Replication Project Write-Up

Written by Kat Pleviak

Group Project completed by Kat Pleviak and Lilly Barham

Introduction

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814.

For this replication study our group looked at the article *Creating False Memories: Remembering Words Not Presented in Lists*. The researchers, Roediger and McDermott, wanted to explore people's tendency to create false memories, the experience of remembering events that never actually happened. Drawing inspiration from an earlier study by James Deese (1959), the authors conducted two experiments confirming that participants will often falsely recall or recognize a central word, the "critical lure", that was never shared with participants in the experiment. In both experiments, participants were verbally given word lists in which all items were related to a single critical lure word. The data they collected in both experiments showed, participants remembered this critical word during both the recall phases and the recognition test demonstrating that false memories can happen.

For our replication project we focused on experiment 1 as it was a bit smaller in scope and more manageable to recreate. In this experiment, participants were read six lists of 12 related words and were immediately asked to recall a many of them as they could, starting with the last words read to them and then remembering all the other words they were able to.

After the recall of all six lists, they had a brief 2–3-minute conversation with the experimenter, followed by a recognition test that included 12 studied items, the 6 critical lures, 12 weakly related lures, and 12 unrelated lures. The participants rated each word on a 4-point scale as follows : 4 = Sure it was old (studied), 3 = Probably old, 2 = Probably new, 1 = Sure it was new.

The main independent variables were the kinds or categories of word represented in the recognition test. These were: Studied words, critical lures, weakly related lures, and unrelated lures. There were two dependent variables, one for each part of the experiment. For the first part, the recall phase, the dependent variable was the recall of words participants were able to remember. Researchers took note of both words that were given to the participants as well as

words they remembered that had not been provided, specifically noting if the participant remembered hearing the critical lure. For the second part, the recognition test, the dependent variable was the confidence ratings using a 4-point scale. This showed whether participants we're remembering the words they studied, the critical lure, related words, or unrelated words and how much confidence they had in their belief that the words they remembered had been presented to them previously. Their replication of Deese study was successful in confirming participants falsely recalled the critical lure.

Methods

Power Analysis: N/A for this project.

Planned Sample: For our study we were provided participants through our academic institution, College of DuPage via Professor Grey, using a platform called Prolific in conjunction with Qualtrics. Prolific is a paid service whose purpose is to provide researchers with quality, study, participants. The participants are paid for their time and are screened by Prolific to help ensure quality control. You can also set filters to ensure the participants fulfill the needs of your experiment. Qualtrics is a survey creating platform specifically for researchers. Our team and I came up with the overall format for the survey we created, and I am the person who programed the final version. Our teacher, Professor Grey, was responsible for setting up and running Prolific. I am confident he provided us with an appropriate demographic range for our project and gave us a sample size of 59, close to double the original study's sample size of 36.

Materials:

Computer: The main and most important material we used. I have an Apple Laptop.

Google Drive/Docs: Where we shared documents across the team and were able to update/edit/adjust them, so everyone had the most current data.

Excell: The spread sheet program I used to sort/store our data

Word: This word processing program is used to write reports and helped us to manage data using word find/replace.

ChatGPT/AI: I used this to sort data, create graphs, break down complicated articles, and edit my writing for grammar and spelling. We also used it to write our Spaghetti story for the "break" between the recall and recognition phases of our study.

Zoom: A virtual app used to live chat with video and hold meetings.

Blackboard: A forum for communication at C.O.D. used for classwork and this project.

Qualtrics: A website/application we use to create our online survey.

Prolific/Participants: a paid service we used to provide us with the participants that took our study.

Voice Memos: The app that was used to record the lists of words participants heard during the study. Lily B provided the voice over work.

Original Study: Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814.

Materials Planned vs. As Conducted: For materials in terms of planned versus as conducted they are the same. There was nothing we tried to use that ended up not working and so there was no need to replace anything with a different product or material. I believe historically finding participants was done using websites such as Reddit. My understanding is Prolific is a new service that has only been available as of this semester. But since we had Prolific, we didn't need to use Reddit which I am grateful for. Prolific made getting participants easy and that's something we as researchers had to put little to no effort into doing. Though I'm sure professor Gray did in setting up the parameters for who could take our survey's and in securing the funds so we could take advantage of Prolific.

Procedures:

For our replication of Roediger and McDermott's study we chose to focus on their first experiment. We created a quiz using Qualtrics. Participants logged in and were read 6 lists consisting of 12 words each. The critical lure of each list was chair, sweet, sleep, mountain, rough, and needle. These words were not given to participants. The words of each list were read at a pace of one word per every 1.5 seconds and were prerecorded as audio files. Each list was its own audio file.

After each audio file finished playing, the slide in Qualtrics it was on, moved forward to a blank slide with a box participants could type in and they were instructed to type as many words as they could remember beginning with the most recent words they had heard and then writing as many words as they could remember after that from the list they had just listened to. Participants had 60 seconds to remember as many words from the list as possible. At the end of 60 seconds the slide they were writing on would automatically move forward and the next audio clip would begin.

This recall section of the experiment which contained the 6 lists and 6 recall slides, was all self-automated so once the participants began with the first slide the entire recall part of the experiment ran and there was no opportunity to pause until all six lists and recall slides had been completed.

After the last recall was completed, participants were shown a short one-page story, " The Great Spaghetti Heist" written by AI, and asked to read it.

Then they were shown a list of 42 words (6 were critical lures, 12 were related, 12 were unrelated, and 12 were studied). The words were presented in one long list but were grouped by the audio lists that were presented to participants in the recall section and ordered as followed. The 1st and 6th words of each section were studied words. The 7th was the critical lure. The 2nd-5th words were related and unrelated words and were listed in random orders.

Ideally, we would have presented these 6 lists of 7 words each next to each other forming a word grid. We could not make that happen in Qualtrics, so stacked them creating one continuous long list.

Participants were asked to select all the words they thought had been presented to them previously during the recall section.

Procedures Planned vs. As Conducted:

Planned	As Conducted
1) Originally, I wanted to do the second study because it seemed more interesting. But Professor Gray pointed out it was going to be way more work and be cost prohibitive for Prolific, because of how long the survey would take. Also, that participants are less likely to stick with a survey that takes too long. Furthermore, this was our first replication experiment so there was no need to overcomplicate things.	We chose experiment 1 and then attempted to not overcomplicate things. Jury is out as to whether we succeeded in doing so.
2) I wanted to choose word lists that had words that I thought were interesting, but Lily B noticed a detail in the article that was not listed on the word list index. The article did list the six lists that had been used in experiment 1.	We use the same lists that the researchers used in experiment 1.
3) Lilly B offered to make recordings of the audio for the project and after she made them, she emailed me the files. We knew the words were supposed to be read 1.5 seconds apart. When I listened to the files the words felt to fast, so I asked Lily how she timed it. She had used a timer on her voice memo app but when we checked	Lilly rerecorded the audio files using her watch as the timer, so we got them to be the correct pace of 1.5 seconds per word.

realized the timer was not using seconds as its marker.	
4) In class we started making our survey together and in Qualtrics you can group questions in a single box or put one box per question. I ended up doing a lot of work on the programming of our survey and when I went in to work on the automation, I noticed that we had built it by putting all the questions in a single box. I was not confident that we would be able to make the automation work with this method and I found it made reordering things very difficult.	I created individual boxes for each question which allowed for easier copying, pasting, and repeating of questions through out the servey. It made the questions easier to move around and manipulate.
5) Originally, we did not think it was possible to make audio automatically play in Qualtrics. So, our plan was to add a timer to each card that had a question on it and add 10 seconds to that time. This would give participants 10 seconds to hit the play button to hear all the audio. It was a good fix for something we didn't believe was possible. But when I tested it myself, I didn't hit the play button in time. I got distracted reading the instructions and when I didn't hit the play button in time the card moved forward cutting off the audio and I didn't get to hear the entire list. I knew this would fundamentally make our data unusable since it was a memory test, and we couldn't count on memory data if people hadn't heard the entire list. To solve this problem, I googled if making audio self-play was possible and found out it could be done if you edited the computer code.	I was able to successfully edit the computer code, and we were able to make all our audio and recall slides self-automate so we know every participant was able to 1) hear the full list, 2) was not able to play it twice, and 3) could not be writing the words down while they were hearing them (at least in the computer). So, I believe we set up our survey to give us the most reliable data possible. YEAH!!!!!!!!!!!!!!
6) When we made our long list for the recognition part of the experiment, we knew the original format had been a grid which we could not replicate. So, we decided to make one long list but had not observed that there was an	We reordered our long list to match the format of the original experiment, so all our lists meet those same criteria even though they had to be presented in a different way.

	order to which the words had been placed within individual lists by topic. While reviewing how they had created the grid I noticed that specific categories of words (critical lures, studied words etc…) were listed in certain places, for example the critical lure was always the seventh word, and the studied words were first and 6th.	

Analysis Plan: Our analysis plan breaks down into two sections one for each part of the experiment.

The first is the recall section in which we are looking at three variables and the participants memory or recall of them. These variables were the studies words, the critical lures, and other intrusions or words participants remembered that had not been included as part of the studied word list.

We tallied the word count for each of the studied words and the critical lure's and then found the percentage of times each individual word was recalled. Then we averaged the percentages of all the words from each list, (chair, sweet, sleep, mountain, rough, and needle) and then averaged those 6 percentages together to find the percentage of times participants remembered the studied words.

Next, we tallied the word count on each critical lure word and got the percentage of times people recalled those and then averaged the 6 critical lure percentages together to find the overall average for people recalling the critical lures.

Finally, we added together all the intrusions and divided that number by participants to get a percentage of times participants recalled intrusions.

For the second part, the recognition section, we were looking to see if participants remembered the critical lure. We were able to measure this by tallying the word count of the word's participants selected and then finding the percentage of times each critical lure was recalled. Then we could average those percentages together to find the percentages of times critical lures was recalled during the recognition test. The data we collected also allowed us to contrast the recognition of the critical lure with the recognition of studied words, related words, and unrelated words.

Analysis Planned vs. As Conducted: We worked on this together in class, with the help of Professor Grey so there was not much of note that we planned different than what was finally conducted. The only issue that came up involved intrusions during the recall part of the experiment and our lack of ability to figure out what was done in the original experiment. But that will be addressed further in our results.

Differences from the Original Study:

Original Study	Our Replication
Had 36 participants.	We had 59, almost double the original study's number.
All their participants in the original study were students in the same class studying memory.	Our participants were people actively looking to take surveys for money, so did not know each other or share a background, other than wanting to take surveys for money.
Word lists were read to participants by a person indicating they were in the same controlled space.	Our word lists were prerecorded and delivered via online survey. So, participants were unmonitored and were in unknow locations to us. They could be anywhere with working Wi-Fi
Since the lists were read there were opportunities for participants to ask questions or stop the survey if needed.	Our survey was automated during the recall part so once the first list began everything auto ran until all 6 list and recall timings were done. There was no chance to pause, get distracted, or ask questions.
Participants wrote their responses on paper.	Participants recorded their responses digitally.
Participants were told not to guess words during the recall phase of the experiment.	We did not specifically tell the participants not to guess. We only told them to remember as many words as they could, starting with the last words they heard and then remembering everything else they could.
For the break between recall and recognition the experimenter had a conversation with the participant.	For the break between recall and recognition the participant read a 1-page story we provided called "The Great Spaghetti Heist". Problem
In the recall section participants were asked to rank all the words listed from 1-4 evaluating their confidence level as to whether they believed the word had been presented to them in one of the previous lists. The lists were also presented in a grid like presentation where the six lists of seven words each were all listed one next to each other.	In our recall section we provided one long continuous list of words and ask participants to select the ones they thought they remembered from previous lists.

Actual Sample: For our study we were provided participants through our academic institution, College of DuPage via Professor Grey, using a platform called Prolific in conjunction with Qualtrics. Prolific is a paid service whose purpose is to provide researchers with quality, study, participants. The participants are paid for their time and are screened by Prolific to help ensure quality control. You can also set filters to ensure the participants fulfill the needs of your

experiment. Qualtrics is a survey creating platform specifically for researchers. Our team and I came up with the overall format for the survey we created, and I am the person who programed the final version. Our teacher, Professor Grey, was responsible for setting up and running Prolific. I am confident he provided us with an appropriate demographic range for our project and gave us a sample size of 59, close to double the original study's sample size of 36.

Differences from Pre-data collection Plan: None

Results

The Recall

Our replication study somewhat mirrored the results of Roediger and McDermott, in the sense that our data came up with the same results conceptually. But our statistics did not match and were lower than the values reached in the study we were replicating for studied words and critical lures.

For unrelated words our numbers were significantly higher. We were not able to figure out how Roediger and McDermott calculated their data and so assume this has something to do with the large discrepancy in our statistics. To try to figure it out we did take one of our categories, sleep, and rather than counting the individual intrusions we noted how many participants wrote an intrusion in that category. We found 17 out of 59 participants wrote intrusions during the sleep section of the test which came out to 28%, far lower than the average of 74.09% which we came to buy adding up all the intrusions by category and dividing by number of participants. But still 28% is double the 14% of the original study and is higher while the other two statistics, studied words and the critical lures came out lower.

Our Study Numbers for Recall

	The average of times studied words were remembered per list.	How many times the critical lure was written.	% of participants who write the critical lure.	How many unrelated words were written per list	% of participants who wrote an unrelated word.
Chair	41.81%	16	27.12%	53	89.83%
Sweet	56.07%	12	20.34%	45	76.72%
Sleep	58.33%	11	18.64%	49	83.05%
Rough	47.82%	10	16.95%	23	38.98%
Mountain	45.20%	5	8.47%	50	84.75%
Needle	71.19%	10	16.95%	42	71.19%

Average	**50.30%**		**18.08%**		**74.09%**

Comparison Between Our Recall and Roediger and McDermott Study Numbers

	The average of times studied words were remembered per list.	% of participants who write the critical lure.	% of participants who wrote an unrelated word.
Our Study	**50.30%**	**18.08%**	**74.09%**
R and M	**65.00%**	**40.00%**	**14%**

The Replication.

It's a little bit more difficult to compare our results to the original study because our study just evaluated whether participants remembered the word or not, we did not rank it with a scale. However using binary data, by allowing the participants to only select the words they felt they remembered, makes our data easy to sort and if we combine the 3 and 4 categories, as well as the 1 and 2 categories of the original study we can break it down into remembered or not remembered. That turns there data binary and places our data on a similar scale so we can compare it. When we do that, this is what it looks like.

Comparison Between Our Recognition and Roediger and McDermott Study Numbers

Category	**Our Study (% Recognized)**	**Original Study (R and M)**	**Notes**
Studied Words	73.0%	86% (rated 3 or 4)	Very similar results. Shows successful recognition of studied items in both cases.
Critical Lures	57.9%	84% false alarm rate	Our false alarm rate is lower but still high, showing strong false memory effect.
Related Lures	19.9%	21% false alarm rate	So close!!! Related lures were mistaken at similar rates.
Unrelated Lures	9.3%	2% false alarm rate	Our rate is higher, possibly because we used binary recognition (no rating scale). The participants had 2 less choices.

What this shows overall is that our study does successfully reproduce the same patterns of false memories seen in the original study even though our percentages in some places are very different. In terms of studied and related words our statistics are close. For critical lures our statistic comes out lower at 57.9% as compared to their 84%. This is a notable difference but still shows a high level of recognition for the critical lure, supporting the finding of the original study. For unrelated words we came up with 9.3% while the original study only produced a 2% false alarm rate. This is significant and worthy of note.

Discussion:

Yes, I think this replication is successful in terms of the data it produced. Though our statistics did not always match the original study they still supported it's findings. Those being that people are prone to fabricating memories. I do not believe our study had any significant problems, however there were differences in our replication from the original study which made it not exact. We learned in class that this is acceptable, and I believe the realities of the world make that true. However, after experiencing this analysis, I think it will always be my choice in the future to be as exact as possible when doing a replication. If we really want to know if something replicates, I think being as identical as possible will give us the most accurate picture.