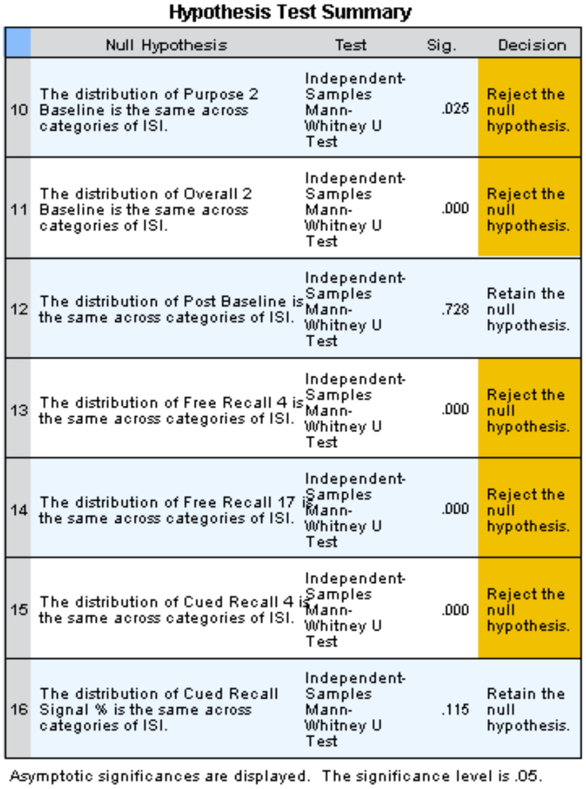
Appendix F



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Pre Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.186	Retain the null hypothesis.
2	The distribution of Design 1 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.156	Retain the null hypothesis.
3	The distribution of Authority 1 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.017	Reject the null hypothesis.
4	The distribution of Content 1 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.028	Reject the null hypothesis.
5	The distribution of Purpose 1 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.130	Retain the null hypothesis.
6	The distribution of Overall 1 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.468	Retain the null hypothesis.
7	The distribution of Design 2 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.355	Retain the null hypothesis.
8	The distribution of Authority 2 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.145	Retain the null hypothesis.
9	The distribution of Content 2 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.002	Reject the null hypothesis.

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



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of Purpose 2 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.025	Reject the null hypothesis.
11	The distribution of Overall 2 Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
12	The distribution of Post Baseline is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.728	Retain the null hypothesis.
13	The distribution of Free Recall 4 is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
14	The distribution of Free Recall 17 is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
15	The distribution of Cued Recall 4 is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
16	The distribution of Cued Recall Signal % is the same across categories of ISI.	Independent-Samples Mann-Whitney U Test	.115	Retain the null hypothesis.

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