Journal of Articles in Support of the Null Hypothesis
Vol. 15, No. 1
 1539-8714
www.jasnh.com

Effects of Spacing and Testing on Inductive Learning

Chanda S. Murphy
Philip I. Pavlik, Jr.
The University of Memphis

The current study aimed to replicate the results of previous studies examining the spacing and testing effect by showing a benefit of spaced practice and repeated testing on inductive learning. Seventy-four participants practiced diagnosing 36 case studies of six psychological disorders and tested in a posttest phase. Although learning occurred, there were no significant differences found in posttest scores between the stimuli that were practiced in a massed versus spaced format. There were also no differences found in posttest scores between stimuli that were practiced as study versus testing trials. The results of the current study necessitate a discussion about how spacing and testing can be most effective and if the effectiveness is conditional on the material being studied.

Keywords: spacing effect, testing effect, inductive learning, retention, study habits

Author Note: Chanda S. Murphy, Department of Psychology, The University of Memphis; Philip I. Pavlik Jr., Department of Psychology, The University of Memphis. Experimental Psychology: Optimal Learning. Correspondence concerning this article should be addressed to Chanda S. Murphy, Department of Psychology, The University of Memphis, 400 Innovation Drive, Memphis, TN 38152. Email: chanda.murphy@gmail.com

Many academics who teach complex and related concepts struggle to help students retain information, and previous research on the topic of learning and memory fails to provide efficient methods for teachers, learners and curriculum designers (Rohrer & Pashler, 2010). However, research on concrete strategies for improving learning and retention has gained momentum in the past few years, specifically in two areas: spacing and its effects on learning (Carvalho & Goldstone, 2012; Kornmeier, McLean, Burt, & Bath, 2012; Spitzer & Sosic-Vasic, 2014; Wahlhein, Dunlosky, & Jacoby, 2011; Zulkiply & Burt, 2013) and testing and its effects on learning (Karpicke & Roediger, 2010; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006; Rowland & DeLosh, 2014; Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014).

Spaced versus Massed Practice

Massed study is defined as any study of a topic without interruption or intervening items (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). An often used example of massed study in academia is cramming for a test or, in general, reviewing material with short or no delays between repetitions. In contrast, spaced study refers to distributed practice in which a measurable amount of time or differing items are interjected into the study (Cepeda et al., 2006). An example of spaced practice would be breaking up study over a period of days or weeks leading up to a test or, in general, having long delays between repetitions of the material.

The study of massed versus spaced practice started as early as the 1800's in association with memory and retention (Ebbinghaus, 1885/1964). Ebbinghaus (1885/1964) found that distributing practice over a span of time provided for better retention in learning a series of syllables. Since then thousands of studies on the spacing effect have been conducted and continue to be conducted in both modern cognitive and educational literature. These previous studies examined a range of stimuli from verbal memory tasks, such as list recall and paired associates (Cull, 2000; Janiszewski, Noel, & Sawyer, 2003; Kornmeier, Spitzer, & Sosic-Vasic, 2014; Pavlik & Anderson, 2005), text comprehension (Reder & Anderson, 1982), and categorical assignment of items (Kornell & Bjork, 2008; Wahlheim, Dunlosky, & Jacoby, 2011; Zulkiply, McLean, Burt, & Bath, 2012). Previous research also focuses on a number of spacing effect variables including interleaving (Carvalho & Goldstone, 2012; Wahlheim, Dulosky, & Jacoby, 2011; Zulkiply & Burt, 2012), embellishment (Reder & Anderson, 1982), the interval of the spaced time (Cull, 2000), age (Kornell, Castel, Eich, & Bjork, 2010) inductive learning (Kornell & Bjork, 2008; Zulkiply et al., 2012) and the testing effect (Cull, 2000; Kornmeier, Spitzer, & Sosic-Vasic, 2014).

Despite all the research that has been done since Ebbinhaus (1885/1964) supporting spaced study over massed study, there is still a disconnect between what is being done in the laboratory and what is being applied in the classroom. In a 1988 article, Dempster suggests this failure stems from the lack of alignment between conditions studied in the laboratory and conditions in a classroom. For example, most of the applied studies on the spacing effect focus on simple tasks like text recall (Dempster, 1986) or vocabulary learning (Dempster, 1987b), whereas classrooms usually require more complex learning, and it is not clear whether beneficial effects of spaced study can be extrapolated to complex learning (Dempster, 1988). Similarly, Pashler, Rohrer, Cepeda, and Carpenter (2007) note that many studies have shown benefits of spacing on learning using vocabulary word tests and math problems. However, they were unable to show similar results when examining the spacing

effect on inductive learning (i.e. checkerboard patterns, dermatological diagnoses). They also conclude that more parallels are required between laboratory variables and classroom conditions and content. Like Dempster (1988), Rohrer and Pashler (2010) argue that benefits seen using limited study variables, like vocabulary learning (Bahrick et al., 1993) and fact or text recall (Carpenter et al., 2009), cannot be generalized to more complex classroom learning. These reviews by Dempster (1988), Pashler et al. (2007) and Rohrer and Pashler (2010) highlight the need to study more complex and applicable stimuli, e.g. categorical assignment or problem solving, in order to establish a better connection between research findings and classroom application.

A study by Kornell and Bjork (2008) was one of the first to test stimuli that better bridged the gap from the lab to the classroom. This paper introduced a new paradigm that showed how spacing affects inductive learning. In contrast to previous research with spacing, they hypothesized that massed practice of category examples is more effective than spaced practice due to massed practice, allowing commonalities to be more easily drawn between concepts and categories. Kornell and Bjork's study required the assignment of paintings to the appropriate artist and included both a practice and testing phase. In the practice phase, paintings were randomly assigned to a massed or spaced presentation, and participants reviewed the painting with the artist's name displayed. In the testing phase, new paintings by the same artists were presented, and participants needed to recognize the correct artist's name from multiple choices. With this inductive learning design, Kornell and Bjork discovered, in contrast to their hypothesis, that spaced practice of examples from a category results in better posttest performance than massed practice.

In an effort to support and generalize Kornell and Bjork's (2008) findings, Zulkiply et al. (2012) replicated the aforementioned study but used case studies of psychological disorders as the categorical stimuli instead of paintings. The use of text-based stimuli by Zulkiply et al. is a notable contribution to the spaced versus massed practice literature due to the educational relevance of text in most academic settings. Zulkiply et al. modeled the design of Kornell and Bjork (2008) in the practice portion by presenting three case studies for each of six psychological disorders in either spaced or massed presentation. In this practice phase, the participant reviewed the correct diagnosis presented on the screen with the case study. The test phase presented unseen case studies where the participant must correctly choose from the same six psychological disorders. To control for prior knowledge, Zulkiply et al. (2012) used novel labels for the disorder names, e.g. Duv for Obsessive Compulsive Disorder, Tem for Schizophrenia, Baj for Phobia Disorder, Pliq for Attention Deficit Disorder (Inattentive type), Hix for Attention Deficit Disorder (Hyperactive and Impulsive type) and Wos for Depression. Zulkiply et al. (2012) replicated the findings of Kornell and Bjork (2008) and similarly conclude that inductive learning benefits from spaced practice.

By testing college students with stimuli they would normally be learning in a classroom, the Zulkiply et al. (2012) study better bridges the gap between laboratory conclusions and classroom applications. The current study replicates elements of Zulkiply et al.'s design and attempts to support their findings. In particular, one limitation to Zulkiply et al.'s findings that the current study attempts to correct is the use of novel names for each of the disorder categories. Because there is a need to bridge the gap between the lab and the classroom, using real names for the disorders should make the results more applicable. The current study uses the actual names for the disorders to increase applicability to the classroom. As we will see in the discussion, the use of the disorder names allows us to better

examine if the spacing effect results are related to the inductive learning or possibly other variables.

Testing Effect

In addition to the spacing effect, the testing effect has also been shown to play an important role in learning and retention and is often used in conjunction with spaced practice (Carpenter, Pashler, Wixted, & Vul, 2008; Karpicke & Roediger, 2010; Pavlik & Anderson, 2005). The testing effect describes the improved retention of material when a student is tested multiple times in comparison to passive study in which a student reviews or re-reads the same material multiple times. Contrary to the benefit found from the testing effect, many college courses include very little testing based practice as part of the requirements. In some cases, courses only incorporate a comprehensive midterm and final, which are not focused on learning benefits but rather assessment. Based on previous research in the classroom by Carpenter, Pashler, and Cepeda (2009), this lack of tests could be hindering students' learning and retention. Thus, similar to the spacing effect, the lack of implementation of the testing effect plagues educational settings even though research has shown beneficial outcomes for this technique.

The testing effect was first studied in 1922 by Gates in what he called a recitation task. Gates (1922) found that when children were given a list of nonsense syllables and asked to recite them at differing intervals, the recitation significantly improved the students' ability to recall all the syllables in a final test. Many laboratory studies have been conducted since Gates research using differing variables such as text comprehension (Spitzer, 1939), word list recall (Hogan & Kintsch, 1971; Thompson, Wenger, & Bartling, 1978; Tulving, 1967) and paired associates (Allen, Mahler, & Estes, 1969; Estes, 1960; Landauer & Bjork, 1978; Pavlik, 2007). All of these findings, no matter the variables, support the same results; tests promote better retention than do additional study trials, and multiple tests further increase performance.

Roediger and Karpicke (2006a) point out in a review that most of the early studies on the testing effect used materials that are not very applicable to the educational setting. They define materials used in previous testing effect research as cognitive stimuli, e.g. word pairs and word lists. To remedy this gap, Roediger and Karpicke (2006b) replicated previous studies but replaced the cognitive stimuli with educational stimuli (i.e. prose). In their study, participants read a prose passage and then either took a free recall test or restudied the entire passage. The participants then completed a free recall posttest after varying delay intervals (i.e. five minutes, two days and one week). Roediger and Karpicke found that testing leads to a significant increase in retention after the two-day and one-week interval conditions. However, in the five-minute condition, repeated studying shows a benefit. Thompson, Wenger, and Bartling (1978) find similar results in their word list recall study. After a 20-minute delay, groups that were tested during practice perform only slightly better on posttest than groups that were given repeated presentations of the words during practice. However, the groups that were tested perform significantly better on the posttest that is given after a 48-hour delay. These studies by Roediger and Karpicke and by Thompson et al. show that the retention interval plays a key role in the impact of the testing effect.

In addition to the lab research, there have been classroom experiments that explore

the testing effect as well. Carpenter et al. (2008) studied the retention of academic facts over a nine-month time period in an eighth grade history class. Some of the facts taught were reviewed by testing, whereas other facts were reviewed by re-studying. At the end of nine months, students were given a final test covering all the facts. Consistent with previous testing effect research, Carpenter et al. found significantly better retention for the facts that were tested than the facts that were restudied. Leeming (2002) found similar results in the classroom setting. Students enrolled in introductory psychology classes were given either short exams at the beginning of every class or four larger exams throughout the course. All students were given a retention test at the end of the semester. The students that were tested daily throughout the course had better performance on the final retention test than the students that only took the four exams. Many of the previous studies examining the testing effect only found results after long term retention delays; however, Pavlik (2007) found significant results for the testing effect at both the short term (after two trials) and long term (after 60 trials) retention intervals. In the current study, we manipulate the amount of testing for each category, in addition to manipulating the spacing effect. Our goal is to determine if there is an advantage of testing when compared to re-studying or passive study. Based on Pavlik's findings, the current study also aims to show a testing effect when only a short-term retention interval is required.

Current study

A primary purpose of the current study was to replicate and support the spacing effect results of Zulkiply et al. (2012). Like Zulkiply et al., the current study also uses categorical stimuli which asks the participants to study symptoms of psychological disorders in order to identify the disorders. Similar to Zulkiply et al. and other spacing effect literature, the following is hypothesized:

> *H1:* The stimuli presented with spacing between repetitions will have higher posttest scores than the stimuli presented massed.

To make the study even more applicable to classroom recommendations, the current study also manipulates the amount of testing during the practice phase. Previous research has shown that including tests with studying of material helps improve one's memory for the material, thereby increasing retention (McDaniel et al., 2007; McDaniel, Roediger, & McDermott, 2007; Pavlik, 2007; Roediger & Karpicke, 2006b). The current study also replicates the educational applicability of Roediger and Karpicke (2006b) by using prose in an attempt to confirm test-based learning advantages in category induction. Further, the current study attempts to replicate the findings of Spitzer's (1939) and Karpicke and Roediger's (2010) by showing that multiple tests can improve learning and retention. Therefore, the following is hypothesized:

> *H2:* The stimuli that are tested in the study phase will have higher posttest scores than the stimuli that are only read.

> *H3:* The stimuli in the two-test condition in the study phase will have higher posttest scores than the stimuli in the one test or no test conditions.

Method

Participants

Seventy-four undergraduates from introductory psychology courses at a small, private university in the mid-south participated voluntarily for extra credit in a course. The average age of the participants was 19 with 44.6% male and 55.4% female. The majority (75.7%) of the participants were in their freshman year of college with the remaining 9.5% being sophomores, 8.1% juniors, and 6.8% seniors.

Design

Replicating Zulkiply et al. (2012), the study was a within subjects design and included the participants completing a practice phase, a distracter task, a posttest phase and a final survey. All tasks were conducted in a computer lab. An element of the study that differed from Zulkiply et al. was the prior knowledge assessment. Zulkiply et al. used novel names to limit the effect prior knowledge would have on the spacing effect results to get a more clear theoretical result; in contrast, we used the actual names to get a more educationally applicable result. Since the current study used the actual names of the disorders, there was concern that prior knowledge of the disorders could influence the results. Therefore we administered a prior knowledge assessment one week prior to the online portion of the experiment to use as a covariate when analyzing the results.

The practice phase consisted of three case studies for each of six different psychological disorders, totaling 18 case study practices. The case study practice for each participant was presented in a two (spacing or massed) by three (0 test, 1 test, 2 tests) design. During the practice phase participants saw one of two sequences for spaced practice, either MSMSMS or SMSMSM (M representing massing 3 trials for a category; S representing spacing 3 items from different categories) to control for ordering effects. For example, one condition may have consisted of the first three case studies having the diagnosis of anxiety (i.e. massed presentation), followed by three case studies in the spaced condition with the diagnoses of schizophrenia, bipolar and depression. This spaced condition would be followed by three case studies presented in the massed format with the diagnosis of OCD. The spaced condition disorders were then repeated in random order: schizophrenia, bipolar and depression. The final massed disorder category was then presented; in our example this would be three dissociative identity disorder items. The spaced diagnoses were then presented for a final time in random order, totaling 18 case studies in the practice phase. For the testing effect manipulation, one disorder in both the massed and spaced condition was randomly assigned to the no test condition in which all three trials would be study trials. One disorder in both the massed and spaced condition was randomly assigned to the one test condition in which the final case study of the disorder would be a test trial. Finally, one disorder in both the mass and spaced condition was randomly assigned to the two test conditions in which the final two case studies of the disorder would be test trials.

The posttest phase included 18 randomly assigned case studies, once again including three case studies per psychological disorder. The case studies were divided among three test blocks with one case study from each disorder represented in each block. Rather than assigning multiple forms of the practice and posttest, a completely random subset of the

total 36 case studies was assigned to the practice and posttest conditions for each participant.

The final survey consisted of a Likert scale inquiry of whether the participants felt massed or spaced study had greater effect on their learning as well as demographic questions such as age, sex and race.

Materials

The materials included a prior knowledge assessment, 36 case studies developed and adapted from different abnormal psychology sources and a vocabulary distractor task. The 36 case studies consisted of six case studies of six different psychological disorders (generalized anxiety, depression, obsessive-compulsive, schizophrenia, bipolar and dissociative identity disorder). The prior knowledge assessment included 24 multiple choice questions pertaining to the six disorders being assessed in the current study. Two examples of questions are as follows: 1.) A generalized anxiety disorder is characterized by: a.) offensive and unwanted thoughts that persistently preoccupy a person. b.) a continuous state of tension, apprehension, and autonomic nervous system arousal. c.) hyperactive, wildly optimistic states of emotion. d.) alternations between extreme hopelessness and unrealistic optimism. 2.) Sluggishness and inactivity are most likely to be associated with: a.) antisocial personality disorder, b.) major depressive disorder, c.) obsessive-compulsive disorder, d.) dissociative identity disorder. To avoid drawing too much attention to the six disorders that would be used later in the study, other unrelated general psychology questions were also included, i.e. 1.) According to Bandura, reciprocal determinism involves multidirectional influences among: a.) thoughts, emotions and actions, b.) behaviors, internal personal factors and environmental events, c.) id, ego and superego, d.) self-concept, self-actualization, and self-transcendence. 2.) Psychodynamic theories emphasize that personality involves a dynamic interaction between: a.) persons and situations, b.) conditioning and observational learning, c.) conscious and unconscious mental processes, d.) unconditional positive regard and self-actualization (Myers, 2008).

Each of the 36 case studies was approximately between 100 and 120 words in length and included descriptions of symptoms related to the disorder being described (see Appendix A). The names of the six disorders described in the case studies were also presented as a cue card on screen. Each disorder on the cue card had an associated abbreviation (i.e. Obsessive Compulsive Disorder OCD) the participant could use when typing in their diagnosis. The case studies of the disorders were randomly assigned by the FaCT system (Pavlik, Presson, Dozzi, MacWhinney, & Koedinger, 2007), a computer software program, to each condition as well as randomly assigned to the study phase and posttest phase.

The distractor task between learning and testing phases consisted of 15 vocabulary multiple-choice questions in which the participants were asked to find the best definition of words such as perjure, illusory, reprove, etc.

Procedure

The study, with exception of the prior knowledge assessment, was presented in its entirety online through the FaCT system. One week prior to the online portion of the study, participants completed the prior knowledge survey in class. Approximately one week later, participants were then tested in groups of 20 in a school computer lab setting. The practice

phase presented 18 of the case studies, and the participants were asked to read and study the cases. For the disorder categories in the no test condition, each case of the disorder was presented on the screen with the label of the disorder displayed underneath for a total of 40 seconds. In the one test condition, the first two cases of each disorder were presented the same as the no test condition; however, the final case study for each disorder was presented on the screen and the participant filled in a blank as to what disorder was being described. If the participant took too long to answer, the screen would advance after 40 seconds. In the two-test condition, only the first case of each condition was presented the same as the no test condition, and in each subsequent case the participant had to identify the disorder. The participants were given a cue card with abbreviations they could use when typing in the names of the disorders. (In contrast, Zulkiply had the students use labeled buttons with novel terms for responding with no key to the actual meanings of the novel labels). Also in the testing conditions, the participant was given feedback that they were correct or incorrect. If the participant was incorrect, he was given a 10-second review period with the correct answer displayed on the screen. Once the 18 case studies were reviewed, the participants were asked to complete a distracter task in which they answered 15 multiple choice English vocabulary questions for the purpose of clearing working memory.

The posttest phase began after the distracter task. Participants were shown the remaining 18 case studies they had not already read. Identical to the testing in the practice phase, the participants were shown one case study at a time on the computer screen with a blank entry box underneath it for the participant to type in the disorder name. Feedback, identical to the practice phase, was given for each response.

After the posttest phase, the participants were given a description of both the terms massed and spaced and asked to rate which method was more effective in their learning on a seven point Likert type scale. The entirety of the experiment, with the exception of the prior knowledge survey, took approximately 40 minutes.

Results

A repeated measures ANOVA was conducted on the data using the prior knowledge score as a covariate. Contrary to previous studies, the findings of the current study did not support the spacing effect or the testing effect. There were no significant differences in performance between massed and spaced study, $F(1, 72) = .02$, $p = .89$, (massed study ($M = .73$, 95% CI [.69, .77]), spaced study ($M = .71$, 95% CI [.67, .75])). See Figure 1 for the overall performance averages of the first trial at posttest in the current study as compared to Zulkiply et al. (2012).

To look at the non-significant result in more detail, we computed the confidence interval for the effect given the pooled standard error of the difference between massed and spaced scores. Our 95% confidence interval ($SE = .027$, 95% CI [-.066,.043]) showed that with 95% confidence the true mean effect will fall between a 4.3% spacing effect and a 6.6% massing effect. Given the large effects in Zulkiply, this narrow range of confidence intervals with low significant benefit for spaced practice provides strong support for the null effect or a much weaker effect than Zulkiply produced. As might be expected, the current study's effect size was small ($\eta_p^2 = .0033$, not significant) and in favor of massing, whereas Zulkiply et al. (2012) had a large effect size of $\eta_p^2 = .52$ and found a significant spacing condition advantage. See Figure 2 for a comparison across posttest trials between the current study and Zulkiply et al. (2012).

There were also no significant differences among the testing conditions, $F(1,72) = 1.6$, $p = .20$. Because the confidence intervals almost entirely overlap, there was also no meaningful trend in the means among the three testing conditions (no test condition ($M = .71$, 95% CI [.66, .77]), one test ($M = .71$, 95% CI [.66, .76]) and two test ($M = .73$, 95% CI [.68, .78])). However, when comparing the means on performance of the test trials during study ($M = .62$) and the posttest trials ($M = .72$), there was a significant increase in performance across both conditions, $t(73) = 3.27$, $p = .0012$ so it can be concluded that learning did occur from learning session to posttest. The posttest scores were also analyzed to ensure there was no ceiling effect occurring. There was a significant distribution in the means of the posttest scores suggesting there was no ceiling effect, $t(73) = 44.10$, $p < .0005$ ($M = .72$, 95% CI [.69, .75]).

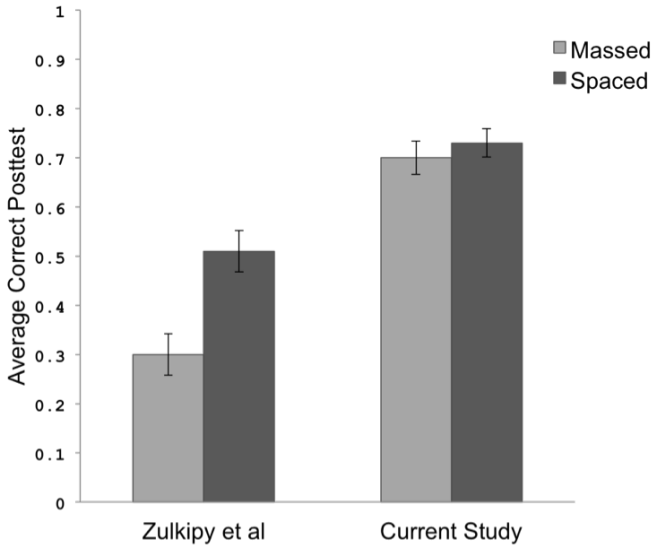

Figure 1. *Comparison of probability correct of the first trial at posttest between massed and spaced performance in both the current study and Zulkiply et al. (2012). Error bars for the current study represent one standard*

The data were also analyzed to examine whether the spacing effect may have had greater impact with either high or low performers. A median split was calculated on the posttest scores with the high performance group having a mean performance above .71 and the lower performing group with a mean performance below .71. There was no spacing effect found in the high performing group, $F(1, 43) = 1.76$, $p = .19$, with the massed condition averaging .83 (95% CI [.80, .87]) and the spaced condition averaging .78 (95% CI [.74, .82]). There was also no spacing effect found in the low performing group, $F(1, 27) = .76$, $p = .39$, with the massed condition averaging .57 ($M = .57$, 95% CI [.52, .61]) and the spaced condition averaging .59 (95% CI [.52, .66]). There was also no significant interaction between the median split and the spacing effect, $F(1, 72) = 1.93$, $p = .17$. A median split was also calculated on the prior knowledge scores with the high prior knowledge group having a mean above .43 ($N = 40$) and the lower prior knowledge group with a mean below .43 ($N = 34$). There was no spacing effect found in the high prior knowledge group, $F(1, 39) = 04$, $p = .85$, with the massed condition averaging .74 (95% CI [.69, .80]) and the spaced condition averaging .74, (95% CI [.69, .79]). There was also no

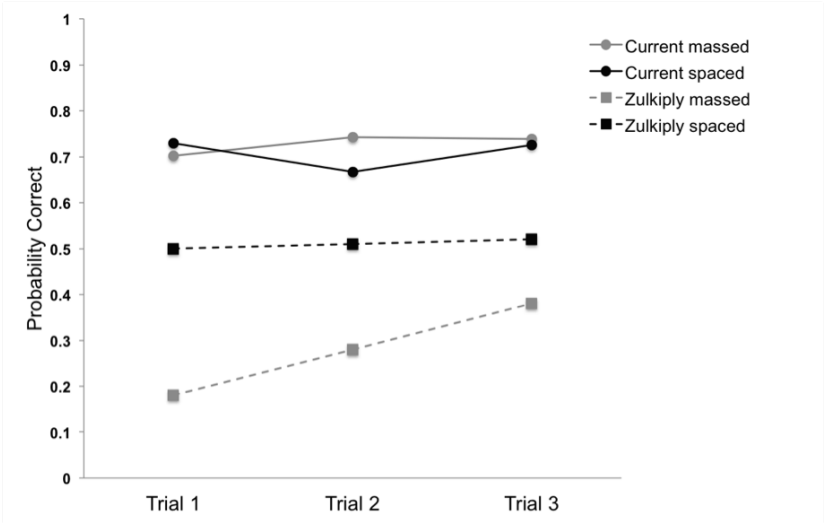

Figure 2. *A comparison of probability correct across post test trials in both massed and spaced presentation between the current study and Zulkiply et al. (2012).*

spacing effect found in the low prior knowledge group, $F(1, 33) = .76$, $p = .34$, with the massed condition averaging .71 (95% CI [.64, .78]) and the spaced condition averaging .67 (95% CI [.60, .74]).

The stimuli that were used in the current study were also analyzed to ensure that properties of the stimuli set were not confounding the results. The average performance across cases was 67% ($SD = .19$) (including practice section trial performance). The mean performance of the majority of the 36 cases fell between 60% and 94% (chance performance being 16.67%). The performance of eight cases fell below 60%. An argument for why these eight cases fell below 60% can be related to the common misdiagnoses that happen due to overlapping symptoms among the disorders. Three of these cases that fell below the average performance belonged to the category of bipolar disorder ($M = .48$, $SD = .22$). Due to the overlapping symptoms of bipolar with symptoms of anxiety and depression, participants misdiagnosed these cases as either anxiety or depression about 30% of the time. Two of these cases belonged to the dissociative identity disorder category. These two cases of dissociative identity disorder were misdiagnosed as schizophrenia 30% of the time. One of the cases belonged to the obsessive compulsive disorder category. This case was misdiagnosed as anxiety 21% and as depression 16% of the time. The final two cases that fell below average performance belonged to the schizophrenia category. The two cases in this category were misdiagnosed as depression 28% of the time. The range in the performance on the cases shows that there was ample room for learning to occur.

After the completion of the study, an a priori power analysis was computed through G*Power software using the effect size from Zulkiply et al. (2012) and the conservative assumption of no correlation for within-subject values. It was found that a sample of only 18 participants was needed for the spacing effect comparison to achieve .99 power. A sensitivity analysis was also conducted, and it was found that with the sample size of the current study ($N = 74$) we should detect an effect size of $\eta^2 = .03$ with .9 power for the spacing effect comparison, further supporting the accuracy of the results of this study.

Discussion

Based on our results, we fail to reject the null hypothesis. While we cannot accept the null hypothesis, the results of the power and sensitivity analysis, which come from a sample larger than Zulkiply used, suggests that the null hypothesis seems very plausible. The confidence intervals we described only allow for a very small effect to have gone undetected in our experiment. Our failure to reject the null hypothesis, in contrast to Zulkiply et al.'s research (2012), is important to the field of learning because it leads us to question the mechanism by which the spacing effects are benefitting learning as reported in prior studies. We are required to ask whether the spacing effect truly helps with inductive learning, or if there exist boundary conditions for the spacing effect as this current study suggests. A first explanation for possible boundaries hinges on the way many prior experiments show inductive spacing effects by using unfamiliar response terms. This hypothesis, which we explore below through examples, claims that the learning of prior unknown response terms benefits from "inductive spacing" as an artifact of the learning of the unfamiliar responses terms rather than the category membership knowledge. A second explanation is less specific and notes that our difficulty level was much lower than previous work in inductive spacing effects. If inductive spacing effects only occur when material is difficult enough, but perhaps

not too difficult, this would also explain our result and limit the applicability of inductive spacing effects to the classroom.

Response Learning Argument

From the perspective of our results, it seems valid to question whether the learning of the conceptual category benefits from spacing, or whether the learning to recognize or produce the category label benefits from spacing. To give a concrete example, does a learner come to understand a particular artist's use of dots (pointillism) to compose larger impressionist scenes better when the examples are spaced, or does the learner come to more easily recognize the name "Georges Seurat" when the repetitions are spaced? This is a difficult question to answer given past research, but becomes very important when the labels are already known as distinct constructs but the categories themselves are not very well known. If "inductive" spacing effects are limited to situations where the labels are poorly known, the educational implications are more limited.

A first example of this issue in a study comes from Kornell and Bjork (2008), who found a spacing effect in their study on learning the classification of artists and their paintings. They found a significant result in the spacing effect with spaced study (M = .61) having a higher performance over massed study (M = .35) and also a large effect size of d = 1.28. However, an issue with interpreting this finding could be in the artist names they were using. This study begs the question of whether the spacing effect resulted from name vocabulary learning rather than from the learning of proper classification of the artist's style. Consider that the artist names chosen were relatively uncommon to those who have not studied art, (i.e. Georges Braque, Henri-Edmond Cross, Judy Hawkins, Philip Juras, Ryan Lewis, Marilyn Mylrea, Bruno Pressani, Ron Schlorff, Georges Seurat, Ciprian Stratulat, George Wexler and Yiemei). It seems plausible to suppose that the main performance increase observed was due to spacing effects contributing to better recognition and discrimination of these previously unfamiliar or unknown names.

A similar example is a study by Birnbaum, Kornell, Bjork and Bjork (2012), in which they found a spacing effect for their testing of recognition discrimination of butterfly species with names such as Admiral, American, Baltimore, Cooper, Eastern Tiger, Hairstreak, Harvester, Mark, Painted Lady, Pine Elfin, Pipevine, Sprite, Tipper, Tree Satyr, Viceroy and Wood Nymph. As with Kornell and Bjork, unless the participant was well versed in butterfly species (an amateur lepidopterist), the spacing effect measured could have been due to the learning of the names of the species rather than the perceptual category. Birnbaum et al. (2012) also had a concern of prior knowledge affecting results, therefore they changed some of the names of the species either to one word, or entirely, if the name of the butterfly described physical characteristics. This seems likely to increase the amount of learning needed for word/name acquisition, making the task even more dependent on verbal learning. Additionally, the effect size of their study was lower (d = .379) in comparison to Kornell and Bjork (2008) and Zulkiply et al. (2012).

In an effort to better understand the inductive spacing effect found by Kornell and Bjork (2008), Walheim, Dunlosky, and Jacoby (2011) studied the learning of bird families. Specifically, Walheim, Dunlosky and Jacoby used bird names such as chickadees, finches, flycatchers, grosbeaks, jays, orioles, sparrows, swallows, thrashers, thrushes, vireos and warblers. Similar to previously mentioned studies, they found a significant spacing effect.

Although some of these names are familiar to many, we think it seems plausible that many college students have no notion of the difference between a chickadee, a finch and a swallow. The current study on the classification of disorders reiterates the question of whether there was actual inductive classification learning benefitting from the spacing effect or simply a learning of the new names of the birds, which would then be replicating previous studies on the spacing effect with vocabulary learning.

In contrast to the typical goals for these inductive paradigms, cognitive psychology work with standard spacing effects has shown how spacing effects are strong and easy to produce even when response terms are known. This fact presumably stems from a beneficial effect of spaced practice on learning the association, stimulus, or both, and not the response term (since it is known). For example, in a study conducted by Pavlik and Anderson (2005), which used active testing to examine the spacing effect in learning English translations of vocabulary (Japanese words in English letters), the response words learned were all common English and a strong spacing effect was found. So, while our experimental response terms were in contrast to prior work with inductive spacing effects they were similar to cases where spacing effects are normally seen, like Pavlik and Anderson (2005), which used known common English responses.

Therefore, we must speculate how our disorder learning task with English labels was somehow different than a classical spacing effect result. Since the response knowledge is ruled out as a primary cause, we might think it is the difference in the learning of the association or concept that drives our lack of result. Perhaps each example of a concept is a variable encoding, creating a unique association, and therefore is not affected by spacing effects. This is an explanation that can be supported by results such as Gartman and Johnson (1972) where different biases (a variability condition) for repeated items negated the spacing effect during learning. This would explain why the repetitions of varying examples in our experiment may not have exhibited spacing effects during learning. Further we propose that learning of the relatively unknown and relatively undifferentiated response terms in prior experiments may have driven their results. There seems no good way to rule out this possible interpretation of the prior results without additional experiments comparing inductive learning conditions with well-known vs. unknown response terms. This is an educationally important question since if spacing effects fade as verbal learning increases, this implies a different approach to using spacing effects in the classroom.

Desirable Difficulty Argument

Another explanation for the findings of the current study is more general and centers on how prior knowledge may block spacing effects by making practice too easy. The overall performance by the participants in this study is relatively high with means ranging from .77 (massed) - .76 (spaced) as compared to previous studies on the spacing effect such as Zulkiply et al. (2012) with performance means ranging from approximately .2 (massed) to .5 (spaced), Wahlheim et al. (2011) with means of .4 - .55 and Kornell and Bjork (2008) with means of .35 (massed) to .6 (spaced). This could be explained by the role prior knowledge plays in these studies. In the previous studies, prior knowledge was mostly ruled out by using names and terminology that participants were not familiar with; however, the current study wanted to also examine the role prior knowledge would play in the spacing effect.

In their review of three different studies, Schmitt and Bjork (1991) support a desirable difficulty argument for practice. Schmitt and Bjork reviewed a study by Shea and Morgan (1979) and two by Landauer and Bjork (1978). All three of these studies presented their variables in both massed and spaced intervals during a practice (or acquisition) phase and also measure learning in a posttest or (retention) phase. Although the three studies used varying stimuli (motor tasks versus verbal tasks), Schmitt and Bjork conclude from these three studies that the similarity in their results were that the tasks that were spaced during practice scored lower in practice however had higher retention scores in the posttest. Schmitt and Bjork argue that the spacing provided a desirable difficulty during acquisition, which caused better retention. They also theorized, based on the results of these studies, that additional difficulty added to the acquisition or learning phase should further enhance retention performance. The current study was not able to replicate these studies in the practice phase with no significant differences between the massed (M = .63) and spaced (M = .62) conditions. Therefore it can be argued that using the actual names of the disorders was not difficult enough overall for the spacing effect to produce better retention in posttest. Therefore, it is possible too much prior knowledge blocked the spacing effect. However, as aforementioned, in the current study there were significant differences in learning session and posttest score suggesting learning was occurring, so the problem here would seem to occur before learning is at ceiling, which suggests spacing effects in this context might not apply to the entire learning function, but rather only for early learning before general proficiency is achieved.

Interleaving Argument

A final possible explanation for the results of the current study relates to the interleaving effect work done by Carvalho and Goldstone (2012). Carvalho and Goldstone conducted two experiments with very similar methods to Kornell and Bjork (2008) but used abstract drawings as stimuli. The methods of the study were the same; however, in the first experiment, the different stimuli drawings had high within category similarity; and in the second experiment, the items had low within category similarity. Through these experiments, Carvalho and Goldstone found that when the drawings were highly similar within category, interleaved study (spaced study) was more effective. They explain how interleaved study assists in learning similar stimuli (i.e. artists or birds) because differences that are hard to detect can be more easily identified when stimuli are interleaved, therefore contrasted, with other stimuli and spaced apart. In contrast, when the drawings had low similarity blocked study (massed study) was more effective. Carvalho and Goldstone state that when similarity is low, showing stimuli in a massed format allows commonalities to be found to make an abstraction. It can be argued that the stimuli used in the current study had lower similarity based on Carvalho and Goldstone's definition. For example, in their study, the abstract drawings in the high similarity group had minor curve differences that were barely noticeable. In contrast, the low similarity drawings had greater differences, such as a curve changed to a straight line or two curves instead of one. In our current study, within each case study and category, we varied the name of the client being diagnosed, the biographical information of the client and the manifestation of the symptoms (i.e. excessive worry versus increased heart and respiration rate). Therefore, the stimuli of the current study could be considered as having low similarity which would support the lack of the

spacing effect. However, because the stimuli we used are analogous to the stimuli used by Zulkiply et al. (2012), this argument about low similarity should also apply to the Zulkilpy (2012) study. Thus, it is unlikely that Carvalho and Goldstone's interleaving results, showing massed study being more effective than spaced study in a low similarity category, can account for the differences we observe between our results and the results of the Zulkiply et al. (2012) study. In Carvalho and Goldstone's discussion, they point out that there may not be one best way for information to be presented for effective retention. They also conclude that it may depend on the content (i.e. bird names, geometric shapes or vocabulary) of the information being studied as to how information should be presented.

Conclusion

The complexities of the results of the current study, as well as the results of the studies mentioned throughout this manuscript, highlight the dangers of a one-size-fits-all prescription for the use and effectiveness of spaced practice. A study by Kost, Carvalho, and Goldstone (2015) provides further evidence for the contradictory and complex results in the domain of conceptual spacing effects. Kost, Carvalho, and Goldstone (2015) looked at how multiple variables influence the inductive spacing effect. Based on their findings, whether or not spacing study is effective is a lot more complicated than previous research has suggested and is dependent on various conditions. Kost, Carvalho, and Goldstone conducted a three part study replicating the same methods and using the same set of artists paintings as Kornell and Bjork (2008); however, they added the variables of repetition and active versus passive study (i.e. testing effect). The first experiment of this three part study was between subjects with one group studying the artists' paintings in a massed format and the other group studying the artists in a spaced format. In both groups the practice phase was repeated twice to add the variable of repetition. This first experiment revealed that with active study (testing) massed practice (blocked) was more effective than spaced practice (interleaved) when there was repetition of the practice phase. Their prediction for this finding, in comparison to previous research, is that the addition of the repetition of practice increased the effectiveness of massed study. To follow up on these results, Kost, Carvalho, and Goldstone (2015) conducted a second experiment replicating the same methods as the first experiment but removing the repetition of practice phase. Interestingly, with the repetition aspect removed, there were no significant differences between massed and spaced study on posttest scores. In an attempt to find the spacing effect that Kornell and Bjork found (2008), Kost, Carvalho, and Goldstone conducted a third experiment of their study in which they completely replicated Kornell and Bjork by only having passive study during the practice phase. In this third experiment, they were able to replicate Kornell and Bjork's results with a significant spacing effect. These results seem to be an anomaly compared to previous research finding that spaced testing usually leads to large spacing effects (e.g. Pavlik & Anderson, 2005). Also, in the current study there was no interaction found between testing and spacing effects; therefore we are left to question why the spacing effect in Kost, Carvalho, and Goldstone's study shows such different results. They conclude, much like Carvalho and Goldstone (2012), that there needs to be further research on how the type of information being learned plays a role in the effectiveness of spaced study.

Kost, Carvalho, and Goldstone's (2015) research, along with the current study and other research that found massed study to be more effective (Carvalho and Goldstone,

2012, Pashler et al., 2007), confirms a need for future research on this topic to examine how various variables such as prior knowledge, active versus passive learning, content being studied (semantic versus categorical learning), level of difficulty and the similarity of the topics being studied may play a role in the effectiveness of the spacing effect for inductive learning. There are best practices currently published and recommendations that have been made to educators that suggest spacing and testing to be the most effective form of study for long term retention (Brown, Roediger, & McDaniel, 2014). However, based on the null results of the current study and the opposing results of other studies (Carvalho & Goldstone, 2012; Kost, Carvalho, & Goldstone, 2015; Pashler et al., 2007), educators should be wary of accepting these overarching suggestions of using the spacing effect in their teaching until more classroom applicable research is done. There is a need to conduct more experiments with natural and applicable inductive learning interventions to better understand when the spacing effect and/or testing effect will be most useful. Research has shown the success of the verbatim repetition spacing effect time and time again in the laboratory; however, the current study shows a need for additional research on how spacing works for tasks and topics closer to what students will be learning in the classroom before the spacing effect can be advocated as the default best practice for all learning situations. If the ideal conditions benefiting massed versus spaced practice can be established through future research, students would then have greater success in real educational settings.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8*(4), 463-470.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316-321.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392-402.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick*. Cambridge, MA: Harvard University Press.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology, 23*(6), 760-771.

Carvalho, P. F., & Goldstone, R. L. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. *Proceedings of the 34th Annual Conference of the Cognitive Science Society, 34,* 186-191.

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*(3), 354.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235.

Dempster, F. N. (1986). Spacing effects in text recall: An extrapolation from the laboratory to the classroom. *Journal of Educational Psychology, 79,* 162-170.

Dempster, F N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology, 79,* 162-170.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43*(8), 627-634.

Ebbinghaus, H. (1885/1964). Memory: A contribution to experimental psychology (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications, Inc.

Estes, W. K. (1960). Learning theory and the new "mental chemistry." *Psychological Review, 67*(4), 207-223.

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetition of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 801-808.

Gates, A. I. (1922). *Recitation as a factor in memorizing* (No. 40). New York: Science Press.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*(5), 562-567.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30*(1), 138-149.

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38*(1), 116-124.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585-592.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498-503.

Kornmeier, J., Spitzer, M., & Sosic-Vasic, Z. (2014). Very similar spacing-effect patterns in very different learning/practice domains. *PloS one, 9*(3), e90656.

Kost, A. S., Carvalho, P. F., & Goldstone, R. L. (2015). Can you repeat that? The effect of item repetition on interleaved and blocked study. *Proceedings of the 37th Annual Conference of the Cognitive Science Society,* 1189-1194.

Landauer, T. K. (1978). *Optimum rehearsal patterns and name learning. Practical aspects of memory*. London: Academic Press.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494-513.

McDaniel, M. A., Roediger, H. L., & Mcdermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206.

Myers, D. G. (2008). *Exploring psychology* (9th ed.). New York: Worth.

Oltmanns, T. F., & Emery, R. E. (1995). *Abnormal psychology*. Englewood Cliffs, N.J: Prentice-Hall, Inc.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*(2), 187-193.

Pavlik Jr, P. I. (2007). Understanding and applying the dynamics of test practice and study practice. *Instructional Science, 35*(5), 407-441.

Pavlik Jr., P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation based model of the spacing effect. *Cognitive Science, 29*(4), 559-586.

Pavlik Jr., P. I., Presson, N., Dozzi, G., Wu, S.-m., MacWhinney, B., & Koedinger, K. R. (2007). The FaCT (Fact and Concept Training) System: A new tool linking cognitive science with educators. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1379-1384). Mahwah, NJ: Lawrence Erlbaum.

Reder, L. M., & Anderson, J. R. (1982). Effects of spacing and embellishment on memory for the main points of a text. *Memory & Cognition, 10*(2), 97-102.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Roediger, H.L., III, & Karpicke, J.D. (2006b). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249-255.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39*(5), 406-412.

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review,* 1-8.

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition, 42*(6), 912-921.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

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

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*(9), 641-656.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*(3), 210-221.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*(2), 175-184.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750-763.

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

Zulkiply, N., & Burt, J. S. (2012). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*(1), 16-27.

Appendix: Sample of a Case Study

Karen Rusa, 30 years old, is a married woman and a mother of four children. For the past several months Karen has been experiencing intrusive, repetitive thoughts that center around her children's safety. Karen also has noted that her daily routine is seriously hampered by an extensive series of counting rituals that she performs throughout each day. She has described herself as tense, jumpy and unable to relax. She has also reported dissatisfaction with her marriage and problems in managing her children. During the past several weeks, she has been spending more and more time crying and hiding alone in her bedroom (Oltmanns et al., 1991).

Psychological Disorder Type: Obsessive Compulsive Disorder

Submitted: 9.4.2017
Revised: 3.13.2018
Accepted: 3.15.2018