CHANG, YOOJIN, M.A. Does the Testing Effect Impact Favorability Judgments? (2013)
Directed by Dr. Peter F. Delaney, 59 pp.

Delayed recall of material is better when the material is retrieved on a previous occasion relative to when the material is restudied -- a phenomenon known as the *testing effect*. The studies reported here aimed to better understand the link between a person's memory and favorability judgments through the testing effect. Both memory and favorability judgments can be enhanced through persuasive messages, but both weaken over time. Considering that being tested decreases forgetting of the material and aids delayed retention, would being tested on a material decrease the decay in favorability judgments? I provided participants with a set of arguments that made the case for a topic and asked them to learn the arguments either through test or restudy. Either an immediate or 2 day delayed favorability judgment task was then given. A marginal testing effect was found in the memory tests. However, there was no testing effect found in favorability judgments.

Key Words: memory, favorability judgments, testing effect, forgetting, persuasion

DOES THE TESTING EFFECT IMPACT FAVORABILITY JUDGMENTS?

by

Yoojin Chang

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
2013

Approved by

Committee Chair

APPROVAL PAGE

This thesis has been approved by the following committee for the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Peter Delaney

Committee Members _____

Dayna Touron

_____

Ethan Zell

_____

Date of Acceptance by Committee

June 27, 2013

_____

Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF FIGURES v

CHAPTER

I. BACKGROUND AND RATIONALE 1

II. OUTLINE OF PROCEDURES 6

Study 1 6
Study 2 11
Experiment 15

III. GENERAL DISCUSSION 30

Why Was There No Benefit of Testing on Delayed Favorability Judgments? 31
Why Was the Testing Effect Present in the Memory Scores? 34
Was the Power Sufficient? 35
Conclusions and Future Direction 36

REFERENCES 38

APPENDIX A. TABLES 42

APPENDIX B. FIGURES 54

LIST OF TABLES

Page

Table 1. Descriptive statistics (mean and standard deviation) for agreement ratings, relevance to the self, and relevance to other UNCG students of 20 topics 42

Table 2. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 4: Everyone should be required to take at least one exercise course per semester 44

Table 3. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 9: Surveillance cameras should be installed around campus 45

Table 4. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 13: The cafeteria should limit servings of fried** food 46

Table 5. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 14: Foreign language courses should not be mandatory 47

Table 6. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 15: Every class should follow the same attendance policy 48

Table 7. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 16: Male and Female students that are not family should be able to be roommates on campus 49

Table 8. Materials used in the test condition for the experiment presented by topics 50

Table 9. Frequency of arguments recall and their persuasiveness judgments 51

Table 10. Correlations between favorability judgments and relearning format of arguments in the immediate condition 52

Table 11. Correlations between favorability judgments and relearning format of arguments in the delay condition 53

LIST OF FIGURES

Page

Figure 1. Scatterplot for self-relevance and participants' agreement with the topics 54

Figure 2. Scatterplot for other student relevance and participants' agreement with the topics 55

Figure 3. Scatterplot for other student relevance and participants' self-relevance of the topics 56

Figure 4. Mean favorability judgments as a function of relearning format and rating test delay 57

Figure 5. Mean favorability judgment latencies as a function of relearning format and rating test delay 58

Figure 6. Mean of recalled arguments as a function of relearning format and rating test delay 59

CHAPTER I

BACKGROUND AND RATIONALE

Providing people with persuasive arguments is one way to change their opinion (Cacioppo & Petty, 1989). However, persuasion does not last forever and messages are forgotten over time. The goal of this study is to determine if being tested on persuasive messages prolongs the initial persuasive state. For example, imagine you have just given a speech as you are a candidate running for congress. Would it be better to ask the audience to talk about the speech with friends or ask them to come to another rally? The former would be a form of a test and the latter would be a form of restudy.

When memory for material is tested after a delay, items that have been retrieved on previous occasions are better recalled than items that have been restudied but not recalled. This phenomenon is called the *testing effect*. For example, let's say a list of words is initially studied. Half of the list is restudied, meaning a person saw identical words again. The other half of the words is tested, meaning a person was asked to retrieve the word. On an immediate memory test, the restudied words are better recalled. After a delay (e.g., 2 days or a week – Roediger & Karpicke, 2006; Karpicke & Roediger, 2007; Karpicke & Roediger, 2008) the overall retention decreases; however, the tested words are less likely to be forgotten than the restudied words.

Given the substantial evidence that the testing effect produces slower memory loss over time compared to restudy (for reviews, see Delaney, Verkoeijen, & Spirgel,

2010; Roediger & Butler, 2011), would testing also affect the duration of one's formed attitudes or judgments? Many of our judgments are constructed using information retrieved from memory (e.g., Hastie & Park, 1986; Begg, Anas, & Farinacci, 1992). I suggest a link between testing effects and judgment because as time passes, not only do we lose access to information stored in memory due to the decay of the encoded material (e.g., Brown, 1958) but we also show less persuasion over time from that information on our attitudes or judgments due to decay (Kumkale & Albarracín, 2004). If we base our judgments on memory, and if memory fades away after a delay, then perhaps strengthening memory for arguments via testing will result in the changes in judgment persisting after testing as well.

In this study, I specifically focused on favorability judgments. As Eagly and Chaiken (2007) stated, one can "evaluate a particular entity with some degree of favor or disfavor." I will refer to how much one favors the object as *favorability judgments*. A favorability judgment is a class of attitude, which is "a person's evaluation of any psychological object" (Ajzen & Fishbein, 1980). The concept of favorability judgment is very close to the conventional definition of attitude. However, in this particular study favorability judgment is a more precise term that is closer to what I intend to understand than attitude.

There are a wide range of opinions about when people form their attitude or make a judgment about something. Indeed, it is likely that there are many different factors that all affect our attitude formation. Zajonc (1980) proposed that attitudes are formed instantaneously and without effort. A recent study by Hepler and Albarracin (2013) found

that attitudes are not only affected by external stimuli but also affected by peoples' pre-existing types of personality. The availability hypothesis proposed by Tversky and Kahneman (1973) suggests that the judgments are based on a heuristic that is the number of available pieces of evidence. Hastie and Park (1986) also showed for judgments that require memory, the available information in memory is a key factor. Such research might suggest that judgment is often not made systematically after careful consideration of factors, but rather instantaneously based on available information.

Earlier, in field of attitude change, McGuire (1968) suggested that people go through a series of stages that include exposure and attention to the statements, comprehension of the statements, and yielding to the statements. Though McGuire did not find retention of the statements to be a fundamental factor, Albarracin (2002) suggested that information (i.e., memory of the argument) is processed differently, as Petty and Cacioppo's (1986) Elaboration Likelihood Model (referred to as ELM throughout the paper) suggested. The model identifies two routes for change in attitude, also known as persuasion. The two routes are the central route and the peripheral route. The central route concerns careful processing of the merits provided by the argument such as the persuasiveness of the argument's content. The peripheral route concerns rather cursory processing of the external cues associated with the argument such as the source of the argument. Previous research shows that the central route leads to more enduring persuasion.

Persuasion occurs with the systematic and detail-oriented processing of information as discussed in the central route of ELM and that this information processing

produces enduring persuasion. The testing effect that also assists detail oriented information processing and benefits delayed retrieval. Therefore, one might reasonably predict that testing would affect favorability judgments at a delay as well.

The current study applied the testing effect paradigm to favorability judgment tasks. By providing people with persuasive arguments and then strengthening them via testing, I hoped to enhance the durability of persuasion over time. A key aspect of the testing effect is that it is most strongly manifested only after a certain amount of time has passed. Being tested on items will enhance later recollection for the item compared to when restudying it. Chan and McDermott (2007) examined the testing effect with recognition tests and suggested that testing will enhance subsequent recollection leaving familiarity unchanged. The rich recollection induced through initial testing enables better memory performance than the restudying on the delayed memory test. Consequently, if the memory information is used to form the judgment, then testing should provide better recollection of the arguments at test and hence sustain persuasion.

Providing material with extra retrieval practice may sometimes elicit inaccurate memory (if the wrong information is retrieved during the test). Retrieving an argument incorrectly might lead to unexpected judgment changes (Braun-Latour & Zaltman, 2006). I therefore provided participants with feedback after each testing trial to avoid this consequence. Interestingly, if feedback – regardless of its presentation duration from the test - is provided after the participants have been tested during learning, the testing effect increases (see Roediger & Butler, 2011).

In the current study, two topics each supported by 8 arguments were presented and participants were asked to study the arguments. One topic's set of arguments were restudied and the other topic's set of arguments were tested. Participants were then asked to judge the favorability toward the topic either on an immediate test or after a 2 day delay. The participants were provided with a topic cue asking whether they favor the topic and were asked to judge their favorability on a 1 to 7 scale. Finally, memory for the arguments was tested using free recall.

I hypothesized that the immediate memory and favorability judgment would be loosely equivalent for both test and restudy conditions, as observed in previous testing effect studies. However, after a delay, I hypothesized that both memory and favorability judgments would decrease in general, with differential decrements determined by testing or restudying condition manipulation. As in previous testing effect studies, I predicted that being tested on an argument would lead to better memory for the material compared to the restudied material after a delay. In addition, I predicted that favorability judgments would show less decay over time in the tested condition relative to the restudy condition after a delay.

CHAPTER II

OUTLINE OF PROCEDURES

Study 1

The experiment described below examined the main hypotheses of the study, namely whether being tested on the provided materials will lead to a more favorable judgment. To conduct the experiment, however, I had to norm the topics and the arguments. Therefore, norming the materials for the experiment was the main purpose of Studies 1 and 2. Study 1 focused on topic sentences and their favorability judgments. A *topic* was defined as a single sentence stating a certain issue that could elicit favorable or unfavorable judgments. The topics that were judged to be neutral in Study 1 and arguments that are determined to be strongly favorable in Study 2 were then employed as stimuli in the experiment.

Producing normed materials for future favorability judgments research was another interest of Studies 1 and 2. Attitude studies use generated statements but hardly have the statements' qualities put to the test. Though some statements may be restricted to issues only regarding University of North Carolina at Greensboro, it is good to have the descriptive statistics of agreement or favorability judgment ratings as a reference. In the experiment, I focused on neutral topics and strong persuasive arguments by selecting the statements based on the stated criteria. However, if I choose to examine the testing effect on strongly agreed or weak non-persuasive arguments I could use the materials that

have been normed through Studies 1 and 2. Other researchers, who would like to generate similar issue statements on their campus, can refer to the normed materials from Studies 1 and 2.

Study 1 focused on finding not too provocative but still relevant issues for students on the UNCG campus. The reason why I chose to use neutral topics was to have baseline favorability judgments to compare with subsequent favorability judgments made during the experiment. The difference between the rated favorability judgments of the experiment and Study 1 would be determined as persuasion (attitude change). Using within subjects attitude measure will interfere because participants often use their memory of their initial favorability judgment instead of making a novel judgment to the topic. To avoid this, it seemed optimal to use a between-subjects design in order to ollect the favorability judgment data. According to the ELM, the more the statements are relevant and interesting to the participants the more likely the central route will affect the persuasion. To understand how relevant the participants think the topics are to themselves and to other UNCG students, the relevance measure was included.

Methods

Participants. Twenty student participants at the University of North Carolina at Greensboro were recruited at open campus areas that attract many students (i.e., Student center EUC: Elliot University Center and dormitory cafeteria). The participants were asked to volunteer their time without expectation of compensation. However, when the experiment ended, a piece of candy was offered as a thank-you. An in-lab pilot with

research assistants estimated the average required time for participation to be about 10 minutes. The participants were asked to volunteer 10 minutes of their time; however, the experimenters did not stop the one participant who needed more time to rate the scales. This participant had difficulty understanding the questions since English was their second language, and finished within 20 minutes. It took half a day for data collection. The participants showed good compliance with the instructions, according to research assistants' reports.

Materials. Twenty topics were generated as stimuli for favorability judgments. The topics that reflect what would interest the participants (undergraduate students) and also would not be too provocative were selected. A topic was a simple, ordinary sentence such as "At least one exercise course per semester should be required." The materials are presented in Table 1.

Procedure. A seven page paper packet of topics was given to the participants. The purpose of the experiment, which was to evaluate the degree of agreement and relevance for each topic, was introduced, as well as the instructions. The participants were also informed about the structure of the packet, which contained 20 topic sentences. After carefully reading each topic the participants were asked to judge how much they agree with the topic by rating the degree of agreement on the provided 7-point Likert scale. Each topic was presented with a rating scale right below the topic sentence. The topics were presented all together on a sheet of paper.

Participants were instructed to rate each topic on three measures and then move to the next topic. The degree of agreement with the statement, the relevance of the statement to themselves, and the relevance of the statement to other UNCG students were rated. Instructions about the scales were verbally provided by the experimenters and an example of an arbitrarily marked scale was provided as visual illustration on the packet. On the agreement scale, 1 indicated that the participant absolutely did not agree with the topic sentence, 4 indicated the participant was neutral towards the topic, and 7 indicated that the participant absolutely agreed with the topic sentence. Ratings for self-relevance and other UNCG student relevance to the topic sentence were also measured on a 7-point Likert scale.

Results and Discussion

I computed descriptive statistics for each topic (see Table 1). I used the mean favorability judgment rating for each topic. As aforementioned, the ratings were on a 7-point Likert scale and the topics with the means closer to neutral (mean = 4 ± 2) were flagged as prospective stimuli. The topics that had a standard deviation greater than 2 were not selected. When adding one standard deviation to the mean the highest mean rating would be a 6 and the lowest would be a 2. In the former case (mean = 6), the topic would be determined strongly favorable since it is only one digit lower than the possible maximum rating, which is 7 on a 7 point Likert scale. The latter case (mean = 2), the topic would be determined strongly unfavorable since it is only one digit higher than the possible minimum rating, which is 1 on a 7 point Likert scale. To minimize the conflicts

in topic selection, the first selection criterion was the mean closest to neutral which was a rating score of approximately 4 ± 2. The topic sentences that fit these criteria were:

- **Topic 4**, which stated, "Everyone should be required to take at least one exercise course per semester."
- **Topic 9,** which stated, "Surveillance cameras should be installed around campus."
- **Topic 13**, which stated, "The cafeteria should limit servings of fried food."
- **Topic 14**, which stated "Foreign language courses should not be mandatory."
- **Topic 15**, which stated, "Every class should follow the same attendance policy."
- **Topic 16**, which stated, "Male and female students that aren't family members should be able to be roommates on campus."

A significant correlation was found between how the participants' thought each topic was relevant to them and how much they agree on the topic, $r(18) = .466$, $p = .038$. Figure 1 shows a scatterplot of this correlation. However, there was no significant correlation (see Figure 2 for a scatterplot) found between the participants' agreement to the topics and other UNCG student-relevance, $r(18) = .246$, *ns.* Also, there was a significant correlation (see Figure 3 for a scatterplot) between how much the participants thought each topic was relevant to them and to other UNCG students, $r(18) = .802$, $p < .001$. This correlation was very strong, suggesting that people may have been assuming that things relevant to them are also relevant to other students

Conclusion

The purpose of Study 1 was to norm the materials for the experiment and for future favorability judgment studies. Out of the 20 topics that were generated, six of the topics were selected to be viable topics for future experiments in that they had an agreement(favorability judgment) near 4 which was pegged as neutral. In Study 2, arguments will be generated to support the selected topics and their persuasiveness will be tested.

Study 2

The main purpose of Study 2 was to identify arguments that are persuasive enough to participants to change their opinion on a topic. As Cacioppo and Petty (1989) mentioned, the arguments will be determined *strong* if the overall judgment of the argument is deemed favorable. Strong arguments were needed because I wanted to persuade the participants about the topic sentence. In Study 1, the neutral topics were selected. To persuade the participants, in other words, to show changes in the judgment of favorability for the topics, the arguments for each topic needed to be strong (larger than 4 on a 7 point Likert scale).

Methods

Participants. Participants were recruited on campus in a fashion similar to how participants were recruited for Study 1. A total of 35 participants' data were collected. Of the 35 participants, research assistants who helped data collection reported that four

participants did not seem to understand the purpose of the survey or did not take it; therefore their data were excluded. All participants were students of the University of North Carolina at Greensboro. The participants were aware that there was no official compensation for their participation. However, the data were mostly collected on or close to Valentine's Day, so a piece of candy was provided as a way of appreciating their participation. Most of the participants finished the task within 20 minutes. The compliance rate was good and the data collection was completed in a day.

Materials. The materials were the 6 arguments from Study 1 that were rated as neutral (mean rating of approximately 4 out of a 7 point scale). The six topics that were preselected through this route were: "Everyone should be required to take at least one exercise course per semester (Topic 4 in Study 1) ," "Surveillance cameras should be installed around campus (Topic 9 in Study 1) ," " The cafeteria should limit servings of d food (Topic 13 in Study 1) ," "Foreign language courses should not be mandatory (Topic 14 in Study 1)," "Every class should follow the same attendance policy (Topic 15 in Study 1) ," "Male and female students that aren't family members should be able to be roommates on campus (Topic 16 in Study 1)."

The topics each consisted of 16 arguments which were made to support the topics. The arguments for each topic were simple and ordinary sentences. An example of the topic could be: "At least one exercise course per semester should be required." Examples of the argument sentences that should make the case for the topic sentences could be: "Regularly exercising can prevent obesity," "Even stretching can promote health by

enhancing blood circulation," or "It is difficult to exercise regularly without accountability." There were 6 topics and each 16 arguments for each. The arguments associated with the topic were all aligned in a table. There were 6 tables and 2 tables were presented on a sheet of paper. There were 3 sheets of paper to this packet. There were small empty boxes after each argument for rating each argument's persuasiveness towards each associated topic on a 7-point Likert scale. On this scale, 1 indicated that the argument was absolutely not persuasive, 4 indicated that the argument was neutral in terms of persuasion, and 7 indicated that the argument was absolutely persuasive.

Procedure. The purpose of the experiment – to evaluate the degree of persuasiveness for each argument associated with the topics –was introduced. The paper packet of topics and its arguments were handed to the participants on individual clipboards with a pen or pencil. The participants were informed about the structure of the packet; the topic sentences (that were preselected from Study 1) and its set of arguments. After carefully reading each argument associated with the topics, the participants were asked to judge the degree to which they were persuaded by the argument by rating the degree of persuasiveness next to the small empty box at the end of each argument. Instructions about the scale were verbally explained by the experimenters. The participants were allowed to take their time on each argument evaluation. They were asked to rate each argument one at a time and advance to the next argument after they had rated an argument.

Results and Discussion

Descriptive statistics for each argument were computed (see Table 2, Table 3, Table 4, Table 5, Table 6, and Table 7 for details). The mean persuasiveness rating for each topic was the information of interest. The ratings were on a 7-point Likert scale, and the arguments with a mean higher than 4 were identified as prospective stimuli. I proposed to use the arguments whose mean score was above 5.5 for the experiment. However, the data suggested I use a less strict selection criterion because there were insufficient numbers of arguments that fulfilled the proposed criteria.

While I originally proposed to use strong arguments which would be determined by a persuasiveness judgment above 5.5 on a 7-point scale, the results showed that not many arguments fit the criterion. I caught a mistake in the making of the experiment sheets after the data was collected. Originally Topic 13 was “The cafeteria should limit servings of fried food.” However in Study 2 the topic sentence presented to the participants read “The cafeteria should limit servings of food.” This topic’s arguments did not meet the criterion for selection but perhaps the omission of the key word “fried” from the topic could have led to the low persuasiveness ratings of the arguments.

In the experiment, I proposed to use at least eight arguments for the paradigm so the topics that had sufficient number (> 8) of arguments whose mean was above 4 were used. Of the six topics, Topic 4 (10 arguments) and Topic 9 (10 arguments) had sufficient numbers of arguments. The arguments with the highest persuasiveness rated in Study 2 were proposed to be used. However, when the experiment was first proposed, 25% of the

words in the arguments were to be left out as blanks (more details are given in the Materials section of the experiment). To match the sentence length between the two topics as well as the average persuasiveness, some sentences were chosen to fit this criterion of convenient number of words per sentence. When a pilot was run for the experiment which led to a change in the testing format, the selected arguments were not switched back to the original standard, which was the most persuasive argument sentence.

Conclusion

Persuasive arguments were successfully normed to fulfill the purpose of potential materials for the experiment. The persuasiveness level for each argument was not as ideal as it was originally proposed. However, the persuasiveness level was still above neutral (Topic 4, Mean = 5.12; Topic 9, Mean = 5.13), thus providing marginal support for the topic sentences. Of the 6 topic sentences selected from Study 1, Topic 4 and Topic 9 each had 10 arguments that could be deemed as strong arguments and those were selected as potential stimuli for the experiment. Future studies in attitude can use the provided arguments normed from Study 2 to test theories regarding strong and weak sentences.

Experiment

The experiment aimed to determine whether tested arguments result in higher favorability judgments after a delay than restudied arguments do. I hypothesized that restudied arguments will be less recalled and judged to be less favorable on a delayed task. I also hypothesized that with feedback provided after the test, the participants will

show similar rates of recall and favorability judgments on an immediate memory and judgment tasks.

Methods

Power analysis. A power analysis was conducted with G-Power (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009) a statistical tool that was used to determine *a priori* the minimum sample size needed to achieve .90 power for a two way ANOVA with 2 between and 2 within subjects factors. The estimated effect size of the within and between subjects factors' interaction was $f = .25$ and the desired power was .9. The estimated correlation among the repeated measures was $r = .40$. The number of participants that need to be recruited to obtain the power of .9 ($f = .25$) is 56 total.

Participants. Participants were 58 English-speaking UNCG students ages 18-30 who received course credit or money for participation. The participants were recruited across the spring semester and summer sessions. Most were recruited through participant pools, but 5 were paid $10 for participating after responding to a flyer. An additional 13 volunteered to participate for free.

Design. The design of the experiment was a 2 relearning format (tested vs. restudied) x 2 time point (delayed vs. immediate) mixed factorial design. The within-subjects factor was the initial relearning format (restudy vs. tested) and the between-subjects factor was the delay before making the favorability judgment and taking the memory test (immediate vs. after a 2 day delay). The order of presentation of the two

topics was counterbalanced, along with whether the arguments for each of the two topics were tested or restudied. For a given topic, the presentation order of the arguments was randomized during the first study session, but then the same order was maintained in the following relearning session.

Materials. The materials are presented in Table 8. They consisted of two opinion topics that were preselected from Study 1 that were overall rated as neutral. The topic sentences were distinguished from the argument sentences because there was a black square line around the topic screen. The topics each were supported by eight persuasive arguments that were preselected from Study 2. The materials for the restudy condition were the same whereas the materials for the test condition had blanks in the sentence for completion.

The blanks were preselected based on their contribution to overall sentence comprehension. The to-be-deleted-words are preselected content words crucial to the overall thesis of the sentence, consistent with previous literature on sentence completion (e.g., Hinze & Wiley, 2011). Initially, I planned to replace 25% of the words with blanks in each of the argument sentences (i.e., if the sentence consisted of 20 words, 5 words would be chosen to be left as a blank). However, the pilot data suggested that filling in 25% of a sentence was too hard of a task, so I limited the blanks to be two for each sentence at the most. Finley, Benjamin, Hays, Bjork, and Kornell (2011) showed that final test recall rates were higher when participants were asked to recall sentences based on diminishing cues (adding more blanks to fill in for the target as trials progress) than when recalled

sentences based on accumulating cues (decreasing the number of blanks to fill in for the target as trials progress). I therefore incorporated this method into the experiment by testing participants on each argument in the testing condition twice, increasing the number of blanks to fill in by one from the first to the second test.

Procedure. After signing the consent form, the participants were seated in front of a computer and were told to attend to the argument sentences presented on the screen. The instructions specified that there was to be a study session and a relearning session. The participants were informed that there would be a subsequent judgment task, but the memory task was a surprise. The participants read through the instructions by advancing the slides with a space bar and were asked to explain to the experimenters what they were supposed to be doing.

All participants then completed the three phases of the experiment. The three phases were the study phase, relearning phase, and the test phase, and I will describe each of them below.

Study phase. First, a topic slide that had the topic sentence written in black on a white background with a black square outline at the edge of the screen was presented for 6 s. Next, the eight arguments that supported the topic were individually presented on the computer screen in black on a white background for 12 s each with a 3 s blank screen as an inter-stimulus interval. The other topic with its arguments list was then presented in the same manner. The order of topic presentation was counterbalanced. The order of argument presentation was randomized. A series of math problems were solved on paper

using a pen/pencil as a distracter task for 60 s after the study phase, which consists of the presentation of the two topic and its arguments.

Relearning phase. During the relearning phase, one topic's arguments were restudied while the other topic's arguments were tested. Which particular topic was assigned to the restudy condition and which was assigned to the tested condition was counterbalanced across participants. Which topic was restudied first was the same as their order of presentation during the study phase. The randomized argument presentation order established during the study phase was constrained to be the same throughout the relearning phase.

In the *restudy condition*, the participants saw the arguments once for 12 s with a 3 s inter-stimulus interval and then saw all the arguments again for the second time in the same fashion. In the *test condition*, the participants typed in the words that completed the sentences. Immediately after giving their answer, a feedback screen appeared for 2 s showing the arguments again followed by a 1s inter-stimulus interval. For the first round of testing, the participants filled in one blank within the argument sentence. For the second round of testing for the same argument, they filled in two blanks in the argument sentence. After the relearning (test or restudy) phase was completed, another series of math problems as a distracter task was provided for 60 s in the same fashion during the study phase.

Test phase. Half of the participants received the test phase immediately after completing the relearning phase, while the rest returned after a 2-day delay. After a brief

instruction was provided, the two topic sentences were presented separately. A 7-point Likert scale arrow was drawn beneath the topic sentence. The slide would only advance if the participants entered a number between 1 and 7. The participants were asked to judge how much they agreed with or favored the topic. On the scale, 1 indicated that they absolutely did not agree with the topic and 7 indicated that they absolutely did agree with the topic. The rating procedure was self-paced. The order of the topics' presentation was randomized.

After the favorability judgment data were collected, a memory task took place. The participants were asked to recall as much as they could about the arguments for each topic. To avoid omission due to poor handwriting, the participants were asked to type their response on a Microsoft Word document. The participants saw both of the topic sentences simultaneously and could choose whatever topic they wanted to start with when recalling the entailed arguments. They were instructed to take their time and to do their best to at least write the gist of the arguments. This memory test was used as a reference when comparing the actual recall of the sentence to the degree of how much memory the participants retained during the favorability judgment task. The memory test was self-paced and participants alerted the experimenter when they were finished.

As in Studies 1 and 2, after the experiment was completed the participants received the debriefing sheets. They were told that the arguments and topics were generated by the experimenters and had nothing to do with the school's policy.

Results and Discussion

Evidence for persuasion. To even begin discussing whether there was a testing effect in persuasion it was fundamental that I ask if there was any persuasion evident. Did the persuasive arguments increase favorability judgments? I investigated this question by comparing the difference between the two topics' favorability judgment scores from Study 1 and the experiment.

In Study 1, the mean favorability judgment (agreement to the statement) of Topic 4 (At least one exercise course should be required every semester) was 3.97(±1.96) and the mean favorability judgment (agreement to the statement) of Topic 9 (Surveillance cameras should be installed around on campus) was 4.37(±1.77). In the experiment, the mean favorability judgment of Topic 4 was 4.81 (±1.52) and the mean of Topic 9 was 5.42 (±1.39). An independent samples *t*-test showed that the mean difference for Topic 4 in experiments 1 and 3 increased by a reliable 0.87 ($t(79) = -2.24$, $p = .028$) and for Topic 9 by 1.43 ($t(79) = -3.47$, $p = .001$). Thus, presenting the arguments apparently increased the favorability judgments in the experiment since Study 1. The difference was significant for both Topic 4 and Topic 9.

Testing effects in favorability judgments. This question was the primary question of the experiment. I originally hypothesized that the favorability judgment data would show a testing effect. That is, on a final test after several days, the favorability judgment for a topic would be higher when it had been tested earlier than when the topic's arguments were restudied. To investigate this question, a mixed factorial ANOVA

was conducted. The relearning format (restudy versus test) was the within-subjects factor because all participants restudied one topic's list of arguments and tested the other topic's list of arguments. Whether the test phase was performed immediately or after 2 days were the between-subjects factor. The means of favorability judgments made for each topic were combined and compared according to the factor.

The means and standard errors of the groups are provided in Figure 4. The main effect of the within-subject factor relearning format (restudy versus test) was not significant, $F(1, 56) = 0.34$, $MSE = 0.70$, $\eta^2 = .006$, $p = .561$. The main effect of the between-subject factor test delay (immediate versus delay) was also not significant, $F(1, 56) = .17$, $MSE = .17$, $\eta^2 = .003$, $p = .681$. With 58 participants, the interaction between relearning format and delay was not significant, $F(2, 56) = .34$, $MSE = .70$, $\eta^2 = .006$, $p = .561$.

Implicit Association Tests that examine the response latency recorded when making judgments suggest that response latency can be a valid measure that is predictive of attitude strength (Maison, Greenwald, & Bruin, 2004). In the process of data collection for the experiment, I programmed the experiment to collect the participants' response times when making favorability judgments. The means of response latency made for each topic was combined and compared according to the factor. Faster response time would mean stronger attitudes toward the topic. The relearning format (restudy versus test) was the within-subjects factor because all participants restudied one topic's list of arguments

and tested the other topic's list of arguments. Whether the test phase was performed immediately or after 2 days were the between-subjects factor.

Preliminary analysis of raw reaction time (from the onset of the topic sentence presentation on screen to the moment the participants pressed a number key to indicate their favorability judgment) was computed. The main effect of the within-subject factor relearning format (restudy versus test) was not significant, F (1, 56) = 1.46, MSE = .00, η^2 = .025, p = .232. The main effect of the between-subject factor test delay (immediate versus delay) was not significant, F (1, 56) = 1.36, MSE = .00, η^2 = .024, p = .249, and there was no interaction, F (2, 56) = .03, MSE = .00, η^2 = .001, p = .865. Therefore, for both reported favorability judgments and reaction times, I obtained no evidence for a testing effect.

Transforming the non-normally distributed raw reaction time data into a log10 data the pattern did not change the raw data analysis. There were neither main effects of format (restudy versus test), F(1, 56) = 1.50, MSE = .09, η^2 = .026, p = .225, nor test delay (immediate versus delay), F(1,56) = .20, MSE = .01, η^2 = .004, p = .653. There was also no significant interaction between the two factors, F (2, 56) = .01, MSE = .06, η^2 = .00, p = .936). The means and standard errors of the groups (transformed back) are provided in Figure 5.

Testing effects in memory[1]. Because the design of the experiment incorporated a typical testing effect paradigm, I hypothesized that fewer arguments would be forgotten after a 2 day delay if participants were tested on them initially rather than restudying them. To investigate this question a mixed-factorial ANOVA was conducted. The relearning format (restudy versus test) was the within-subjects factor because all participants restudied one topic's list of arguments and tested the other topic's list of arguments. Whether the test phase was performed immediately or after 2 days were the between-subjects factor.

Two scorers separately scored each argument on a scale of 0 to 10 independently of each other. The scorers did not know which recalled argument data was for which participant to avoid biased scoring due to knowledge of the condition. The scorers read through the recalled arguments and coded the arguments and scored the quality of the recalled arguments. First they gave the argument a code For example, for an argument of Topic 4 that read, "It can lower obesity rates," the code would be *obesity*. An argument that was recalled verbatim would be scored as a 10 and nothing recalled would be rated a 0. When the two raters' opinion had a difference score of 3 or larger, the two raters

[1] *NOTE*: I proposed that two scorers would collect the free recall data of arguments for each topic and assign one point for each argument that was recalled correctly. The maximum score should be 8 for each topic since there were 8 arguments to be recalled within each topic. The minimum score would be zero. The argument did not have to be recalled verbatim; however, the two scorers would have to agree that the gist of the argument was properly recalled. Synonyms were accepted as answers. The raters were trained to code the recalled arguments following the scoring criteria described above. Each argument was given a code if the raters considered the content of the argument to be sufficient enough to be labeled as recalled. The inter-rater reliability was calculated at the item level and since there were two raters the Pearson correlation was used. For Topic 4, the inter-rater reliability was r (58) = .829, p <.001, and for Topic 9, the inter-rater reliability was r(58) = .926, p < .001. The means and standard errors of the groups are provided in Figure 6. The main effect of the within-subject factor relearning format (restudy versus test) was significant (F (1, 56) = 6.90, MSE = 1.32, η^2 = .11, p = .011). The restudied topic (M = 3.41, SD = .22) entailed significantly more recalled arguments than the tested topic (M = 2.88, SD = .18). The main effect of the between-subjects factor, test delay (immediate versus delay), was significant (F (1, 56) = 20.95, MSE = 3.01, η^2 = .27, $p < .001$). The memory test given immediately (M = 3.85, SD = .21) produced more recalled arguments than when the test was given after 2 days (M = 2.44, SD = .21). However, the interaction between relearning format and delay was not significant (F (2, 56) = .28, MSE = 1.32, η^2 = .005, p = .601). The answer to the question whether there was a significant testing effect found in quantitative memory was no.

discussed their reasoning and made changes to their ratings. When the arguments were coded and scored, I added the total scores rated for each participant and divided that by 8, which was the total number of arguments per topic. The highest number that could have been achieved would be a 10 because (10 + 10 + 10 + 10 + 10 + 10 + 10 + 10)/8 is 10. The lowest would be a 0. Nobody recalled all 8 arguments for either topic; therefore a ceiling effect was avoided. However, there were a few participants that had a floor effect and did not recall at all. For the arguments that were not coded and therefore pegged as "not – recalled" were assigned a 0 in the equation. The inter-rater reliability was calculated at the item level, and since there were two raters the Pearson correlation was used. For Topic 4 the inter-rater reliability was r (58) = .775, $p < .001$, and for Topic 9 the inter-rater reliability was r (58) = .712, $p < .001$.

Using the memory scores (the average score for each recalled item rated by the raters), a mixed factorial ANOVA was run with the relearning format as the within-subjects factor and the delay as the between-subjects factor. The means and standard errors are provided in Figure 7. The main effect of the within-subject factor relearning format (restudy versus test) was not significant, F (1, 56) = 1.44, MSE = 1.42, η^2 = .025, p = .235. The main effect of the between-subject factor test delay (immediate versus delay) was significant, F (1, 56) = 40.38, MSE = 3.39, η^2 = .419, $p < .001$. After a 2 day delay (M = 2.23, SD = .24) people recalled significantly fewer arguments than when the memory test was provided immediately (M = 4.27, SD = .242). The interaction between relearning format and delay was marginally significant, F (2, 26) = 3.27, MSE = 1.42, η^2 = .055, p = .076. In the immediate condition, the tested topics (M = 4.70, SD = 2.01) had

marginally more arguments recalled than the restudied topics (M = 4.03, SD = 1.53), t (28) = 1.98, p =.064. In the delayed condition, the tested topics (M = 2.58, SD = 1.20) had significantly more arguments recalled than the restudied topics, (M = 1.90, SD = 1.01), t (28) = 3.09, p = .004. The answer to the question whether there is a marginally significant testing effect found in qualitative memory is yes.

Memory for Arguments and Favorability Judgments. Was there a relationship between the memory of arguments and its originally scored persuasiveness? Would the arguments that were frequently remembered be those that were rated as strong persuasive arguments in Study 2? To answer these questions, I analyzed the frequency of codes that were made when scoring the arguments' memory scores. The basic frequency of all arguments and their persuasiveness found in Study 2 are presented in Table 9. For Topic 4, the most frequently recalled argument was "College is about becoming a well-rounded *person: mind, body, and spirit.*" Being recalled 18.75%. The persuasiveness of this argument found in Study 2 was 5.03, which was not the highest. For Topic 9, the most frequently recalled argument was *"The footage won't go public so your privacy will still be protected."* being recalled 17.98% of the time. The persuasiveness of this argument found in Study 2 was 4.45, which was also in the rather lower bound of the persuasive arguments. A correlation between the arguments' frequency rate and their initial persuasiveness was computed. For Topic 4, there were no significant correlation detected, r (8) = -.083, p = .846. However, for Topic 9, there was a significantly negative correlation found, r (8) = -.752, p = .032. For Topic 9, the more frequently recalled arguments were those whose persuasiveness was closer to neutral (4).

In the introduction, I suggested that episodic memory can aid in judgments. Though it is hard to make a causal claim regarding the relationship between favorability judgments and memory, a correlation can be investigated to see whether there is a relationship and if so to what degree it is present. A correlation between qualitatively sensitive memory scores and favorability judgments that were segmented into their restudy, test, immediate, and delay was computed. The results are presented in Table 10 and Table 11.

In the immediate condition, at the significance level of $p < .001$ there was a correlation found between rand the quality of tested arguments' memory and the quantity of restudied arguments' memory, $r\ (29) = .852$, $p < .001$. At the significance level of $p < .01$ there was a correlation found between restudied topics' favorability judgments and the favorability judgment latency for the tested topics, $r\ (29) = .550$, $p = .002$, a negative correlation between the quality of the tested arguments' memory and the reaction time to the tested topics' favorability judgment, $r\ (29) = -.481$, $p = .008$, and a correlation between the quality of the tested arguments' memory and the quality of the restudied arguments' memory, $r\ (29) = .511$, $p = .005$. Other correlations found at the significance level of $p < .05$ were between favorability judgment latency of the tested topics and restudied topics, $r\ (29) = -.369$, $p = .049$, quality of restudied arguments' memory and favorability judgment latency for tested topics, $r\ (29) = -.391$, $p = .036$, and the quality and quantity of the restudied arguments' memory, $r\ (29) = .427$, $p = .021$.

In the delay condition, at the significance level of $p < .001$ there was a significant correlation found between the quality of tested arguments' memory and the quantity of restudied arguments' memory, r (29) = .921, $p < .001$, and the quantity of the tested arguments' memory and the quality of the restudied arguments' memory, r (29)=.879, $p < .001$. At the significance level of $p < .01$ there were correlations found between restudied topics' favorability judgments latency and the quality of the tested arguments' memory, r (29) = .473, $p = .01$. Other correlations found at the significance level of $p < .05$ were between favorability judgment latency of the restudied topics and the quantity of the restudied arguments' memory, r (29) =.467, $p = .011$, quality and quantity of restudied arguments' memory and r (29)=.429, $p = .020$, and the quality of the restudied and tested arguments' memory, r (29)=.425, $p = .022$.

Conclusions. Though I did find a marginal testing effect in the qualitatively sensitive memory scores, there were no significant testing effects found in the quantitatively sensitive memory scores or in favorability judgments. The experiment was expected to show that the tested arguments would lessen the degree of decline in persuasion. What was interesting about the results of the experiment was that testing arguments during relearning did not affect the favorability judgment at the delay. The latency results seemed to replicate the favorability judgment data pattern.

However, there was a testing effect in memory – specifically, for those that were qualitatively sensitive when being scored. One concern with the experiment results was that the arguments chosen from Study 2 might not be as strongly persuasive as I had

hoped them to be. However, the arguments (disregarding their relearning format) seemed to have an effect on later favorability judgments in that there was significant increase in favorability judgments compared to Study 1 for both topics. More general thoughts and interpretation of the data will be discussed in the general discussion section.

CHAPTER III

GENERAL DISCUSSION

In Studies 1 and 2, the materials were normed to fulfill the purpose of the experiment. Neutral topics were found in Study 1 and marginally persuasive arguments were found in Study 2. In the experiment, there was difference in favorability judgments but there was no decline in favorability judgments after a delay. Also, being tested on the materials did not seem to have an effect on persuasion decline at the delay as hypothesized. However, I did find a marginally significant testing effect on a measure of memory that incorporated the quality of the recall of the arguments, although there was no testing effect on the number of arguments that were recalled.

Studies that incorporate episodic memory with judgments and attitude change such as advertisement studies using the spacing effect (e.g., Braun-Latour & Zaltman, 2006) often mention that the stimuli are best utilized by participants if they are personal or internalized. One of the main goals of this study was to norm statements and to determine the average persuasiveness and favorability judgments for the arguments and topics respectively. In order to achieve this goal, the topics were rated by different people in Study 1 from those who rated arguments in Study 2. Similarly, the participants in the experiment were not the same people from Studies 1 and 2.

The correlations between the quality of the tested arguments and the quantity of the restudied arguments, r (29) = .921, $p < .001$, as well as the correlations between the

quality of the restudied arguments and the quantity of the tested arguments, r (29) = .880, $p < .001$, in the delayed condition might suggest a limitation to the final memory test procedure in the experiment. The topics were simultaneously presented on the same page at the final free recall task. This was intended to avoid the order of a certain topic affecting the recall process. Though the participants were instructed to not restrict their response to the order in which they input their answers, they could have tried to level the amount of sentences they retrieve for the two topics regardless of how much or how well they remember the arguments.

Why Was There No Benefit of Testing on Delayed Favorability Judgments?

Simply put, the overall results suggest that being tested on persuasive arguments led to lasting memory advantages, but not necessarily to sustained persuasion as measured by favorability judgments. This study was the first to attempt to expand the notion of the testing effect impacting judgment beyond previously found effects on memory. Nevertheless, the results of the study did not reject the null hypothesis. Therefore, it is possible that there is no effect of testing on delayed favorability judgments.

Alternatively, perhaps unlike memory in which testing which is basically retrieval practice during testing helps later memory, merely restudying the topic can impact favorability judgments. According to the *mere exposure effect* (Zajonc, 1968), even neutral statements when repeatedly exposed to the participant can elicit increased favorability for that statement. The arguments were presented mainly for memory

purposes. However, multiple exposures of the arguments that were supporting the topics could have made the topics extremely familiar to the participants. Whether they were being tested on it or restudying it, the frequency of the argument presentation would equally trigger mental exposure of the related topic sentences. Perhaps that could be why I did not find a testing effect in favorability judgments: the restudy condition produced a mere exposure effect on the topics, and canceled out the effects of testing. Perhaps the addition of a control condition in which the arguments are not relearned at all would help prevent this phenomenon. I tried to minimize this effect of multiple exposures on the topic sentences by limiting the exposure of the actual topic sentences prior to the rating tests. The topic sentences were presented only once, immediately before receiving the arguments (during the study phase). However, it still could be considered as a caveat in that the opinions for the topics could be made at the initial exposure of the topic sentences when studying the arguments during the study phase. Future studies may omit the topic sentence at the initial exposure of the arguments and provide the topic sentences for the first time when they are evaluating the topics.

The blanks in the argument sentences were made by the experimenters and were piloted to enhance sentence comprehension and recollection. Perhaps if the participants were to do a free recall during the initial relearning phase instead of filling in the blanks or making their own fill in the blanks tests much like the flash cards students make when studying could help the materials become more internalized and thus reveal testing effects in favorability judgments. One reason to suspect this is that quantity of arguments recalled was unaffected by testing. Perhaps if we used free recall, the practice on

retrieving the entire argument would improve the number of arguments as well as their quality. If people use just the number of arguments retrieved to make their judgments and pay little attention to their quality (cf. Tversky & Kahneman, 1973), and if the fill-in-the-blanks testing mostly enhanced argument quality and not quantity, then perhaps a stronger manipulation of testing like using free recall testing might enlarge the effects of testing on the persuasiveness of the arguments.

Another explanation as to why there was no testing effect in favorability judgment could be that the participants' motivation was not at its highest. Previous research showed that a decrease in motivation and ability of the participants' judgment making will restrict and impair processing of the argument content (Petty & Cacioppo, 1986; Alba, Marmorstein, & Chattopadhyay, 1992; Albarracin, 2002). Instead, low motivated participants will resort to a more heuristic oriented processing of peripheral information which may not be related to the content of the arguments. To minimize a decrease in motivation or interest for the participants, I chose topics that would most likely arise on campus and measured perceived relevance of the arguments to them and to other UNCG students. However, it could be that this might not have been enough motivation for the participants to utilize the supportive information of the arguments that were provided multiple times and in various formats during the relearning phase. Perhaps reframing the question to lead the participants to think that their opinion towards the topic will affect the school policy could help in future investigations related to this study.

A correlation was found for the latency of the favorability judgments for the tested topics, which was negatively correlated with the quality of arguments recalled for the tested arguments in the immediate condition, r (29) = - 0.481, p = .008. Though I hypothesized this similar pattern would occur in the delayed condition, it seemed that the tested arguments were aiding faster judgments of favorability only immediately.

Why Was the Testing Effect Present in the Memory Scores?

Literature on the testing effect suggests that having retrieval practice during relearning enhances the recollection of the item whereas restudying the item enhances the familiarity of the item (Chan & McDermott, 2007). Familiarity is the feeling that occurs when you think you have seen an object before. Recollection is fully remembering the object in detail. When scoring the recalled arguments, the details recollected about the arguments were counted. Therefore, when the participants were free recalling the arguments, the arguments that had been tested on showed a benefit even after a 2-day delay. This finding of a testing effect in memory data supports Chan and McDermott's claim about the testing effect enhancing recollection and therefore will benefit recollection laden task performances. Chan, McDermott, and Roediger (2006) showed that testing facilitates recall for items that were not even initially tested after a 24 hour delay. However, mainly studied with recalled items and usually the biggest effect occurs with the recalled items (Delaney, Verkoeijen, & Spirgel, 2010). Therefore, the fill-in-the-blank memory test that served as a cued recall task may have provided more information within the sentence, but did not give practice in actual retrieval of the argument

sentences. Thus, a free recall type of test might be a better relearning format in a future study.

Was the Power Sufficient?

In my proposal, I proposed a repeated measures correlation of $r = .4$ and an effect size of $f = .25$ for the interaction for the between and within subjects factor. For a power of .9 I proposed I would need at least 56 data points to examine the idea of testing effect impacting favorability judgments.

The observed correlation of the repeated measures for the favorability judgment was $r\ (58) = .14$, $p = .309$. The observed correlation of the repeated measures for the latency of favorability judgments was $r\ (58) = -.27$, $p = .044$. The observed correlation of the repeated measures for the qualitatively sensitive memory scores was $r\ (58) = .61$, $p < .001$. Compared to the proposed correlation $r = .4$, it makes sense that a marginal testing effect is found for memory but not for the favorability judgments or its latency measures. The effect size for the interaction for the qualitative memory for which I did get a marginal testing effect was only $\eta^2 = .055$ (observed power = .05). Taking together the smaller effect size compared to previous studies and the much lower correlation between measures observed for favorability judgments, the power observed is far below what was proposed. It would not be wise to conclude that the testing effect has no impact on favorability judgments until after a well powered experiment is conducted.

Conclusions and Future Direction

In Studies 1 and 2, the materials were normed to fulfill the purpose of the experiment. Neutral topics were found in Study 1 and marginally persuasive arguments were found in Study 2. In the experiment, there was positive change in favorability judgments but there was no decline in favorability judgments after a delay. Also, being tested on the materials did not seem to have an effect on favorability judgments whereas it did on the memory tests. Nevertheless, being tested on the arguments or merely restudying them aided in the overall increase of favorability judgments in that it can help persuasion. The current study suggests that we base our favorability judgments by utilizing any available cues (which would be observed through the quantity of the recalled arguments) or systematically evaluating the pros and cons that could be of issue for the topic (which would be observed through the quality of the recalled arguments). Thinking back to the example of the election speech rally, it seems that whether you attend to multiple speeches or discuss among yourselves regarding the speech, it will enhance the overall favorability judgment of the speaker.

Previous research investigating memory and judgment have usually tapped on the spacing effect or lag effects. The current study extends this line of literature by looking into the effects of testing. Though with the current procedure and materials, no testing effect was found in the delayed favorability judgments it seems premature to conclude that there is no testing effect in favorability judgments in general. As a start, it would be nice to have a larger sample to obtain a more reasonable power. Given that the

correlations between repeated measurements were quite low on the favorability judgment measures (compared to the memory measures), a larger sample is likely necessary to detect any effects. As mentioned in the general discussion, having the study phase without exposing the topic sentences until their actual time of evaluation could be one avenue. Changing the fill-in-the-blank cued recall memory test to free recall test format during the relearning phase could also be an avenue. Future research can also be conducted to work on the norming of materials to modify some limitations of Studies 1 and 2. For example, changing the wording of the sentences, finding more persuasive arguments, and finding topics that will be relevant and interesting but not thought of much as on-campus topics could be another avenue.

Modern information is more interactive than ever before. Modern advertisements, media, and communication are increasingly two-way and electronic. Therefore, understanding persuasion in the context of testing effect could enrich social communication as well as broaden knowledge of the link between memory and judgment.

REFERENCES

Alba, J . W., Marmorstein, H., & Chattopadhyay, A. (1992). Transitions in preference over time: The effects of Memory on Message Persuasiveness. *Journal of Marketing Research, 29*, 406-416

Albarracin, D. (2002). Cognition in Persuasion: An analysis of information processing in response to persuasive communications. *Advances in Experimental Social Psychology, 34,* 61-130.

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. *New York, NY: Prentice Hall.*

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*, 446-458

Braun-Latour, K. A. & Zaltman, G. (2006). Memory change: An intimate measure of persuasion. *Journal of Advertising Research, 46,* 57-72

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*, 12-21.

Cacioppo, J. T., & Petty, R. E. (1989). Effects of message repetition on argument processing, recall, and persuasion. *Basic and Applied Social Psychology*, *10,* 3-12.

Chan, J. C., McDermott, K. B., Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related

material. *Journal of Experimental Psychology: General, 135,* 553–571.

Chan, J. C., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 431-437.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply criticial, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63-147

Eagly, A., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition: Special Issue: What is an Attitude?,* 582-602

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Finley, J., R., Benjamin, A. s., Hays, M. J., Bjork, R. A., & Norenell, N. (2011). Benefits of accumulating versus diminishing dues in recall. *Journal of Memory and Language, 64*, 289-298.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*, 258-268.

Hinze, S. R. & Wiley, J. (2011). Testing the limits of testing effects using completion

tests. *Memory, 19*, 290-304

Hepler, J. & Albarracin, D. (2013). Attitudes without objects: Evidence for a dispositional attitude, it's measurement, and its consequences. *Journal of Personality and Social Psychology, 104,* 1060-1076.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long term retention. *Journal of Memory and Language, 57*, 151-162.

Karpicke, J. D., & Roediger, H. L. (2008).The critical importance of retrieval for learning. *Science, 319,* 966-968.

Kumkale, G. T. & Albaracín, D. (2004). The sleeper effect in persuasion: A meta-analytic review. *Psychological Bulletin*, *130*, 143-172.

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive validity of the implicit association test in studies of brands, consumer attitudes, and behavior. *Journal of Consumer Psychology, 14*, 405 – 415.

McGuire, W. J. (1968). Personality and attitude change: An information-processing theory. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), Psychological Foundations of attitudes (pp. 171-196). San Diego, CA: Academics Press.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and Persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Roediger, H. L., & Butler, A. (2011). The critical role of retrieval practice in long term retention. *Trends in Cognitive Sciences*, *15*, 20-27.

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

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–233.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1-27.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151-175.

APPENDIX A

TABLES

Table 1. Descriptive statistics (mean and standard deviation) for agreement ratings, relevance to the self, and relevance to other UNCG students of 20 topics.

	Topic	*Agree	*SELF	*UNCG
1	Campus police should be eliminated.	1.97 (±1.69)	4.62 (±2.31)	5.31 (±1.89)
2	Cumulative final exams should be mandatory.	2.60 (±1.30)	4.66 (±2.29)	4.9 (±2.0)
3	Soda vending machines should be banned on campus.	2.37 (±1.99)	4.52 (±1.98)	4.31 (±2.19)
4	Everyone should be required to take at least one exercise course per semester.	3.97 (±1.96)	4.45 (±2.06)	3.97 (±1.64)
5	Air conditioning on campus should be regulated separately in individual classrooms.	4.90 (±1.37)	4.41 (±2.03)	4.41 (±2.00)
6	All electronic equipment should be banned during class.	2.57 (±1.83)	4.24 (±2.21)	4.72 (±2.23)
7	Studying abroad should be abolished.	1.47 (±0.86)	2.86 (±2.2)	4.34 (±2.26)
8	Advertisement slots available between classes should be sold to private companies who need them for campus revenue increases.	2.77 (±1.89)	2.34 (±1.63)	2.59 (±1.88)
9	Surveillance cameras should be installed around campus for security.	4.37 (±1.77)	4.69 (±1.93)	5.28 (±1.91)
10	Students who drive to campus should not be required to pay public transportation fee included in tuition.	5.23 (±1.87)	5.21 (±2.42)	5.52 (±1.45)

Table 1. Descriptive statistics (mean and standard deviation) for agreement ratings, relevance to the self, and relevance to other UNCG students of 20 topics. (Continued)

	Topic	*Agree	*SELF	*UNCG
11	There should be length requirements for short skirts or dresses.	2.23 (±1.33)	3.07 (±2.42)	4.21 (±2.3)
12	Smoking should be restricted to the quad.	3.30 (±2.44)	3.62 (±2.72)	4.38 (±2.09)
13	The cafeteria should limit servings of fried food.	3.77 (±2.01)	4.14 (±2.53)	4.41 (±2.2)
14	Foreign language courses should not be mandatory.	3.87 (±2.45)	5.31 (±2.32)	5.48 (±1.33)
15	Every class should follow the same attendance policy.	3.57 (±2.11)	5.45 (±1.64)	5.07 (±1.93)
16	Male and Female students that are not family should be able to be roommates on campus.	4.23 (±2.39)	3.14 (±2.32)	4.17 (±2.02)
17	Personal water bottles should be mandatory on campus.	2.73 (±1.66)	2.86 (±2.0)	4.03 (±2.06)
18	All classes at the 200 level and above should require either a final project or a term.	3.07 (±1.93)	3.93 (±2.36)	4.389 (±2.26)
19	Portable hand sanitizers should be installed in every classroom.	5.37 (±1.40)	4.79 (±2.26)	4.83 (±1.71)
20	Retaking classes to rectify poor grades should be banned from the University policy.	1.57 (±1.14)	3.83 (±2.52)	4.69 (±2.38)

*Agree: Favorability judgments rated on a 7 point scale, SELF: Self relevance of the topic sentence, UNCG; Relevance of the topic to UNCG campus.
**Topic 4, Topic 9, Topic 13, Topic 14, Topic 15, and Topic 16 were selected to be used in Study 2. ***Topic 4 and Topic 9 were selected to be used in the experiment.

Table 2. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 4: *Everyone should be required to take at least one exercise course per semester.*

	Argument	Mean	(SD)
1.	It can facilitate social interaction b/w different majors.	3.65	(1.70)
2.	Fees for the school can help facilitate the REC center.	3.58	(1.63)
3.*	As a state University, the health of citizens should be a concern of the state.	4.55	(1.96)
4.*	It can lower obesity rates.	5.29	(1.72)
5.*	It can help with stress relief.	5.48	(1.39)
6.*	It can teach people healthy lifestyle habits.	5.32	(1.35)
7.	It can motivate people to exercise.	4.74	(1.67)
8.	It will broaden the students' knowledge on types of exercise.	4.87	(1.68)
9.*	It can broaden the students' knowledge on the benefits of exercise.	5.06	(1.63)
10.	It can facilitate students' accountability of exercise.	3.94	(1.53)
11.*	Strengthening the body strengthens the mind.	5.00	(1.57)
12.*	College is about becoming a well-rounded person: mind, body, and spirit.	5.03	(1.89)
13.	Benefits UNCG to be known as the most health conscious University.	3.48	(1.96)
14.	Build talent/appreciation for UNCG sports teams.	3.04	(1.85)
15.	Increases alertness for other classes.	3.61	(1.89)
16.*	Some courses such as swimming or self-defense might save lives.	5.19	(1.60)

*Arguments that were selected to be used in the experiment.

Table 3. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 9: *Surveillance cameras should be installed around campus.*

	Argument	Mean	(SD)
1.*	Surveillance cameras can identify those involved in crimes on campus.	5.97	(1.11)
2.*	Surveillance cameras would increase the safety of our campus.	5.71	(1.53)
3.	More students would focus on grades instead of illegal activities.	3.10	(2.04)
4.	The campus is getting more dangerous year after year.	4.06	(1.69)
5.*	It will help the police to track robbers instead of letting them walk away.	5.55	(1.48)
6.	This method has been used in London to good effect.	3.23	(1.80)
7.*	Sometimes safety is better than privacy.	3.30	(1.97)
8.*	On campus crime rates will decrease.	4.97	(1.54)
9.*	Speedy response to campus shooting.	5.53	(1.59)
10.	It is cheaper than hiring campus police to walk the beat.	3.35	(1.78)
11.	It will keep people from stealing your bicycles.	3.77	(2.12)
12.	It will fight assaults.	4.35	(1.92)
13.	It will reduce cheating.	2.94	(1.86)
14.*	It's a permanent record for police brutality.	4.47	(2.03)
15.*	Having late night walk buddies isn't enough for safety.	4.42	(1.67)
16.*	The footage won't go public so your privacy will still be protected.	4.45	(1.89)

*Arguments that were selected to be used in the experiment.

Table 4. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 13: ***The cafeteria should limit servings of fried** food.***

	Argument	Mean(SD)	
1.	A good cook relies on herbs and spices or hot oil.	3.23	(1.93)
2.	The chain restaurants on campus should develop healthier menus that can spread off campus.	3.87	(1.96)
3.	University is not like the "real world" as they care for the under-aged minors.	2.97	(1.73)
4.	Frying is a dirty way of cooking.	2.84	(1.66)
5.	Limiting servings of fried food would reduce cholesterol.	4.19	(1.62)
6.	Restricts the intake of unhealthy foods.	3.90	(1.92)
7.	Limiting servings of fried food will reduce the risk of high blood pressure.	4.37	(1.94)
8.	Limiting servings of fried food would allow people to have more self-control when eating.	3.29	(1.75)
9.	Encourages the intake of healthy foods.	3.97	(1.62)
10.	May decrease the obesity rates on campus.	4.23	(1.78)
11.	It can encourage students to experiment with new foods.	4.10	(1.74)
12.	Many kitchen accidents involve hot oil spills.	3.10	(1.97)
13.	People who are health conscious have more options.	3.74	(1.97)
14.	If you want fried foods, you can always get them off campus anyway.	3.48	(2.03)
15.	The school only fries food because it's cheap at the expense of us.	3.61	(2.08)
16.	We're not suggesting we take away the type of food, just alter the preparation method.	4.26	(1.81)

Table 5. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 14: *Foreign language courses should not be mandatory*.

	Argument	Mean	(SD)
1.	It's hard to choose what language to learn because you don't know where you'd visit.	2.97	(1.70)
2.	Thanks to the internet translation tools, we can get by without learning a language.	2.87	(2.01)
3.	In the future, instant translation will make language learning pointless.	2.71	(1.87)
4.	Immersion is a better way to learn a language than courses.	4.90	(1.64)
5.	Foreign languages should be required only for students studying abroad.	3.53	(2.05)
6.	That time can be spent to build their career with more career related courses.	4.00	(2.02)
7.	Most people forget the material from foreign language courses.	4.55	(1.77)
8.	Many jobs don't require multi-linguists.	3.48	(1.71)
9.	Not everyone is going to use a foreign language.	4.16	(1.81)
10.	It's already required at public high schools.	4.23	(1.71)
11.	It is difficult to learn a language after puberty.	3.16	(1.92)
12.	Bilingual students have an unfair advantage over others.	2.74	(2.07)
13.	This is just another example of the government forcing us to do something we don't want to.	2.13	(1.52)
14.	We should focus more on strengthening our English.	3.58	(1.75)
15.	Most people around the world already use English.	3.32	(1.87)
16.	UNCG doesn't provide effective foreign language courses.	2.61	(1.89)

Table 6. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 15: *Every class should follow the same attendance policy.*

	Argument	Mean	(SD)
1.	Some professors don't realize that attendance policies can motivate students.	3.23	(1.93)
2.	It can prevent students from abusing their professors' time.	3.87	(1.96)
3.	It can help the University to manage attendance and help more students graduate.	2.97	(1.73)
4.	Life is unpredictable enough as it is.	2.84	(1.66)
5.	A centralized policy will be a well thought out one compared to those made by just one person.	4.19	(1.62)
6.	Some professors make abusive policies because they think they're more important.	3.90	(1.92)
7.	It would ensure order and organization of the student body in classes.	4.37	(1.94)
8.	There would be no arguments about the attendance policy.	3.29	(1.75)
9.	It would stop the professors from setting unfair policies.	3.97	(1.62)
10.	There will be less confusion for students.	4.23	(1.78)
11.	Every class is equally important so missing any class is equally important.	4.10	(1.74)
12.	If I had to miss a class, it should be up to me.	3.10	(1.97)
13.	Having different policies makes it a burden to keep track.	3.74	(1.97)
14.	Students would take the courses for their contents rather than their policies.	3.48	(2.03)
15.	A centralized attendance policy will give the students input on the class.	3.61	(2.08)
16.	It will reassure parents that their children are attending classes and not wasting the tuition.	4.26	(1.81)

Table 7. Descriptive statistics (mean and standard deviation) of arguments generated for Topic 16: *Male and Female students that are not family should be able to be roommates on campus.*

	Argument	Mean	(SD)
1.	People should be free to choose who they wish to associate with.	4.74	(2.19)
2.	We should stop pretending that sexual relations don't exist on campus.	4.61	(2.19)
3.	Some students basically already live with their significant others.	4.52	(2.05)
4.	In some European countries this is already the case.	2.94	(1.95)
5.	It would show that UNCG has more trust in the students than other schools.	3.71	(2.21)
6.	A policy like this can attract students who might otherwise not have considered enrolling at UNCG.	3.97	(2.17)
7.	Married couples can actually be together and stay on campus.	5.06	(1.93)
8.	Will allow people the opportunity to change and possibly develop deep platonic relationships with the opposite sex.	3.58	(2.14)
9.	Can encourage open mindedness and empathy for people who are unlike you.	4.13	(2.35)
10.	Allows students to make adult decisions about whom they choose to live with.	4.77	(2.11)
11.	Not every student may feel comfortable sleeping in a room with the same sex.	4.42	(2.14)
12.	The current policy is an outdated relic of the last century.	3.77	(2.22)
13.	This is just another form of segregation.	3.23	(2.40)
14.	If homosexuals can room together without problems why can’t heterosexuals be equally responsible?	4.45	(2.46)
15.	They can already be roommates off campus, why restrict on campus?	4.65	(2.36)
16.	Some members of the opposite sex are already so close to you they might as well be family members.	4.03	(2.14)

Table 8. Materials used in the test condition for the experiment presented by topics.

	Topic 4*	Topic 9*
Argument	As a State University, the health of citizens should be a concern of the state.	Surveillance cameras can identify those involved in crimes on campus.
trial 1	As a ____ University, the health of citizens should be a concern of the state.	Surveillance cameras can _____ those involved in crimes on campus.
trial 2	As a ____ University, the health of citizens should be a _____ of the state.	Surveillance cameras can _____ those involved in crimes on _____.
Argument	It can lower obesity rates.	Surveillance cameras increase the safety of our campus.
trial 1	It can _____ obesity rates.	Surveillance _____ increase the safety of our campus.
trial 2	It can _____ _____ rates.	Surveillance _____ increase the _____ of our campus.
Argument	It can teach people healthy lifestyle habits.	It will help the police to track robbers instead of letting them walk away.
trial 1	It can teach people healthy lifestyle _____.	It will help the police to ____ robbers instead of letting them ____ away.
trial 2	It can teach people _____ lifestyle _____.	
Argument	It can broaden the students' knowledge on the benefits of exercise.	On campus crime rates will decrease.
trial 1	It can broaden the students' _____ on the benefits of exercise.	On campus ____ rates will decrease.
trial 2	It can broaden the students' _____ on the _____ of exercise.	On campus ____ rates will _____.
Argument	Strengthening the body strengthens the mind.	Speedy response to campus shooting.
trial 1	Strengthening the body strengthens the ____.	Speedy _____ to campus shooting.
trial 2	Strengthening the body _____ the ____.	_____ _____ to campus shooting.
Argument	College is about becoming a well-rounded person: mind, body, and spirit.	It's a permanent record for police brutality.
trial 1	College is about becoming a ____-_____ person: mind, body, and spirit.	It's a permanent _____ for police brutality.
trial 2	_____ is about becoming a ____-_____ person: mind, body, and spirit.	It's a permanent _____ for police _____.
Argument	Some courses such as swimming or self-defense might save lives.	Having late night buddies is not enough for safety.
trial 1	Some _____ such as swimming or self-defense might save lives.	Having late night buddies is ___ enough for safety.
trial 2	Some _____ such as swimming or self-defense might save _____.	Having late night buddies is ___ enough for _____.
Argument	It can help with stress relief.	The footage won't go public so your privacy will still be protected.
trial 1	It can help with stress ____.	The footage won't go public so your _____ will still be protected.
trial 2	It can help with _____ ____.	The footage won't go public so your _____ will still be _____.

*Topic 4; *Everyone should be required to take at least one exercise course per semester.*
*Topic 9; *Surveillance cameras should be installed around campus.*

Table 9. Frequency of arguments recall and their persuasiveness judgments.

		Frequency **(%)**	Persuasiveness (1 – 7)
Topic 4 At least one exercise course should be required per semester.			
Argument 1	As a State University, the health of citizens should be a concern of the state.	13.54	4.55
Argument 2	It can lower obesity rates.	18.23	5.29
Argument 3	It can teach people healthy lifestyle habits.	7.29	5.32
Argument 4	It can broaden the students' knowledge on the benefits of exercise.	4.17	5.06
Argument 5	Strengthening the body strengthens the mind.	10.94	5.00
Argument 6	College is about becoming a well-rounded person: mind, body, and spirit.*	18.75	5.03
Argument 7	Some courses such as swimming or self-defense might save lives.	16.15	5.19
Argument 8	It can help with stress relief.	10.94	5.48
Topic 9 Surveillance cameras should be installed on campus.			
Argument 1	Surveillance cameras can identify those involved in crimes on campus.	7.02	5.97
Argument 2	Surveillance cameras increase the safety of our campus.	5.26	5.71
Argument 3	It will help the police to track robbers instead of letting them walk away.	14.04	5.55
Argument 4	On campus crime rates will decrease.	15.79	4.97
Argument 5	Speedy response to campus shooting.	12.28	5.53
Argument 6	It's a permanent record for police brutality.	13.60	4.47
Argument 7	Having late night buddies is not enough for safety.	14.04	4.42
Argument 8	The footage won't go public so your privacy will still be protected.*	17.98	4.45

**The most frequently recalled arguments.*

***Frequency; How frequently the arguments were recalled in the experiment.*
Persuasiveness; The mean persuasiveness ratings that were measured in Study 2 on a 7 point Likert scale.

Table 10. Correlations between favorability judgments and relearning format of arguments in the immediate condition.

	attitudeT	attitudeR	attRT_T	attRT_R	qualT	qualR	quanT
attitudeR	-0.005						
attRT_T	-0.195	0.550**					
attRT_R	-0.004	-0.054	-0.369*				
qualT	0.314†	-0.134	-0.481**	0.248			
qualR	-0.011	-0.112	-0.391*	0.157	0.511**		
quanT	-0.073	0.074	-0.049	-0.256	0.319†	0.285	
quanR	0.151	-0.021	-0.351†	0.270	0.852***	0.427*	0.040

NOTE. ***attitudeT***; favorability judgment for the tested arguments, ***attitudeR***; favorability judgment for the restudied arguments, ***attRT_T***; favorability judgment latency measures for the tested arguments, ***attRT_R***; favorability judgment latency measures for the restudied arguments, ***qualT***; qualitatively sensitive memory scores for the tested arguments, ***qualR***; qualitatively sensitive memory scores for the restudied arguments, ***quanT***; quantitatively sensitive memory scores for the tested arguments, ***quanR***; quantitatively sensitive memory scores for the restudied arguments

The degrees of freedom on the correlations was 29.

†*Correlation is marginally significant at the .10 level (2-tailed)*
**Correlation is significant at the .05 level (2-tailed).*
***Correlation is significant at the .01 level (2-tailed).*
****Correlation is significant at the .001 level (2-tailed).*

Table 11. Correlations between favorability judgments and relearning format of arguments in the delay condition.

	attitudeT	attitudeR	attRT_T	attRT_R	qualT	qualR	quanT
attitudeR	0.241						
attRT_T	- 0.209	0.011					
attRT_R	-0.219	-0.284	-0.289				
qualT	0.062	0.202	0.020	0.473**			
qualR	- 0.095	0.011	0.061	0.198	0.425*		
quanT	-0.114	0.153	0.076	0.028	0.247	0.880***	
quanR	0.195	0.161	-0.032	0.467*	0.921***	0.429*	0.232

NOTE. ***attitudeT***; favorability judgment for the tested arguments, ***attitudeR***; favorability judgment for the restudied arguments, ***attRT_T***; favorability judgment latency measures for the tested arguments, ***attRT_R***; favorability judgment latency measures for the restudied arguments, ***qualT***; qualitatively sensitive memory scores for the tested arguments, ***qualR***; qualitatively sensitive memory scores for the restudied arguments, ***quanT***; quantitatively sensitive memory scores for the tested arguments, ***quanR***; quantitatively sensitive memory scores for the restudied arguments

The degrees of freedom on the correlations was 29.

†*Correlation is marginally significant at the .10 level (2-tailed).*
**Correlation is significant at the .05 level (2-tailed).*
***Correlation is significant at the .01 level (2-tailed).*
****Correlation is significant at the .001 level (2-tailed).*

APPENDIX B

FIGURES

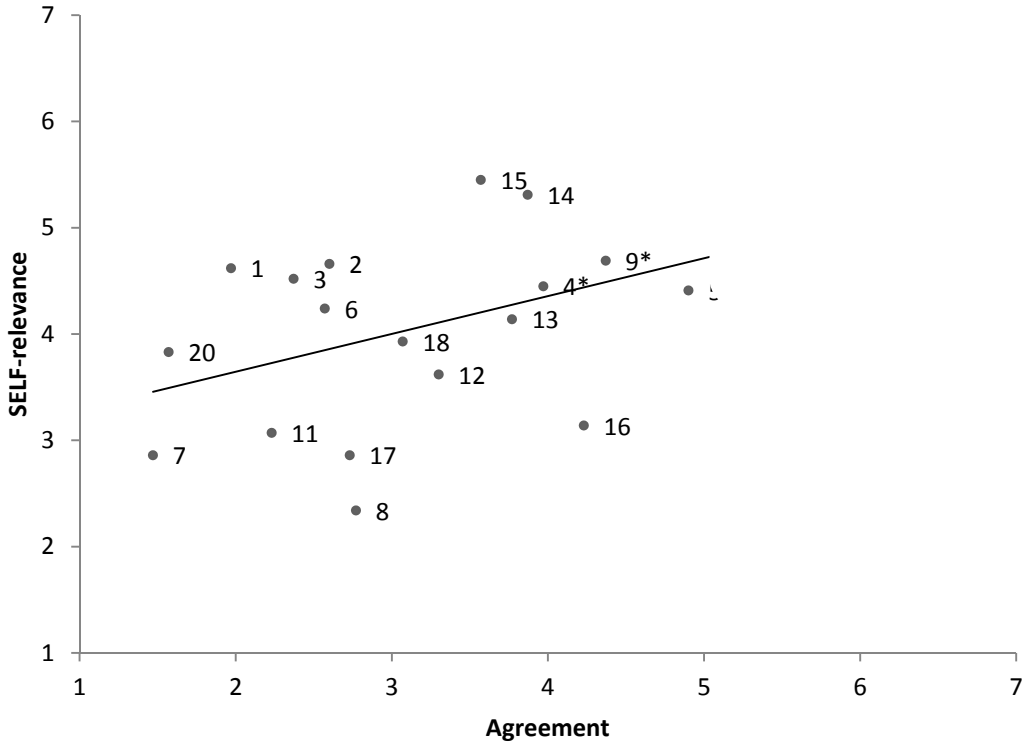

Figure 1. Scatterplot for self-relevance and participants' agreement with the topics.

NOTE: The numbers on the scatter plot are the topic's number. Topics 4* and 9* were used as materials in the experiment.

**Agreement: Favorability judgments rated on a 7 point scale

**SELF-relevance: Self relevance of the topic sentence

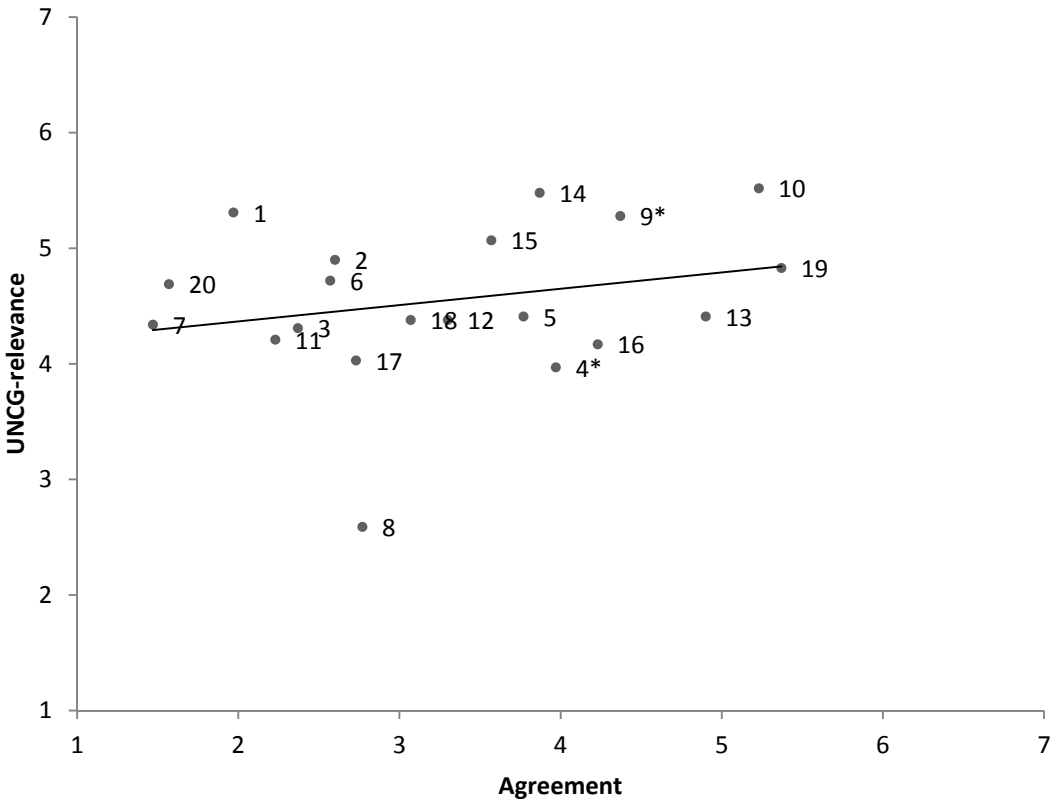

Figure 2. Scatterplot for other student relevance and participants' agreement with the topics.

NOTE: The numbers on the scatter plot are the topic's number. Topics 4* and 9* were used as materials in the experiment.

**Agreement: Favorability judgments rated on a 7 point scale

**SELF-relevance: Self relevance of the topic sentence

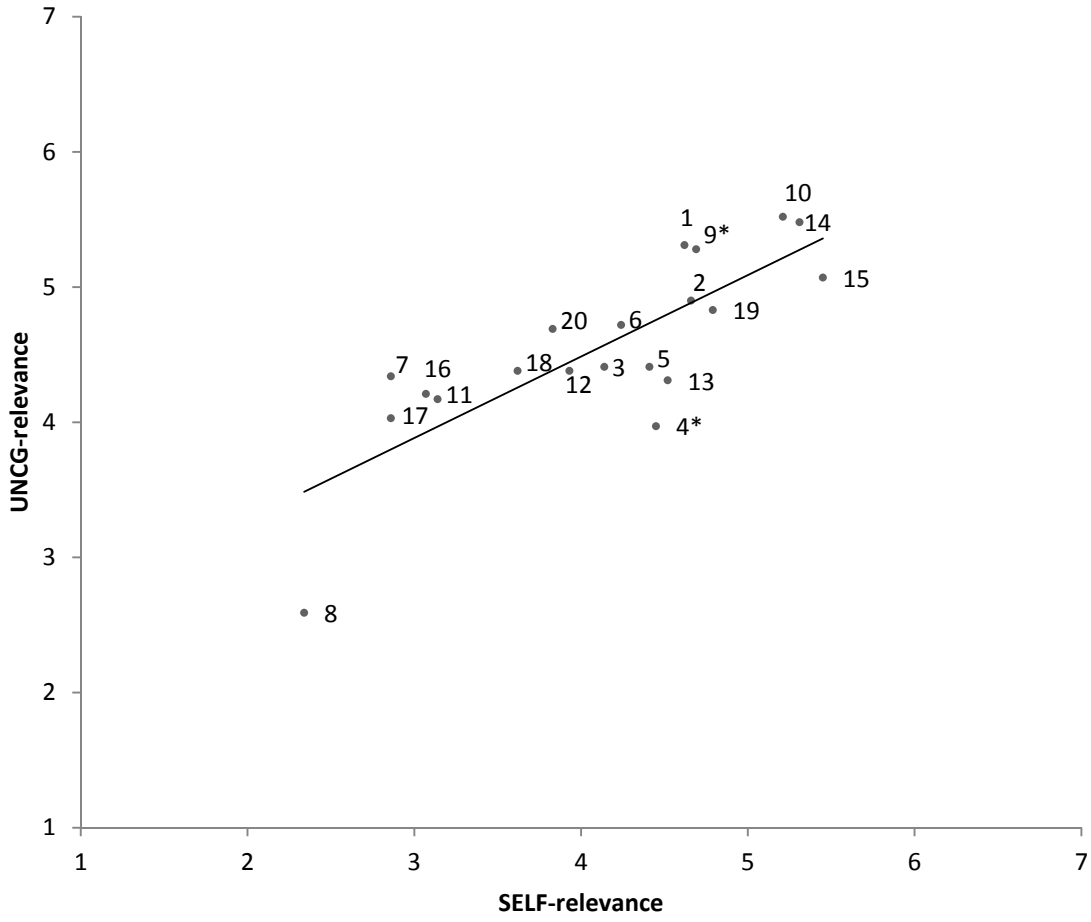

Figure 3. Scatterplot for other student relevance and participants' self-relevance of the topics.

NOTE: The numbers on the scatter plot are the topic's number. Topics 4* and 9* were used as materials in the experiment.

**Agreement: Favorability judgments rated on a 7 point scale

**SELF-relevance: Self relevance of the topic sentence

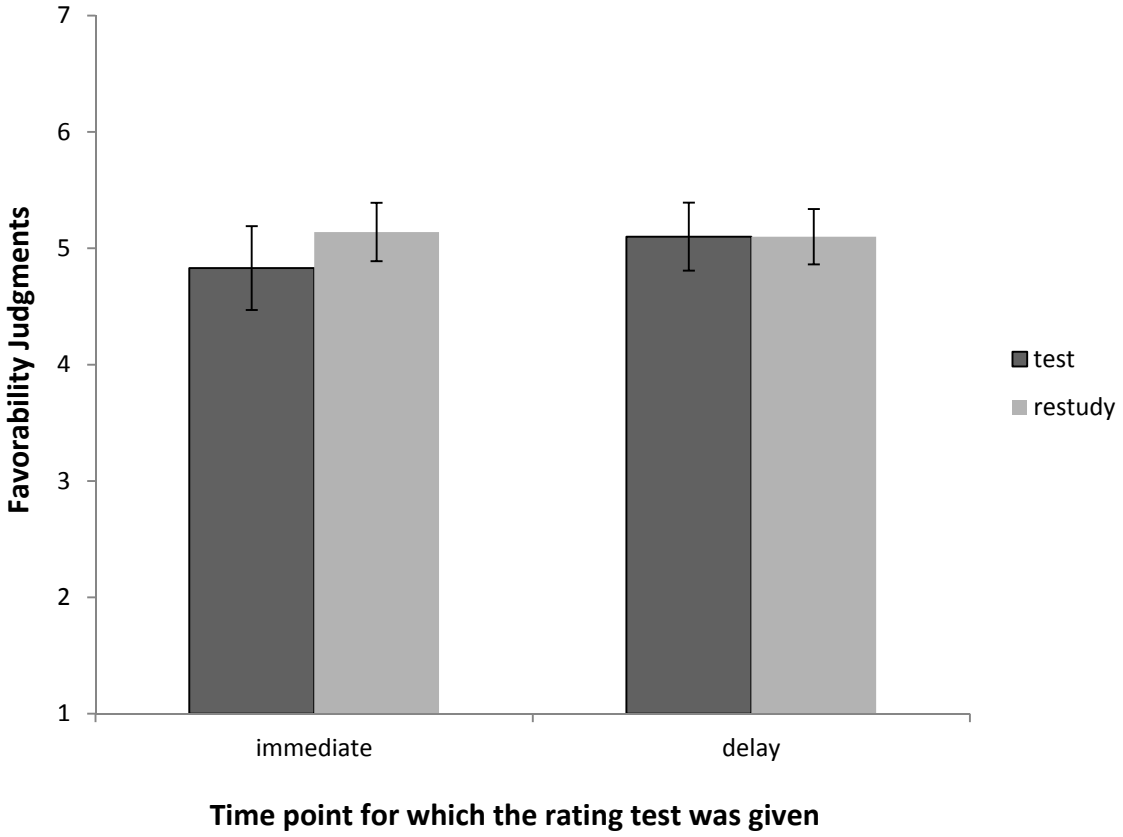

Figure 4. Mean favorability judgments as a function of relearning format and rating test delay. Error bars represent ±*SE*.

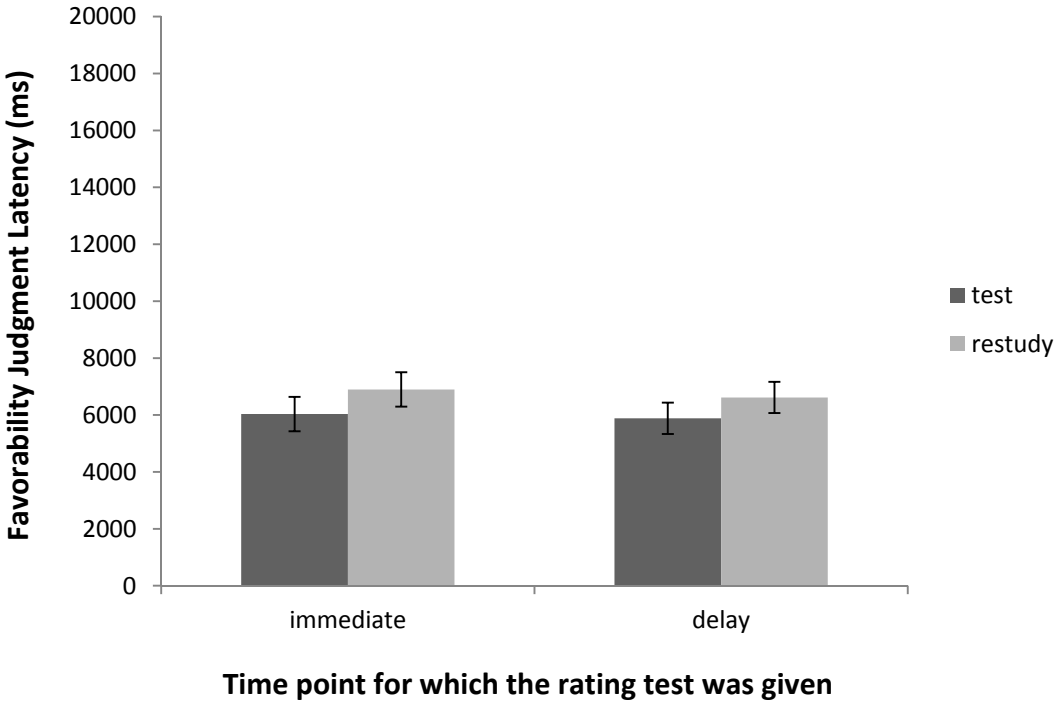

Figure 5. Mean favorability judgment latencies as a function of relearning format and rating test delay. Error bars represent ±*SE*.

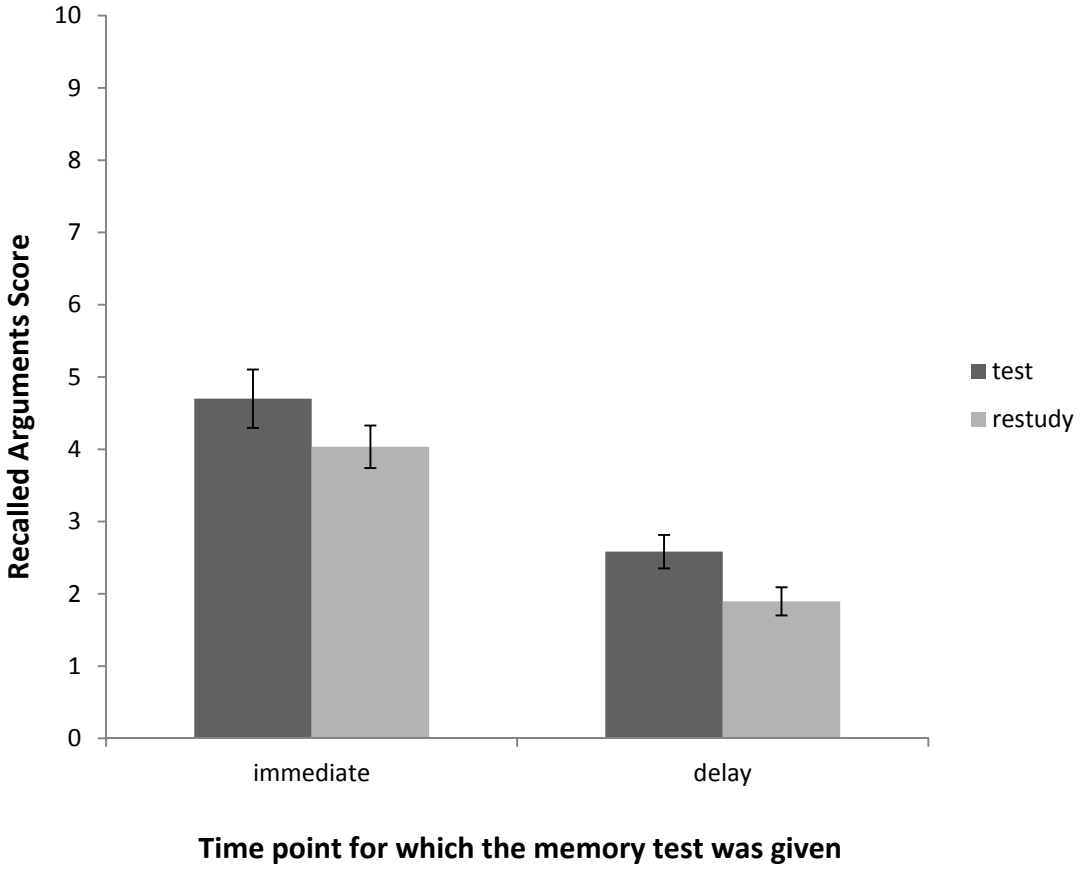

Figure 6. Mean of recalled arguments as a function of relearning format and rating test delay. Error bars represent ±*SE*.

**The two topics were not split in this analysis.*