A Two-Effects Model of Explanation on Exposing the Illusion of Understanding

by

Ethan Andrew Meyers

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2024

© Ethan Andrew Meyers

Examining Committee Membership

External Examiner	Dr. Philip Fernbach Associate Professor Marketing Division University of Colorado at Boulder
Supervisor(s)	Dr. Jonathan Fugelsang Professor Department of Psychology University of Waterloo Dr. Derek Koehler Professor Department of Psychology University of Waterloo
Internal Members	Dr. Sam Johnson Assistant Professor Department of Psychology University of Waterloo Dr. Evan Risko Associate Professor Department of Psychology University of Waterloo
Internal-external Member	Dr. Selcuk Onay Associate Professor Department of Management Science and Engineering University of Waterloo

Author’s Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The work in Chapter 1 was reported in a journal article in *Judgment and Decision Making* and features contributions from the authors: Jeremy Gretton, Joshua Budge, Jonathan Fugelsang, and Derek Koehler. While I led the project in terms of conceptualization, design, execution, and writing, the work would not have come to be without the efforts of these team members.

I am solely responsible for the work presented in the other chapters of this thesis.

Citations:

Chapter 1:

A version of the research presented in Chapter 1 appears in:

Meyers, E. A., Gretton, J. D., Budge. J. R. C., Fugelsang, J. A., & Koehler, D. J., (2023). Broad effects of shallow understanding: Explaining an unrelated phenomenon exposes the illusion of explanatory depth. *Judgment and Decision Making*, *14*(6), 658-670.

Abstract

People often overestimate their understanding of how things work. For instance, people believe that they can explain even ordinary phenomena such as the operation of zippers and speedometers in greater depth than they really can. This is called the illusion of understanding (originally known as the illusion of explanatory depth). Fortunately, a person can expose the illusion by attempting to generate a causal explanation for how the phenomenon operates (e.g., how a zipper works). This might be because explanation makes salient the gaps in a person's knowledge of that phenomenon. However, recent evidence suggests that people might be able to expose the illusion by instead explaining a different phenomenon. Across six preregistered experiments and one secondary data analysis, I examined whether explaining one phenomenon (e.g., how a zipper works) leads individuals to lower their self-assessed knowledge of unrelated phenomena (e.g., how snow forms). My findings demonstrated that participants consistently revised their understanding downwards, not only for the item they explained but also for other items they did not explain. For instance, participants reported reduced understanding of speedometers after explaining helicopters or zippers. Contrary to prior research, participants did exhibit the illusion for familiar movie plots (Experiment 4), but consistent with prior research, participants did not exhibit the illusion for common procedures (Experiment 5). Additionally, when common procedures were included in the experimental design used in Experiments 2 and 3, participants showed no illusion whatsoever (Experiment 6). Finally, an analysis of explanation quality using ChatGPT to code the explanations revealed that the reduction in perceived understanding after explaining (compared to before) correlated with the difference between how well the participant thought they understood the item and how well they actually explained it, but only for explained items. These findings challenge the common framework of how the illusion of understanding operates. Throughout the thesis I evaluate alternative models of the illusion and ultimately find the most support for a two-effects model of explanation, wherein failing to explain a phenomenon temporarily makes people recognize the gaps in their knowledge of the item they explained and makes them feel less knowledgeable about most other things.

Acknowledgements

I'd like to thank my supervisors, Derek Koehler and Jonathan Fugelsang, for the immense intellectual freedom they provided me. It gave me plenty of room to make mistakes (which I am still doing), but also to pursue what fascinates me. I am and will be forever grateful for that. They are two of the kindest and most thoughtful individuals I've met. It was an absolute pleasure to be supervised by them.

To my lab mates and friends in other labs, thanks for the coffee breaks, which sometimes bordered on excessive, your collective problem-solving, timely pieces of advice, and for being great people to share a beer (or two) with.

I am deeply grateful to Ori Friedman, Igor Grossman, Michał Białek, Richard Eibach and David Mandel, who, while not my official supervisors, have acted in that capacity throughout my graduate journey. Your support, mentorship, and challenges have been instrumental.

Dedication

This thesis is dedicated to my mother.

Table of Contents

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xi

Chapter 1 Broad effects of shallow understanding 1

1.1 Experiment One 6

1.1.1 Methods 6

1.1.2 Results 12

1.1.3 Discussion 17

1.2 Experiment 18

1.2.1 Methods 18

1.2.2 Results 20

1.2.3 Discussion 21

1.3 Experiment 3 22

1.3.1 Methods 22

1.3.2 Results 22

1.3.3 Discussion 24

Chapter 2 On the boundaries the illusion of understanding 28

2.1 Experiment 4 28

2.1.1 Methods 29

2.1.2 Results 32
2.1.3 Discussion 34
2.2 Experiment 5 35
2.2.1 Methods 35
2.2.2 Results 38
2.2.3 Discussion 40
2.3 Experiment 6 41
2.3.1 Methods 42
2.3.2 Results 43
2.3.3 Discussion 46
Chapter 3 Evidence of two distinct effects of explanation 48
3.1 Secondary Data Analysis of Reported Experiments 48
3.1.1 Methods 49
3.1.2 Results 54
3.1.3 Discussion 58
Chapter 4 Summary and Conclusions 60
4.1 General Discussion 60
4.2 Relevance of the two-effects of explanation model of the illusion of understanding 65
4.2.1 Additional insights of this research 67
4.3 Limitations and Conclusion 69
References 70
Appendix A Experiment 1 Supplemental Analyses 75
4.3.1 Confirmatory Analyses 75
4.3.2 Similarity Analyses 75

Appendix B Experiment 2 Supplemental Analyses 78

4.3.4 Confirmatory Analyses 78

Appendix C Experiment 3 Supplemental Analyses 79

4.3.5 Confirmatory Analyses 79

Appendix D Experiment 7 Supplemental Analyses 81

List of Figures

Figure 1: The differences in understanding ratings for each Phenomenon Rated by Phenomenon Explained for Experiment 1.....14

Figure 2: The differences in understanding ratings for each Phenomenon Rated by Phenomenon Explained for Experiment 2.....20

Figure 3: The differences in understanding ratings for each Phenomenon Rated by Phenomenon Explained for Experiment 3.....23

Figure 4: Change in understanding from Time 1 to Time 2 by Narrative, split by Condition.....33

Figure 5: Change in understanding from Time 1 to Time 2 by Procedure, split by Condition.....39

Figure 6: Mean understanding ratings by Condition and Time by item of Experiment 6 45

Figure 7: Distribution of explanation quality and understanding ratings over time, split by item from Experiment 2.....55

List of Tables

Table 1: Participant Demographics……..…….……7

Table 2: Primary Task Instructions…….……9

Table 3: Stimuli used in Experiments 1 – 3 and 6…..……...11

Table 4: Summary statistics of understanding ratings by condition for each Experiment…..13

Table 5: Mean (SD) of Similarity ratings between all device pairs..……..……...16

Table 6: Comparison of movie stimuli between original and current experiment……31

Table 7: Procedures stimuli used in current experiment and in Rozenblit & Keil (2002)..…47

Table 8: Sample ChatGPT-4 rubric explanations and evaluation instructions…..……..51

Table 9: correlations with ChatGPT coded explanation quality across each experiment…..57

Chapter 1
Broad effects of shallow understanding

People have a natural tendency to try to understand the world around them, and for good reason. Having a functional model of how various systems operate can lead to improved material outcomes. This awareness can be as minor as knowing how to make a pot of coffee, or as major as knowing which foods are nauseating - or even deadly. In all sorts of situations, it pays to know. However, people often act on what they think they know, which is imperfectly correlated with what they actually know. It is arguably better for a person to be calibrated and recognize what they don't know, than to be miscalibrated and think they know more than they do. For example, it is better to admit to not knowing how to use a coffeemaker than to cause the machine to malfunction. Similarly, it is better to skip a meal than to end up in the emergency room - or morgue. While it pays to know, it can pay more to know what one knows. This reality makes overconfidence particularly consequential and suggests that people should strive to maximize knowledge calibration. Yet, people seem readily equipped to be miscalibrated: to fool themselves into thinking they understand the world more than they really do. One example, and the focus of this thesis, is the illusion of explanatory depth.

The illusion of explanatory depth, also known as the illusion of understanding, refers to the phenomenon in which a person falsely believes they have a better understanding of a topic than they actually do (Rozenblit & Keil, 2002). This phenomenon applies to various kinds of topics, including mechanical devices like toilets and zippers, and natural phenomena such as how snow forms and how earthquakes occur. However, certain topics, such as knowledge of the world's capital cities, seem to be immune to this illusion (Mills & Keil, 2004; Rozenblit & Keil, 2002). The existence of these exceptions distinguishes the illusion from general overconfidence (see Moore & Healy, 2008 for a review of overconfidence).

When a person holds an illusion of understanding, they can expose it by attempting to explain how the target phenomenon (of the illusion) works. For instance, if a person overestimates their knowledge of zippers, then having them explain how zippers work may lead them to reduce their estimate. This might be because the act of explaining forces the person to recognize the gaps in their knowledge of what they are explaining (Mills & Keil, 2004; Rozenblit & Keil, 2002; Vitriol & Marsh, 2018). In turn, this constrains the person’s tendency to overestimate knowledge. Consider how

the illusion of understanding is demonstrated empirically. In a standard procedure, people rate their knowledge of some phenomenon (e.g., how a zipper works), then attempt to explain how the phenomenon works (e.g., how a zipper attaches two pieces of material together), and then re-rate their knowledge of the phenomenon. The observed effect is that understanding ratings are lower after explaining compared to before explaining.

There are several proposed mechanisms for why people tend to subjectively overestimate their knowledge of a phenomenon. I will address only some. When information is socially desirable to know, people are more likely to overestimate the degree to which they know it (Gaviria et al., 2017; cf. Rozenblit & Keil, 2002). For example, people overestimate their knowledge of how inflation works more than their knowledge of how the stock market works, as knowledge of how inflation works is judged to be more socially desirable (Gaviria & Corredor, 2021). People may overestimate their understanding of a device, on the other hand, because they overfocus on its visible parts and neglect its hidden parts or because they mistake knowing what an object does for knowing how it works (Rozenblit & Keil, 2002). Construing an object abstractly rather than concretely may also increase a person's initial knowledge estimate as the comparatively lower fidelity of the abstract construal is likely to obscure gaps in their knowledge (Alter et al., 2010). Alternatively, people might overestimate their knowledge because they mistake having access to knowledge (e.g., having a cellphone with access to the internet) with holding knowledge (Rabb et al., 2019). As is clear from the possible mechanisms noted above, the illusion is complex, and the variety of mechanisms that contribute to it underscores its intricacy.

Despite not definitively understanding how it works mechanistically, researchers continue to examine the consequences of exposing an illusion of understanding (e.g., Cadario et al., 2021; Crawford & Ruscio, 2021; Littrell et al., 2022; Meyers et al., 2020; Sloman & Vives, 2022). This is likely based on the intuitive plausibility of the assumptions underlying the illusion. One assumption, noted above, is that explaining makes people recognize the gaps in their knowledge of the phenomenon they are explaining, and this exposes the illusion of understanding. This claim is often found in the literature. For example, Vitriol and Marsh (2018) state that "… trying to explain a phenomenon reveals to participants how little they actually understand about the workings of that phenomenon…" and Mills and Keil (2004) note that "…when explaining how this might work, the participant recognizes a lack of awareness of the mechanism". The basic idea is that explaining

something makes us recognize that our knowledge of that thing is not as deep as previously thought. I term this the "specificity principle".

The specificity principle has high face validity. The reader can probably recall a time when they suddenly felt less knowledgeable of a topic after trying to explain it. However, the principle also raises an interesting question: Is the illusion only exposed by explaining the specific phenomenon for which the illusion is held? The principle suggests the answer is yes, unless we expect people to generalize their experience of being ignorant about one phenomenon to another. This potential generalization struck me as unlikely. For example, after failing to explain how a zipper works, we would expect a person to question their knowledge of zippers but not their knowledge of natural events (and vice versa). Also, people are often unaware of their biases (e.g., Pronin, et al., 2002) and might not generalize perceptions of bias without that bias being explicitly described as broad or pervasive (Gretton, 2017). Generalization would also appear to be inconsistent with the dearth of robust far-transfer effects in which a person's training in one area, such as chess, supposedly increases their capabilities in other areas, like working memory or concentration (Detterman, 1993; Kassai et al., 2019; Sala & Gobet, 2017; cf. Bart, 2014). In other words, people are not very good at transferring skills and knowledge from one context to another. This casts further doubt on people's ability to correct their overestimation of one phenomenon after explaining a different one.

Recent work from my lab suggests that people may indeed transfer their feelings of ignorance to a phenomenon other than the one explained. In this work, we examined the downstream psychological consequences of exposing an illusion of understanding (Meyers et al., 2020). We demonstrated that people revise their opinions on economic issues more to professional economists than to random members of the public only after failing to explain how the issue worked (compared to when no explanation was required). For example, failing to explain how trading with China affects the U.S. economy led to greater revision of opinion on the issue when provided the consensus among professional economists than when provided identical consensus among members of the public. Counterintuitively, we also found the same pattern when participants had previously been asked to explain completely unrelated phenomena such as how recycling works in a modern U.S. city and how a helicopter takes flight. That is, regardless of what was explained, explaining led people to revise more to experts than the public. This suggests the possible existence of a "breadth principle" wherein the illusion of understanding can be exposed by explaining something other than the target of the illusion, contra the "specificity principle".

Empirical research on learning provides additional evidence supporting the possibility of a breadth principle. Explaining a topic is often a more effective way to improve learning than other methods, such as reading the material twice (Chi et al., 1994; Wong et al., 2002; Williams & Lombrozo, 2010a). This is in part because explanation promotes generalization, as it motivates a person to identify the underlying structure that unifies phenomena (Williams et al., 2010b). Participants who explained why an item may belong to a certain category were more likely to detect the subtle rules that determine category membership than those who provided a careful description of the item (Williams & Lombrozo, 2010b). In a category learning task, explanation increases the chance of discovering the rules of category membership. However, in an illusion of understanding task, generalization may increase the likelihood of recognizing the overestimation of one's knowledge. Therefore, when attempting to explain a phenomenon, a person may generalize their sense of ignorance about that phenomenon to their general knowledge (i.e., questioning knowledge of one phenomenon transfers to other phenomena). Although this mechanism does not favor one principle over the other, it provides a mechanistic explanation for how a breadth principle could operate.

If the illusion of understanding is based on the breadth principle rather than the specificity principle, the implications extend beyond the immediate literature. For instance, if a student fails to explain a concept in one subject, such as Chemistry, and generalizes their sense of ignorance to their overall knowledge, they may be more open to engaging with unrelated subjects, like Philosophy, at least temporarily. This could be particularly useful in educational settings where reducing overconfidence can promote a more balanced and engaged approach to learning (Meyers et al., 2020). Moreover, fostering intellectual humility through explanation could mitigate the impact of misinformation and improve decision-making in various domains (Lewandowsky et al., 2012; Pennycook et al., 2018).

Furthermore, the coexistence of both the specificity and breadth principles raises intriguing possibilities. If explaining a concept makes a person realize their lack of knowledge specifically about that concept (specificity principle) while also inducing a broader intellectual humility (breadth principle), we are looking at a dual mechanism of cognitive debiasing. This dual mechanism preserves the uniqueness of the illusion of understanding, maintaining that it operates under the specificity principle, while simultaneously recognizing the additional benefit of the breadth effect.

The importance of distinguishing between these principles is important. If the illusion of understanding is merely a specific instance of general overconfidence, then focusing exclusively on it might lead to inefficient research and misallocation of resources. Conversely, if there are indeed two distinct effects, understanding this dual mechanism could significantly enhance theoretical models and practical applications. By delineating the conditions under which each principle operates, we can develop more targeted interventions for improving judgment and decision-making across various domains. This would not only validate the continued study of the illusion as a unique phenomenon but also underscore the broader utility of explanation as a powerful tool for cognitive debiasing.

Given the promising benefits of establishing a breadth principle for education and belief change and our previous observation of similar downstream consequences (e.g., opinion revision to expert consensus) when exposing both topic-relevant and -irrelevant illusions, I aimed to systematically test whether similar upstream consequences (e.g., reporting less understanding of a phenomenon than originally believed but did not explain) would also occur. One simple way to test this is to have participants re-evaluate their understanding of items they previously rated but did not explain. To my knowledge, this had not been attempted before.

After completing this work, I became aware of an unpublished manuscript by Roeder and Nelson (2015), who explored the same question. They conducted three experiments with methods similar to mine (that are reported in Chapter One), especially their Experiment 2, and found that people reported knowing less about one phenomenon after attempting to explain another. For instance, explaining how a national flat tax might work led to a reduction in reported understanding of how to institute merit-based pay for teachers. However, they found mixed results as to whether explaining the same phenomenon or explaining a different one is more effective for exposing the illusion. One limitation of their work is that their experiments did not include a no-explanation control condition. Although they could compare participants who explained the same or different phenomena, they could not compare them to those who did not explain anything. It is possible that not explaining anything but having to provide two ratings would lead to a reduction in reported understanding due to conversational norms (e.g., "I exaggerated my first rating of understanding"; Schwarz & Bless, 1992). Research on the initial elevation bias also suggests that ratings on subjective measures are often higher when first assessed compared with subsequent re-ratings, even without an intervention (Shrout et al., 2018). The first chapter of this thesis primarily Experiments 2 and 3, which included a no-explanation comparison condition, thus extends the work of Roeder and Nelson

(2015). I therefore believe my work represents the strongest test of whether attempting to explain a phenomenon can lead to reporting knowing less about another. I will further discuss Roeder and Nelson (2015) in the General Discussion.

1.1 Experiment One

1.1.1 Methods

The materials, data, and analysis script for each experiment are available on the Open Science Framework (OSF) website (https://osf.io/6huwq/). I preregistered this experiment. The preregistration can be found on the OSF (https://osf.io/jvfhe). All experiments reported in this manuscript received ethics clearance by the University of Waterloo's Office of Research Ethics (ORE#43203).

1.1.1.1 Participants

Following my preregistered plan to obtain approximately 90% power to detect a within-subjects effect of $d = 0.2$, I recruited 240 participants via Mechanical Turk. The participants were residents of the U.S. or Canada, had a HIT approval rating of at least 99%, were over the age of 18, and passed a two-question prescreening questionnaire prior to beginning the experiment (Littrell et al., 2021). See Table 1 for descriptions of participant demographics in each experiment.

Table 1. Participant Demographics

Demographic Variable	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6
N	240	327	264	322	322	338
Age Mean (*SD*)	40.07 *(13.36)*	39.18 *(13.16)*	38.50 *(11.11)*	40.9 *(11.7)*	41.4 *(13.2)*	42.4 *(11.5)*
Gender (%)						
Male	59.5	50.3	53.4	41.9	52.4	64.4
Female	39.3	48.8	45.5	43.2	46.3	35.6
Other	0.1	0.3	0.4	0.3	0	0
Education (%)						
High School Diploma	22.7	20.9	25.9	20.5	11.0	11.0
College Diploma	11.6	10.0	8.6	7.14	4.11	4.1
Bachelor's Degree	45.0	45.5	44.0	41.6	61.3	61.3
Master's	16.9	19.1	16.2	13.7	22.9	22.9
Professional Degree	1.7	3.0	2.6	1.24	0.7	0.7
PhD	1.2	0.9	1.9	2.17	0	0
Ethnicity (%)						
American Indian	0.0	0.6	0.0	-	-	0.7
Asian	14.5	10.9	6.0	-	-	6.2
Black or African Amer.	10.7	12.1	14.7	-	-	5.5
Hispanic or Latino/a	4.5	4.8	3.0	-	-	1.4
Native Hawaiian/Pacific Islander	0.0	0.3	0.0	-	-	0.0
White	68.2	70.3	69.5	-	-	85.6
Other	1.2	0.3	0.4	-	-	0.7

Note: The response options for each demographic variable except for Age are contained in the table. Age was an open-ended response. The ethnicity columns do not sum to 100. Any deviation reflects a combination of non-responses and a small amount of rounding error. Ethnicity was not collected for Experiments 4 and 5.

1.1.1.2 Procedure

I modified a standard illusion of understanding task to test the specificity principle. I will first explain the standard task and then highlight the modifications.

A standard task consists of three phases. In Understanding Rating 1, participants rate their understanding of how various mechanical devices[1] (e.g., zipper, can opener, speedometer) work on a scale ranging from 1 (very little understanding) to 7 (very thorough understanding). Prior to rating the items, participants read a description of how to use the scale.[2] Following this scale orientation, participants rate their understanding of each device provided. After providing their ratings participants enter the Explanation phase. They are prompted to explain, in as much detail as possible, how one of the devices works. Sometimes, they are prompted to explain more than one device. After explaining, participants proceed to Understanding Rating 2, where they re-rate their understanding of the item(s) that they just explained. For example, a participant may have initially rated their understanding of how a zipper, a can opener, and a speedometer work, but was only prompted to explain how the can opener works. If that were the case, then in Understanding Rating 2 they would only re-rate their understanding of how a can opener works. After participants re-rate their understanding of relevant item(s), the task concludes. The expanded task followed the same three-phase process with minor modifications to Understanding Rating 1 and Understanding Rating 2. In Understanding Rating 1, I omitted the scale orientation information and instead instructed participants to rate their understanding of the six devices with no further instruction (see Table 2 for the instructions used in experiments 1 - 3 and Table 3 for the list of stimuli used in experiments 1 - 3). The Explanation phase remained unchanged: participants were asked to explain how one of the devices they just rated works. In Understanding Rating 2, participants re-rated their understanding of each device they rated in Understanding Rating 1 instead of just the device they explained. Participants rated their understanding of each device on the same page of the experiment. The devices were displayed in a random order. These were the only differences between the expanded task and the standard task.

---

[1] Although mechanical devices are most common, other domains such as natural phenomena and political systems are also used (Rozenblit & Keil, 2002, Experiment 10).

[2] See Rozenblit and Keil (2002), page 7, for an example of a scale orientation using a crossbow.

Table 2: Primary Task Instructions

Task	Instructions
Understanding Rating 1	"We are going to present you with 6 items. For each item, we want you to rate on a 7-point scale how well you feel you understand how each one works, from 1: Very little understanding to 7: Very thorough understanding." *(Additional instructions of Experiments 2, 3 & 6)* *"When we ask you to rate your understanding of how something works, we are asking how well you would be able to describe all the details you know about how the item works, going from the first step to the last, and providing a connection between each step. That is, in rating your understanding of how the item works, you should consider your ability to produce an explanation that states precisely how each step causes the next step in one continuous chain from start to finish."*
Understanding Rating 2	"Please rate, once again, how well you feel you understand how each item works." *(Additional instructions of Experiments 2, 3 & 6)* *"When we ask you to rate your understanding of how something works, we are asking how well you would be able to describe all the details you know about how the item works, going from the first step to the last, and providing a connection between each step. That is, in rating your understanding of how the item works, you should consider your ability to produce an explanation that states precisely how each step causes the next step in one continuous chain from start to finish."*
Rating scale example:	"How a quartz watch keeps time" 1. (*Very little understanding*) to 7. (*Very thorough understanding*)
Explanation Condition	"Now, we'd like to probe your knowledge in a little more detail about how a quartz watch keeps time. Please describe all the details you know about how this phenomenon works, going from the first step to the last, and providing a connection between each step. That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish." "Explain how a quartz watch keeps time, from the point when the second hand moves, to the point when it moves again."
Control Condition (*Experiments 2 - 6*)	"Please copy the text below by typing it into the box." Above the door to the bedroom is a small wooden rack with red beads dangling from it. Just to the right of the rack is a window framed by white curtains. Next to the window is a plain wooden desk. On the desk are a lamp and various pens and pencils. An ordinary black chair accompanies the desk. Closely behind the chair is a bed not fully made. Near the bed is a closet containing both formal and informal wear. The walls are a light beige color, and the floors are wooden.

Note. The italicized Understanding Rating instructions were additionally included in Experiments 2, 3 and 6. The control condition required participants to replicate the text given in an image. This text's length approximates the average length of written explanations in my prior research. Experiment 1 did not contain a control condition.

Following the expanded illusion of understanding task, participants rated the general similarity between all possible combinations of item-pairs, on a scale from 1 (not at all similar), to 5 (extremely similar). Finally, participants provided demographic information, and the experiment concluded.

Table 3: Stimuli used in Experiments 1 – 3 and 6

Experiment	Domain	Phenomenon
Experiment 1 & 2	Mechanical Device	How a zipper works
		How a speedometer works
		How a piano key produces sound
		How a helicopter flies
		How a sewing machine works
		How a quartz watch keeps time
Experiment 3	Mechanical Device	How a zipper works
		How a speedometer works
	Natural Phenomena	How snow forms
		How earthquakes occur
	Political/Economic System	How legal immigration affects the U.S. job market
		How trade with China affects the average U.S. citizen
Experiment 6	Mechanical Device	How a zipper works
		How a sewing machine works
	Procedure	How to drive a manual transmission
		How to make scrambled eggs

Note. I obtained all "Mechanical Device", "Natural Phenomena" and "Procedure" items from Rozenblit and Keil, (2002, Experiments 1-6, Experiment 10, & Experiment 8, respectively). I obtained "Political/Economic System" from Meyers and colleagues (2020).

1.1.2 Results

I deviated from the preregistration protocol to present a simpler analysis.[3] After attempting to explain how a device works, participants reported knowing less about the device, $t(239) = -4.62$, $p < .001$. However, after attempting to explain how a device works, participants also reported knowing less about different devices, $t(239) = -10.53$, $p < .001$. In other words, participants reported greater understanding of a zipper (and other devices) at Time 1 (pre-explanation) compared to Time 2 (post-explanation) despite only explaining one of them. Table 4 presents the means and standard deviations of each understanding ratings across conditions, including the effect size of the within-subjects comparison of the two ratings. The larger effect size for the Specific condition suggests that the reduction in understanding of, for example, a zipper (device) was somewhat greater when something other than a zipper was explained. I caution against such an interpretation. This is because understanding ratings for some items appear to decrease more when the same object is explained, and others decrease more when something else is explained. This can be observed in Figure 1 below, which displays the average difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained.

[3] I preregistered computing mixed-effects models for my analysis. However, following a suggestion from Jonathan Baron (the editor handling the Meyers et al., 2023 manuscript) I instead report a simplified analysis that is more descriptive than inferential in nature. The results of my modelling are consistent with the descriptive results reported here. Interested readers can consult my mixed effect-modelling in Appendices A-C. This applies to the analyses of each experiment in this thesis.

Table 4. Summary statistics of understanding ratings by condition for each Experiment

Experiment	Condition	*n*	Rating 1 Mean *(SD)*	Rating 2 Mean *(SD)*	Cohen's *d* of difference
Experiment One					
	Explanation – Specific	240	3.99 *(1.93)*	3.60 *(1.99)*	.20
	Explanation – General	240	3.99 *(1.39)*	3.45 *(1.51)*	.37
Experiment Two					
	Explanation – Specific	156	3.47 *(1.83)*	3.11 *(1.80)*	.19
	Explanation – General	156	3.47 *(1.28)*	3.03 *(1.38)*	.44
	Control	172	3.56 *(1.35)*	3.37 *(1.43)*	.13
Experiment Three					
	Explanation – Specific	119	4.07 *(2.00)*	3.86 *(1.95)*	.11
	Explanation – General	119	4.01 *(1.46)*	3.59 *(1.54)*	.22
	Control	145	4.10 *(1.28)*	3.95 *(1.35)*	.08
Experiment Four					
	Explanation – Specific	159	4.63 *(1.37)*	4.25 *(1.38)*	.27
	Control	163	4.68 *(1.31)*	4.83 *(1.69)*	-.09
Experiment Five					
	Explanation – Specific	163	4.33 *(1.37)*	4.54 *(1.27)*	-.16
	Control	159	4.22 *(1.43)*	4.34 *(1.40)*	-.08
Experiment Six					
	Explanation – Specific	126	5.21 (*1.52)*	5.18 (*1.42)*	.02
	Explanation – General	126	5.01 (*1.10)*	4.92 (*1.11)*	.08
	Control	159	5.03 *(1.03)*	5.05 *(1.04)*	-.02

Note. Within each experiment, across conditions, participants rate the same six phenomena. Participants were randomly assigned to either the Explanation or the Control condition. The further levels of the Explanation condition (Specific, General) were within-subjects and hence have the same *n* value. Because only one item was explained, the Explanation – Specific cell has only one phenomenon to be rated that fits the specific criterion (the one explained), whereas the Explanation – General cell has five phenomena to be rated that fit the general criterion (the ones not explained) and the Control cell has six phenomena to be rated (nothing explained). This means there are five times as many observations in Explanation – General and six times as many observations in Control compared to Explanation – Specific. This has implications for the variance of each condition, there will be less variance in conditions with more observations. Rating 1 represents Understanding Rating 1. Rating 2 represents Understanding Rating 2. Each rating was made on a 1 (little understanding) to 7 (very thorough understanding).

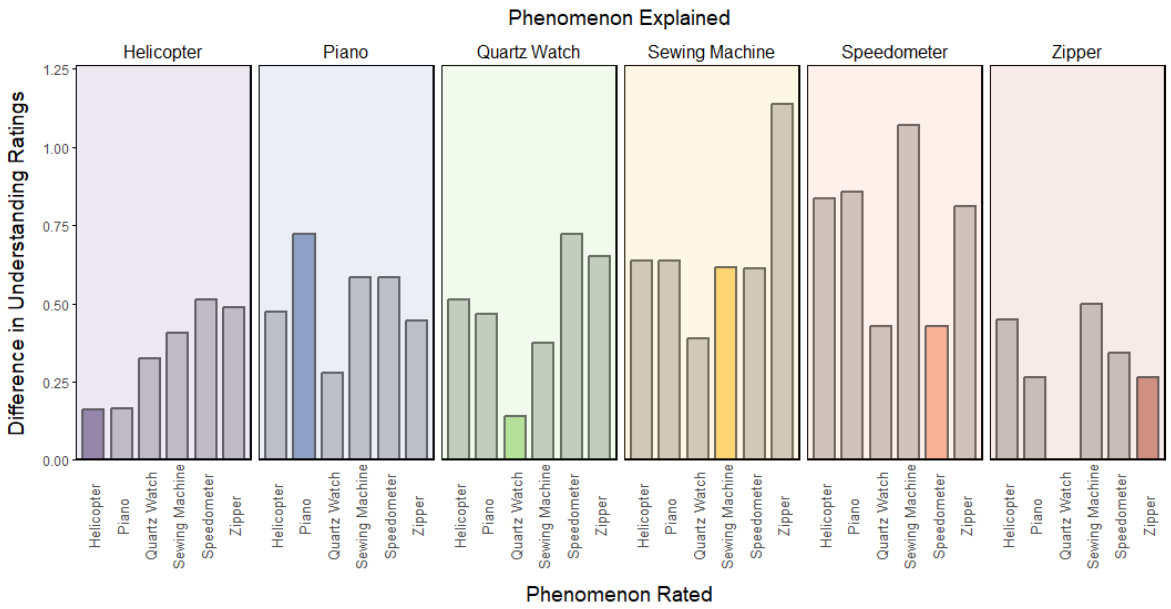


Figure 1. The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1 minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. The panels (and respective participants) are separated by the randomly assigned Phenomenon Explained. Each gently colored panel represents one of the phenomena which is correspondingly labelled at the top of the panel. Each participant was asked to explain how their assigned phenomenon worked. As each participant rated all six phenomena regardless of what they explained, each Phenomenon Rated has a difference-score in each of the plots. Notably, whenever the Phenomenon Rated matched the Phenomenon Explained, the bar is appropriately colored (and all non-matching phenomena are grey colored). The difference-score for the phenomenon Quartz Watch when Zipper was explained was 0.00 and so is not visible in the plot.

Table 5 displays the mean and standard deviations of the similarity ratings of each phenomenon. I found no evidence of an effect of similarity on the reduction in understanding across explanation types (see Appendix A, Table A.1). This might be due to the devices being rated as generally similar. That said, an ANOVA revealed that participants judged at least some of the device pairs (e.g., Zipper and Helicopter) as less similar than other device pairs (e.g., Speedometer and Quartz Watch), $F(14, 3346) = 39.46$, $p < .001$.

Table 5. Mean (*SD*) of Similarity ratings between all device pairs.

Device	Helicopter	Piano	Quartz Watch	Sewing Machine	Speedometer	Zipper
Helicopter	—	3.31 (*0.84*)	3.58 (*0.97*)	3.62 (*1.00*)	3.83 (*1.01*)	3.31 (*0.82*)
Piano		—	3.63 (*1.00*)	3.64 (*1.05*)	3.44 (*0.93*)	3.44 (*0.91*)
Quartz Watch			—	3.58 (*0.96*)	4.05 (*1.06*)	3.43 (*0.96*)
Sewing Machine				—	3.55 (*0.91*)	4.13 (*1.08*)
Speedometer					—	3.37 (*0.88*)
Zipper						—

Note. Each device pair was judged on a 1 (not at all similar) to 5 (extremely similar) scale.

1.1.3 Discussion

Participants reported knowing less about how a device works after attempting to explain it, and this effect was also observed when they attempted to explain a different device. Importantly, the reduction in reported knowledge was similar regardless of whether the participants explained the same or a different device. For example, participants reported a reduced understanding of speedometers after explaining how they work, as well as after explaining how a piano produces noise, how a sewing machine works, or how a quartz watch keeps track of time. I found no evidence that the similarity between devices predicted the reduction of reported knowledge.

While some items were considered more similar than others, I did not find the degree of similarity between device pairs predicted differences in understanding judgments. This finding is unexpected if explaining a device's operation leads to recognizing gaps in one's knowledge of that device. For instance, a zipper and a sewing machine share some properties that are likely to be revealed when explaining their operation, which is less true for a zipper and a piano key. However, it is worth noting that the similarity judgments measured the general similarity between two devices, rather than the similarity of how they work. Asking the latter question might have yielded different results. Given that item similarity was not my primary focus (and to minimize costs), I did not include similarity judgments in subsequent experiments.

One limitation of Experiment 1 is that it did not include a control condition, such as a condition where participants make two judgments without an intervening explanation task. This leaves room for alternative explanations for the observed effects. One such explanation is that providing a second understanding judgment could inherently lead to a reduction in reported understanding. When making a second judgment, a participant may consider more about the device than they did when providing their first judgment. As they have more time to realize what they do not know about the device, the drop in understanding may simply reflect the participant having second thoughts. This parallels the two-response paradigm often employed in reasoning experiments (Thompson et al., 2011), wherein participants first provide a speeded 'intuitive' response and then are given unlimited time to provide a second 'reflective' response. Often, the additional time to reflect leads to a participant changing their answer (Newman et al., 2017; see also Shrout et al., 2018). Thus, the reported reduction of understanding across judgments may have little to do with the explanation process and instead be the product of reasonable adjustments made after having additional time to reflect. This concern also applies to the work of Roeder and Nelson (2015). On the other hand, this

concern is inconsistent with the illusion failing to apply to all domains of knowledge (Mills & Keil, 2004; Rozenblit & Keil, 2002) as well as the usually null effects produced by various control conditions previously implemented (Crawford & Ruscio, 2021; Kleinberg & Marsh, 2020; Fernbach et al., 2013; Zeveney & Marsh, 2016). Because my design is meaningfully different from the designs that previously implemented a control condition, I cannot meaningfully rely on these previous findings. Therefore, I will attempt to rule out this explanation as well.

Another possible explanation is that participants' interpretation of the understanding rating scale changes after the explanation task (Ostrom & Upshaw, 1968; Petty & Wegener, 1993; Schwarz & Bless, 1992). During the pre-explanation judgment, participants are unlikely to consider understanding as including their ability to explain the precise causal chain of how the device works. However, a causal chain is what they are asked to provide when probed for an explanation. During the post-explanation judgment, participants may hold an updated definition of understanding, such that the results could reflect a change in participants' interpretation of the scale between the two ratings. This is especially plausible as I largely replaced the scale-rating instructions that are normally included, which could have otherwise prevented this potential issue. Therefore, in a preregistered Experiment 2, I will attempt to replicate the results of Experiment 1 while ruling out the two aforementioned alternative explanations.

1.2 Experiment

1.2.1 Methods

Experiment 2 followed the same procedure as Experiment 1 with an added between-subjects control condition and a clarified set of rating-scale instructions. This experiment was preregistered on the OSF (https://osf.io/h3trk).

1.2.1.1 Participants

Following the preregistered plan to obtain approximately 95% power to detect a within-subjects effect of $d = 0.2$, I recruited 327 participants via Mechanical Turk (see Table 1 for demographics). This experiment used the same recruitment criteria as Experiment 1 (minimum age of 18, US resident, HIT approval rating of 99% or greater and passing two “botchas”; Littrell et al., 2021).

1.2.1.1.1 Procedure

I made three changes to the procedure of Experiment 1. First, I added a between-subjects control condition (see Table 2). In this condition, instead of explaining how a device works, participants reproduced a paragraph of text manually into a text box (the paragraph of text was contained in an image and could not be copied and pasted).[4] The text passage was descriptive in nature, detailing a room in a home, and reproducing it did not require any effort from the participant beyond reading and typing the exact words (i.e., no explanation of any sort was generated). Second, I added more detail to the understanding rating scale instructions to reduce the likelihood that the scale would be interpreted differently during the post-explanation judgment. These instructions included a sentence stating that when rating understanding, the participant should consider their ability to produce an explanation of the casual chain of how the device works. These changes can be reviewed in Table 2. Third and finally, I removed the similarity ratings. This produced a 2 (Writing Condition: Control/Experimental; between-subjects) x 2 (Time: Pre-explanation/Post-explanation; within-subjects) mixed design.

1.2.2 Results

Consistent with Experiment 1, I deviated from the preregistration protocol to present a simpler analysis. After attempting to explain how a device works, participants reported knowing less about the device, $t(155) = -3.92$, $p < .001$. However, after attempting to explain how a device works, participants also reported knowing less about different devices, $t(155) = 7.08$, $p < .001$. In other words, participants reported more knowledge of a zipper (and other devices) at Time 1 (pre-explanation) compared to Time 2 (post-explanation) despite only explaining one device. This can be observed in Figure 2 below. On average, explaining the specific device versus explaining any device produced a similar reduction in understanding (see Table 4) but this varied substantially by item. Surprisingly, participants who did not explain anything also reported knowing less about the device, $t(171) = -5.63$, $p < .001$. This suggests that some portion of the reduction in understanding is not attributable to having explained something. Importantly, the reduction in reported knowledge in the control condition was half the size of the reduction in the explanation conditions.

[4] I used the control condition from Meyers and colleagues (2020).

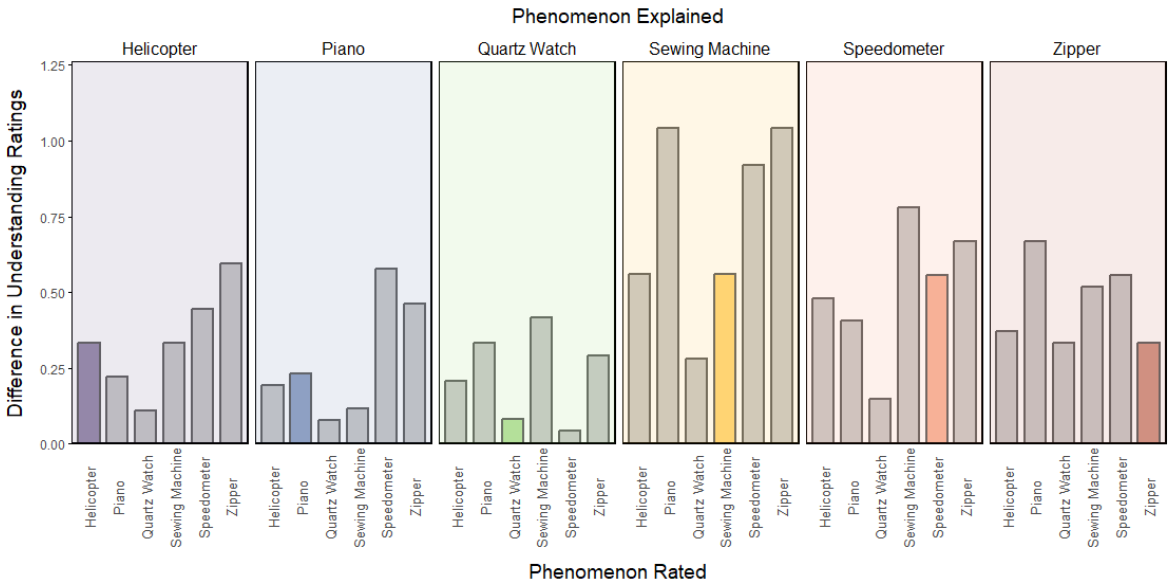


Figure 2. The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1 minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. The panels (and respective participants) are separated by the randomly assigned Phenomenon Explained. Each gently colored panel represents one of the phenomena which is correspondingly labelled at the top of the panel. Each participant was asked to explain how their assigned phenomenon worked. As each participant rated all six phenomena regardless of what they explained, each Phenomenon Rated has a difference-score in each of the plots. Notably, whenever the Phenomenon Rated matched the Phenomenon Explained, the bar is appropriately colored (and all non-matching phenomena are grey colored).

1.2.3 Discussion

Consistent with Experiment 1, participants reported knowing less about how a device works after attempting to explain it. Interestingly, they also reported knowing less about how a device works after attempting to explain how a different device works. For instance, participants reported knowing less about speedometers after attempting to explain how they work, but they also reported knowing less about speedometers after attempting to explain how a piano produces noise, how a sewing machine works, or how a helicopter flies. Additionally, I found that participants who copied a block of irrelevant text instead of explaining a device also subsequently reported lower understanding. However, the reduction in reported understanding was greater when something was explained compared to when nothing was explained.

Experiment 2 calls into question two alternative explanations for these results. Specifically, I tested whether the reduction in understanding was due to people reinterpreting the understanding scale after attempting to explain something. Despite adding further detail to the judgment scale instructions to minimize this possibility, participants still reported lower understanding of the devices after attempting explanation. I also tested whether the reduction in understanding judgments across all items was simply the product of having to provide a second judgment. While I found that providing two judgments led to a decrease in reported understanding, it did not fully explain the effect.

The results thus far suggest that what is being explained might be irrelevant for exposing an illusion of understanding. However, these results were obtained using stimuli from only one domain of knowledge: mechanical devices. I reasoned that if attempting to explain completely unrelated phenomena could expose an illusion of understanding, then even attempting to explain something like how snow forms should lead a person to revise how much they think they know about, for example, how a zipper works. It is possible, on the other hand, that the illusion of understanding applies to certain domains of knowledge but not to others, which means that the illusion is exposed at the domain level. This would imply that an illusion of understanding for a mechanical device (e.g., how a zipper works) could be exposed by attempting to explain any type of mechanical device (e.g., how a sewing machine works), but not any other phenomenon (e.g., how snow forms). Similarly, an illusion held for a natural phenomenon (e.g., how snow forms) could be exposed by attempting to explain other natural phenomena (e.g., how earthquakes happen), but not other types of phenomena (e.g., how a zipper works). I tested these competing predictions in a preregistered Experiment 3.

1.3 Experiment 3

1.3.1 Methods

This experiment was preregistered on the OSF (https://osf.io/hvqes).

1.3.1.1 Participants

Following my preregistered plan to obtain approximately 90% power to detect a within-subjects effect of $f = 0.1$ (approximately equivalent to $d = 0.2$), I recruited 264 participants via Mechanical Turk (see Table 1 for demographics). This experiment used the same recruitment criteria as Experiments 1 and 2 (minimum age of 18, US resident, HIT approval rating of 99% or greater and passing two "botchas").

1.3.1.2 Procedure

I made a single modification to the procedure of Experiment 2. I replaced four of the six device items with four items from two different domains of knowledge. The domains and items were natural phenomena (items: "How snow forms", "How an earthquake occurs") drawn from Rozenblit and Keil, (2002), and political systems (items: "How trade with China affects the average U.S. citizen", "How legal immigration affects the U.S. job market") drawn from Meyers and colleagues (2020).

1.3.2 Results

Consistent with the prior experiments, I deviated from the preregistration protocol to present a simpler analysis. After attempting to explain how a phenomenon works, participants reported knowing less about that phenomenon, $t(118) = -1.76$, $p = .040$. However, after attempting to explain how a phenomenon works, participants also reported knowing less about the other phenomena, $t(118) = -6.11$, $p < .001$. Participants tended to report knowing even less about a phenomenon when they explained a different phenomenon rather than the same one (see effect size column in Table 4), however, as illustrated in Figure 3, there is substantial item-level variance in this effect. Surprisingly, but consistent with Experiment 2, participants who did not explain anything also subsequently reported knowing less about the phenomenon, $t(144) = -3.62$, $p < .001$, even though those who had to attempt explanation reported knowing even less about the phenomenon.

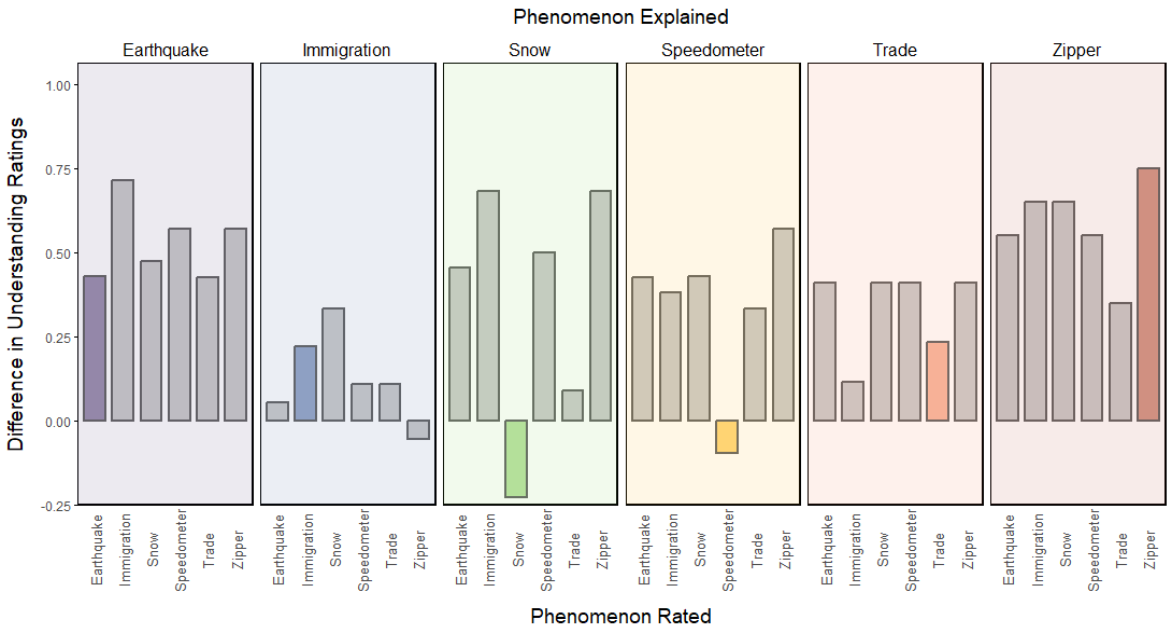


Figure 3. The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. A negative value represents an increase in reported understanding. The panels (and respective participants) are separated by the randomly assigned Phenomenon Explained. Each gently colored panel represents one of the phenomena which is correspondingly labelled at the top of the panel. Each participant was asked to explain how their assigned phenomenon worked. As each participant rated all six phenomena regardless of what they explained, each Phenomenon Rated has a difference-score in each of the plots. Notably, whenever the Phenomenon Rated matched the Phenomenon Explained, the bar is appropriately colored (and all non-matching phenomena are grey colored).

1.3.3 Discussion

I replicated the results of Experiments 1 and 2 using a diverse set of items spanning three domains of knowledge. Participants reported knowing less about how a phenomenon works after attempting to explain it. They also reported knowing less about how a phenomenon works after attempting to explain a different phenomenon. For instance, participants reported knowing less about zippers after attempting to explain how they work. However, participants also reported knowing less about zippers after attempting to explain how a speedometer works, how snow forms, or how earthquakes occur. I also found that participants who copied a block of irrelevant text instead of explaining a phenomenon subsequently reported lower understanding. However, the reduction in reported understanding was greater when the intervening task involved attempting an explanation. These findings have important implications for understanding the functioning of the illusion of understanding and the role of explanation in exposing it.

My findings, which are generally consistent with those of Roeder and Nelson (2015), challenge a Narrow Model that posits the illusion only applies to a single phenomenon and can only be exposed by failing to explain how that particular phenomenon works. While this Narrow Model seems reasonable—we typically do not expect people to report knowing less about how snow forms after explaining how a speedometer works—my data do not support it. Hence, alternative models of the illusion warrant further consideration.

An alternative model suggests that the illusion of understanding applies specifically to individual phenomena but is exposed generally. In other words, while the illusion might apply to different phenomena individually, attempting to explain just one phenomenon exposes a broader set of illusions held. According to this model, a person's lack of knowledge about one phenomenon is generalized to other phenomena. This could result in a temporary increase in intellectual humility (Porter et al., 2022), where people recognize the extent of their ignorance about the explained phenomenon (Meyers et al., 2020), which they then extend to their knowledge of other phenomena. For example, failing to explain how a speedometer works might lead someone to question their understanding of how snow forms because they associate a general feeling of ignorance with the rest of their knowledge. This model does not view the illusion as an expression of general overconfidence, although reductions in confidence might occur alongside the exposure of the illusion. I refer to this model as the Single Broad Effect Model.

A third model, the Two-Effects Model posits that there are two distinct effects when failing to explain something: a target-specific effect and a breadth effect. The target-specific effect involves recognizing the limits of one's knowledge about the specific item being explained. In contrast, the breadth effect induces a general sense of intellectual humility, leading individuals to reassess their understanding of other unrelated phenomena. This model suggests that the target-specific effect is immediate and focused, while the breadth effect is more diffuse, affecting general knowledge calibration. In a nutshell, this model considers the traditional narrow model and the single broad effect model to be accurate and compatible with each other.

Another model suggests that the illusion of understanding represents a manifestation of general overconfidence and not a distinct phenomenon. According to this model, the illusion broadly applies, meaning that no domain of knowledge is immune to it at baseline. Like the Single Broad Effect Model, this model also assumes that failing to explain just one phenomenon should expose the entire illusion and make people report knowing less about most phenomena. This model is iconoclastic in that if it is supported, it suggests there is no unique phenomenon that we currently understand as the illusion of understanding. I refer to this as the General Overconfidence Model.

The Single Broad Effect Model and General Over Confidence Model differ in the material that the illusion applies to and in the material that can expose the illusion. This raises the question of whether the illusion acts at a more general level applying to all or most phenomena, or at a more specific level applying only to certain phenomena. While the illusion is traditionally considered a specific type of overconfidence separate from general overconfidence, as it has been observed to predominantly apply to phenomena with a rich causal structure (e.g., how a speedometer works; Rozenblit & Keil, 2002; Mills & Keil, 2004), it has also been observed in domains lacking such structure (e.g., how to administer a flat tax; Alter et al., 2010; Fernbach et al., 2013). Therefore, it is unclear whether the illusion is an expression of general overconfidence. If there is not a privileged group of phenomena immune to the illusion, then my further efforts would focus on establishing the validity of the illusion of understanding. If there is a privileged group of phenomena immune the illusion, then I could reject the General Overconfidence Model and use those stimuli to assess the limits of the breadth effect, which could be used to tease apart whether the effect of explanation of the illusion of understanding is indeed one broad effect or two (or more) distinct effects.

Before progressing to the next experiment, it's worth considering that these results may be artifacts of the experimental design. For example, in Experiments 2 and 3 I found small (i.e., $d = .10$) but significant reductions of reported knowledge in the control conditions. This suggests that some portion of the observed reduction in understanding is not due to explaining anything. This could be because people tend to take advantage of the opportunity to change their answer after having additional time for reflection (Pennycook et al., 2014; Thompson et al., 2011). It may also be because people's understanding of the response scale (their understanding of the word "understanding") shifts across ratings (see Ostrom & Upshaw, 1968; Petty & Wegener, 1993). However, neither mechanism can fully account for why the process of explanation reduces self-reported understanding. One possibility is that due to conversational norms, a person may assume there is a specific reason why they are being asked the exact same question again (e.g., "I exaggerated my first rating of understanding"; Schwarz & Bless, 1992). Alternatively, the phenomenon might relate to the finding that the average of two guesses tends to be better than one guess alone (Vul & Pashler, 2008). If people are initially overestimating their knowledge, then it is likely that a second guess will be lower than the first. This does not suggest that the second guess is necessarily more accurate, but rather that the person's true knowledge of the phenomenon might sit at the average of their two guesses.

Another concern is that the results may be attributed to demand effects. While the no-explanation comparison condition eliminates several potential demand explanations, it is plausible that the demands on the participants were not uniform across conditions. In other words, participants who provided an explanation for a phenomenon may have felt compelled to report a similarly adjusted understanding for the phenomena they did not explain. This could be because after explaining one of the six phenomena the participants were asked to re-evaluate their understanding of all six phenomena, which may have made it particularly salient that they should make the same adjustments for each phenomenon. However, if this were the case, then for each phenomenon explained I should observe a similar reduction (or increase or no change) in understanding when that phenomenon is re-rated compared to when other phenomena are re-rated. The phenomena Speedometer and Snow in Experiment 3 suggest this is not the case. Figure 4 demonstrates that participants reported an increased understanding of speedometers after explaining how they work, and yet reported a decreased understanding of every other phenomenon. This same pattern emerged for the participants who explained how snow forms. These cases suggest that even if explaining a phenomenon does not make a person feel they understand it less, they may still feel they understand

other phenomena less. Furthermore, it is not clear what demand effect would account for the observed variation in the magnitude and direction of the change in understanding of a phenomenon due to explaining it. Therefore, I do not believe that the evidence for demand effects explaining the results is strong. Still, I will assess the validity of demand effects throughout the thesis.

Chapter 2
On the boundaries the illusion of understanding

2.1 Experiment 4

I have been investigating one component of breadth in the illusion of understanding, specifically the breadth of reducing the illusion. The previous experiments did not find a consistently larger reduction in perceived understanding for the item that was explained compared to the items that were not explained, although I still found a significant (and occasionally larger) reduction for the item that was explained. This finding casts doubt on a fundamental principle of the illusion of understanding—that it is distinct from general overconfidence (Rozenblit & Keil, 2002; Fernbach et al., 2013). However, evidence consistent with an interpretation does not make the interpretation definitive (Popper, 1974), especially when the evidence is a lack of result. Proponents of the idea that the illusion of understanding is distinct from overconfidence can point to seemingly robust evidence: demonstrated cases where the illusion fails to extend to certain types of knowledge.

People have exhibited an illusion of understanding in a wide variety of domains. For instance, the experiments reported thus far used mechanical devices, natural phenomena, and social phenomena to test the illusion. However, there is evidence that the illusion does not apply to some domains of knowledge. For example, the illusion has failed to extend to narratives (e.g., the plots of popular movies) and procedures (e.g., basic tasks such as making scrambled eggs or changing a tire; Rozenblit & Keil, 2002). After rating their knowledge of popular movies (e.g., Jaws, Titanic) and explaining the plot of one of the movies, participants did not report knowing less about the movie they explained (Experiment 9 of Rozenblit & Keil, 2002). Similarly, other participants did not report knowing less about common procedures after explaining how to perform them (Experiment 8 of Rozenblit & Keil, 2002). Understanding why the illusion applies to certain domains but not others is a worthwhile focus of study, but it is not the primary focus of this thesis. My primary interest lies in the fact that these studies constitute the bulk of the evidence for the illusion's failure to apply to other domains.

The aforementioned work was conducted with small samples from a restricted population pool (i.e., about 40 Yale graduate and undergraduate students). While this was an acceptable standard at the time, standards for statistical power and sampling have risen in response to the replication crisis

(Camerer et al., 2018; Open Science Collaboration, 2015; Simmons et al., 2018). Indeed, prior claims such as the idea that exposing the illusion decreases political extremism (Fernbach et al., 2013) have been challenged and supported by recent direct and indirect replications (see Kruger et al., 2022 for challenge; see Meyers et al., 2020 for support). This underscores the need to replicate the experiments that evaluated whether the illusion extends to narratives and procedures.

The next two experiments are modestly updated near-direct replications of Experiments 8 and 9 of Rozenblit and Keil (2002). I have adhered to the original designs as closely as possible while greatly improving statistical power to more accurately test whether the illusion extends to these domains. The original work did not find the illusion in either domain.

2.1.1 Methods

The materials, data, and analysis script for this experiment are available on the Open Science Framework (OSF) website (https://osf.io/apvbh). I preregistered this experiment, and the preregistration can be found on the OSF (https://osf.io/umf5p).

2.1.1.1 Participants

Following the preregistered plan to obtain a sample size at least 2.5 times the original, I recruited 324 participants via Mechanical Turk. This number was determined to ensure at least 130 participants per cell after accounting for an estimated 10% exclusion rate, affording more than 80% power to detect a small effect (e.g., $d = 0.2$). Participants were residents of the United States or Canada, had a HIT approval rating of at least 95%, were over the age of 18, and passed a prescreening questionnaire prior to beginning the experiment.

2.1.1.2 Procedure

This experiment replicates Experiment 9 of Rozenblit and Keil (2002), which assessed whether the illusion of understanding extends to narratives. I closely followed the original procedure with three notable changes: administering the experiment online rather than in-lab, making minor updates to some stimuli (see Table 6 for the original and updated stimuli), and adding the control condition used in this thesis. These changes are not expected to significantly impact the experiment's conclusions.

The experiment began with movie selection, where participants were shown a list of 10 popular movies, such as "Star Wars: Episode IV - A New Hope" and "The Lion King," and asked which ones they had seen. Participants then rated their understanding of each movie's plot on a scale from 1 (very little understanding) to 7 (very thorough understanding). They rated their understanding based on their ability to describe the storyline's details from start to finish, ensuring each plot point connected in a continuous chain.

Participants were then randomly assigned to either the explanation condition or the control condition. In the explanation condition, participants detailed the plot of one of the movies they had rated. In the control condition, participants copied a block of text from an image. After the explanation phase, participants re-rated their understanding of the movie plot. Those in the explanation condition re-rated the movie they had explained, while those in the control condition re-rated a randomly selected movie plot they had rated earlier. Finally, participants provided demographic information, sex, age, and education level.

Table 6. Comparison of Movie Stimuli Between Original and Current Experiment

Original Experiment	**Current Experiment**
American Beauty (1999)	Star Wars: Episode IV - A New Hope (1977)
The Spy Who Shagged Me (1999)	Avengers: Endgame (2019)
Analyze This (1999)	The Dark Knight (2008)
Armageddon (1998)	The Lion King (1994)
The Wedding Singer (1998)	Avatar (2009)
American Pie (1999)	Titanic (1997)
The Phantom Menace (1999)	Jurassic Park (1993)
Saving Private Ryan (1998)	Ghostbusters (1984)
The Matrix (1999)	Inception (2010)
Titanic (1997)	The Wolf of Wall Street (2013)
Enemy of the State (1998)	
The Truman Show (1998)	
A Civil Action (1998)	
The X Files (1998)	
The Sixth Sense (1999)	
Lethal Weapon 4 (1998)	
You've Got Mail (1998)	

***Note*:** This table compares the stimuli used in Rozenblit and Keil's (2002) Experiment 9 with the updated stimuli used in the current study. The updates were made to ensure the relevance and recognition of the stimuli by modern audiences. Participants in both studies were asked to rate their knowledge of each movie's plot, following a similar procedure to assess the illusion of explanatory depth.

2.1.2 Results

I computed a mixed effects analyses treating Time and Condition as fixed effects and participant as a random effect with Understanding Rating as the dependent variable. Participants self-reported understanding of movie plots did not significantly change across time, $F(1,1363) < 1$, $p = .486$. However, this was qualified by a significant interaction between time and task type, $F(1,1363) = 10.23$, $p = .001$, indicating that the change in understanding over time depended on the type of task performed. Planned contrasts revealed participants in the explanation condition reported a significant decrease in understanding from Time 1 ($M = 4.65$) to Time 2 ($M = 4.34$), $t(166) = 2.63$, $SE = 0.12$, $p = .009$, whereas those in the control condition reported a non-significant increase in understanding from Time 1 ($M = 4.64$) to Time 2 ($M = 4.84$), $t(154) = -1.86$, $SE = 0.11$, $p = .064$. These results demonstrate an illusion of understanding (and its exposure) for movie plots. The change in understanding rating for each item, split by writing Condition, is depicted in Figure 4 below.

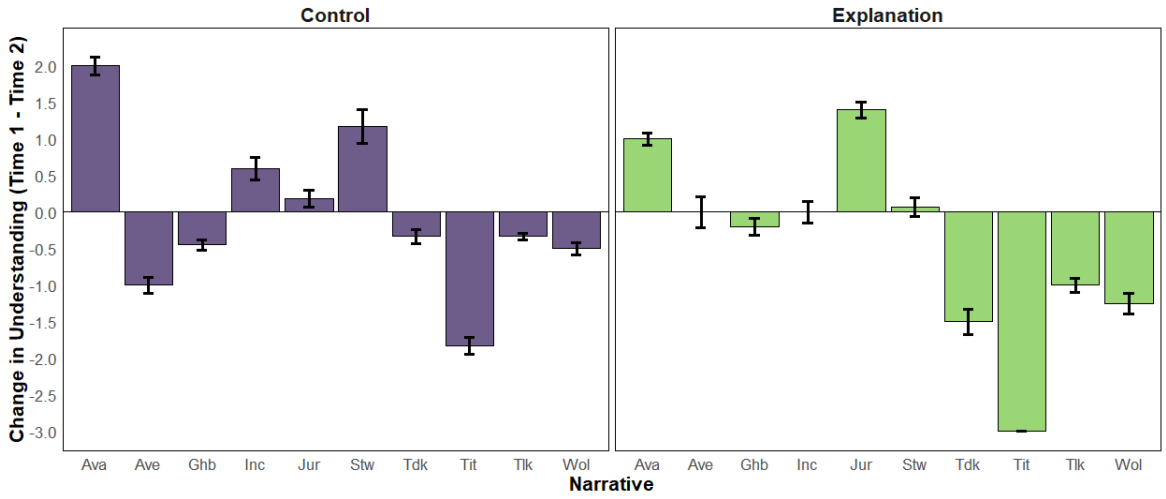


Figure 4. Change in understanding from Time 1 to Time 2 by Narrative, split by Condition. The Y-axis represents the change in understanding (Time 2 rating minus Time 1 rating) for each narrative item. Positive values indicate increased understanding at Time 2 relative to Time 1, while negative values indicate decreased understanding. Error bars represent standard errors of the mean. The solid horizontal line at zero indicates no change in understanding between Time 1 and Time 2. Narratives included in the analysis are "Avatar" (Ava), "Avengers" (Ave), "Ghostbusters" (Ghb), "Inception" (Inc), "Jurassic Park" (Jur), "Star Wars" (Stw), "The Dark Knight" (Tdk), "Titanic" (Tit), "The Lion King" (Tlk), and “Wolverine" (Wol). The control group (left panel) did not provide any explanations, while the experimental group (right panel) wrote explanations for a narrative assigned at random from the movies they reported knowledge of.

2.1.3 Discussion

I attempted to replicate Experiment 9 of Rozenblit and Keil (2002), which originally demonstrated that the illusion of understanding did not apply to narratives. They found that attempting to explain a movie plot after initially rating one's knowledge of it did not lead to a significant reduction in reported knowledge of that movie. Despite a similar design and larger sample size, I did not reproduce this effect. Indeed, while participants in the control condition reported a minor increase in reported knowledge, those in the explanation condition reported a decrease in reported knowledge, consistent with the illusion of understanding. This effect suggests the illusion is broader than previously assumed, and further challenges the fundamental claim that the illusion is distinct from general overconfidence

Could small changes to the design have affected the results? It seems unlikely. Despite updating the stimuli, many of the original stimuli were included as they remain popular and culturally relevant films today, leaving little room for item-effects driving the change in result. Moreover, the experiment differed procedurally only in that it was administered online rather than with pen-and-paper in person. If either of these factors were responsible for the change in results, it would cast doubt on the robustness of the illusion. A more reasonable explanation for the difference in observed results is the increased sample size increasing the likelihood of finding a genuine effect. While it is possible that the participant demographics of the original work (Yale graduate and undergraduate students) were not conducive to finding the effect in this domain, even the original authors suggested their population might be especially susceptible to the illusion of understanding (Rozenblit & Keil, 2002). Furthermore, it is unclear why a more diverse set of adults would necessarily exhibit different results. Instead, the larger sample size permitted a more powerful examination of the phenomenon, leading me to find the effect. Moreover, the inclusion of a control condition, which was absent in the original experiment, allows me to more concretely show the effect that explanation had.

Notably, this experiment exhibited a null effect in the control condition. In the previous two experiments, I observed a small but significant reduction in understanding across time in the control conditions. In this experiment, I have light evidence for the opposite. When considered in combination with exposing an illusion of understanding in the explanation condition, this further supports the argument against a demand interpretation. However, given my plans to run another similarly designed experiment, it is prudent to withhold further discussion of the control condition until I have more concrete results from the next experiment.

To return to the issue at hand, I have failed to reproduce the findings of one of the supposedly robust pillars of the illusion of understanding—that is, the domains of knowledge to which it does and does not apply. If I cannot reproduce a domain that does not exhibit the illusion, I have little support to claim it is distinct from general overconfidence when considered in combination with the breadth results from the earlier experiments. Thus, I now turn to replicating Experiment 8 from Rozenblit and Keil's (2002) work, which failed to observe an effect of explanation in reducing reported knowledge of how common procedures work.

2.2 Experiment 5

This experiment replicates Experiment 8 of Rozenblit and Keil (2002), which assessed whether the illusion of understanding extends to procedures.

2.2.1 Methods

The materials, data, and analysis script for this experiment are available on the OSF (https://osf.io/apvbh/). I preregistered this experiment, and the preregistration can also be found on the OSF (https://osf.io/umf5p).

2.2.1.1 Participants

Following the preregistered plan to obtain a sample size at least 2.5 times the original, I recruited 322 participants via Mechanical Turk. This number was determined to ensure at least 130 participants per cell after accounting for an estimated 10% exclusion rate, affording more than 80% power to detect a small effect (e.g., $d = .20$). Participants were residents of the United States or Canada, had a HIT approval rating of at least 95%, were over the age of 18, and passed a prescreening questionnaire prior to beginning the experiment (see Table 1 for demographics).

2.2.1.1.1 Procedure

This experiment aimed to investigate the illusion of understanding for procedural knowledge. The procedure was as follows:

Participants were first shown a list of 10 common procedures listed in Table 7 below and asked to indicate which ones they were familiar with. For each procedure they were familiar with, participants then rated their understanding on a scale from 1 (very little understanding) to 7 (very thorough understanding). They were instructed to rate their understanding based on their ability to

describe all the details of performing the procedure, from the first step to the last, ensuring each step connected in a continuous chain.

Participants were then randomly assigned to either the explanation condition or the control condition. In the explanation condition, participants detailed one of the procedures they had rated in as much detail as possible. In the control condition, participants copied a block of text from an image.

After the explanation phase, participants re-rated their understanding of the procedure. Those in the explanation condition re-rated their understanding of the procedure they had explained, while those in the control condition re-rated their understanding of a randomly selected procedure they had rated earlier.

Finally, participants provided demographic information, sex, age, and education level.

Table 7. Procedures Stimuli Used in Current Experiment and in Rozenblit & Keil (2002)

Original Experiment	Current Experiment
The correct procedure for how to tie a bow tie	The correct procedure for how to tie a bow tie
The correct procedure for how to set a table	The correct procedure for how to set a table
A correct procedure for how to make pasta	A correct procedure for how to make pasta
The correct procedure for how to file your taxes	The correct procedure for how to file your taxes
A correct procedure for how to make scrambled eggs	A correct procedure for how to make scrambled eggs
A correct procedure for how to make chocolate chip cookies from scratch	A correct procedure for how to make chocolate chip cookies from scratch
The correct procedure for how to make an international telephone call	A correct procedure for how to make a group video call via the internet
A correct procedure for how to drive from New Haven to Chicago	The correct procedure for how to dance a Tango
A correct procedure for how to drive from New Haven to New York City	The correct procedure for how to drive a vehicle with a manual transmission
The correct procedure for how to tie a bow-tie	A correct procedure for how to make a bed

Note: This table compares the procedures stimuli used in Experiment 5 of this thesis (Current Experiment) with those used in Experiment 8 of Rozenblit & Keil (2002) (Original Experiment). The stimuli were found in Appendix B.1 of the original publication which included the duplicate how to tie a bow-tie item (one of which was written as “bow tie” and the other as “bow-tie”, both are included for completeness).

2.2.2 Results

I computed a mixed effects analyses treating Time and Condition as fixed effects and participant as a random effect with Understanding Rating as the dependent variable. Participants' self-reported understanding of procedures did not significantly change across time, $F(1, 1847) = 2.28$, $p = .131$. This lack of significant change was consistent across both the explanation and control conditions, as indicated by the non-significant interaction between time and task type, $F(1, 1847) = 0.22$, $p = .636$. Indeed, planned contrasts revealed that participants in the control condition did not show a significant change in understanding from Time 1 ($M = 4.23$) to Time 2 ($M = 4.33$), $t(162) = -0.73$, $p = .463$. Similarly, participants in the explanation condition did not show a significant change in understanding from Time 1 ($M = 4.32$) to Time 2 ($M = 4.53$), $t(157) = -1.40$, $p = .162$. These findings suggest that the act of explaining how a procedure works does not impact self-reported understanding of that procedure. The change in understanding rating for each item, split by writing Condition, is depicted in Figure 5 below.

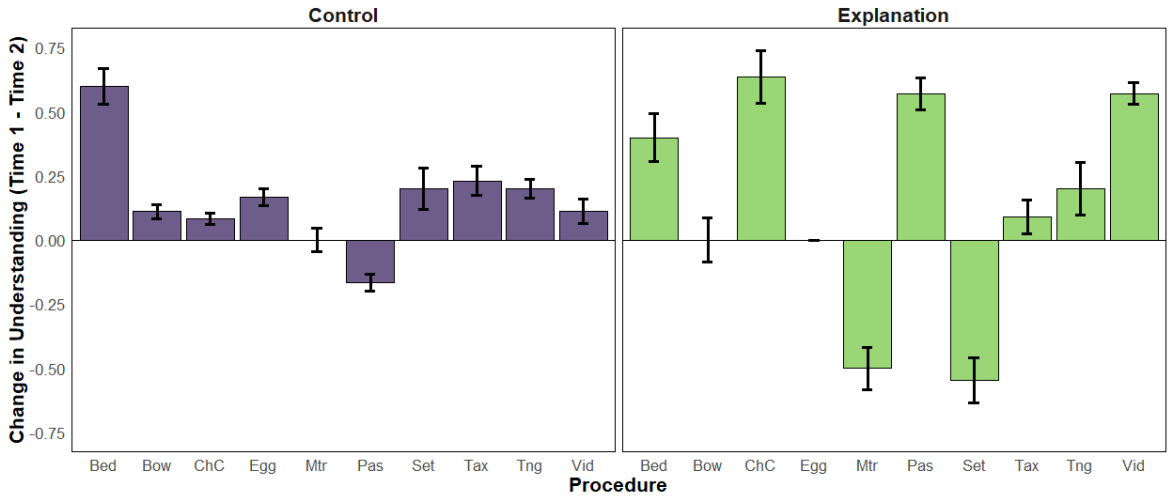


Figure 5. Change in understanding from Time 1 to Time 2 by Procedure, split by Condition. The Y-axis represents the change in understanding (Time 2 rating minus Time 1 rating) for each procedure. Positive values indicate increased understanding at Time 2 relative to Time 1, while negative values indicate decreased understanding. Error bars represent standard errors of the mean. The solid horizontal line at zero indicates no change in understanding between Time 1 and Time 2. The procedures included in the analysis are Bed Making, Bow Tying, Baking Chocolate Chip Cookies, , , Cooking Scrambling Eggs, Driving a Manual Transmission, Cooking Pasta, Setting a Table, Filing Taxes, Dancing the Tango, Making a Video Call. The control group (left panel) did not provide explanations, while the experimental group (right panel) was required to explain the procedure.

2.2.3 Discussion

I attempted to replicate Experiment 8 of Rozenblit and Keil (2002), which originally demonstrated that the illusion of understanding applied to common procedures. They found that attempting to explain a procedure after initially rating one's knowledge of it did not lead to a significant reduction in reported knowledge of that procedure. I reproduced this (lack of) effect. Administering the experiment online, minorly updating the materials, and recruiting a larger, more diverse sample led to the same result: participants did not report knowing less about a procedure after trying to explain it. In fact, participants non-significantly tended to report knowing more about the procedures after explaining them. This supports the claim that at least common procedures are exempt from the illusion of understanding, which bolsters confidence that the illusion of understanding is a construct distinct from general overconfidence.

This finding reveals a limit to the breadth of the illusion of understanding, at least in one sense. Leveraging this knowledge, I can refocus the investigation: does the breadth principle represent a distinct effect from the specificity principle, or are they manifestations of the same underlying mechanism? The Single Broad Effect Model suggests that the breadth and specificity principles are manifestations of the same general mechanism that induces intellectual humility. In contrast, the Two-Effects Model posits that there are two distinct effects: a target-specific effect that reduces confidence in one's knowledge of the explained item, and a breadth effect that induces a more general sense of intellectual humility. By replicating a boundary condition of the illusion, I have provided evidence against the General Overconfidence Model, which posits that these principles are the same and the illusion is not distinct from general overconfidence. With this new information, I can refine the focus to the two competing models. Investigating whether items not previously shown to exhibit an illusion of understanding (e.g., how to cook scrambled eggs) will exhibit a breadth effect after explaining an item previously shown to demonstrate the illusion (e.g., how a zipper works) can help pit these two explanations against each other and further elucidate the mechanisms underlying the illusion of understanding.

Consistent with the previous experiment, no reduction in reported knowledge was observed in the control condition. However, inconsistent with the previous experiment, I also failed to find an effect of the illusion's exposure. These results contrast with Experiments 2 and 3, where a reduction in reported knowledge was observed across time in the control condition. This discrepancy raises questions about potential causes. Two notable differences between the sets of experiments are the

design and the stimuli used (Simmons et al., 2011). Since I did not manipulate these variables independently, it is unclear which might be responsible for the differing results. My results suggest that the change in design may play a larger role. For instance, in the previous two experiments, I observed a null effect in the control condition, but not all nulls in the explanation conditions. This indicates that the exposure effect can be found even when there is no 'exposure' effect in the control condition, pointing more towards design rather than item differences. However, this remains uncertain, and a design that incorporates procedure stimuli in a breadth design, where multiple domains of items are included along with a control condition, would provide further insight into this matter.

2.3 Experiment 6

In previous experiments, I explored the breadth effect of explanation and its impact on participants' perceived understanding of various phenomena. Experiments 1-3 demonstrated that explaining a concept can lead to a reduction in self-reported understanding of items that were not explained, termed the breadth effect of explanation. I hypothesized that this effect occurs in addition to a separate, target-specific effect, where explaining an item leads to reporting less understanding of that item afterward. The significant variance observed between the average reduction in understanding for the items explained versus not explained suggests that these two effects are distinct. Sometimes, the target-specific effect would be significantly larger, while other times, the breadth effect would dominate.

If the two effects are not separate, it challenges the idea that the illusion is distinct from general overconfidence. In recent experiments, I explored the limitations of the illusion and confirmed its distinctness from general overconfidence by reproducing the lack of effect when applied to common procedures, though not for movie plots. This suggests that the breadth effect of explanation might be separable from the target-specific effect.

This experiment aims to assess whether the breadth effect can apply to items previously shown to be immune to the illusion. Specifically, I examine whether participants will report knowing less about procedures after explaining mechanical devices, even if they do not show a target-specific effect for procedures. This question tests whether the breadth effect extends to common procedures, which have not previously demonstrated the illusion of understanding. Furthermore, this experiment tests for limitations to the breadth of the explanation effect. If the breadth effect applies where the

target-specific effect does not (e.g., people report knowing less about mechanical devices after explaining procedures and not reporting knowing less about them), I will have strong evidence of their separability. Conversely, if I do not find this, it suggests that the two effects are more intertwined than the Two-Effects Model assumes.

Additionally, this experiment serves as a test for the possibility of demand effects specific to the breadth design. In Experiments 4 and 5 that followed a more traditional design, I found no effect in the control condition. However, in Experiments 2 and 3 that featured the breadth design, I found a small but significant reduction in understanding in the control condition. This raises the possibility that the breadth procedure induces demand effects or that the new stimuli in Experiments 4 and 5 are less prone to such effects. By using one of the stimuli from Experiment 5 (i.e., procedures) in a breadth design, I aim to clarify whether the observed results are due to demand effects or inherent properties of the stimuli or neither.

2.3.1 Methods

The materials, data, and analysis script for this experiment are available on the OSF (https://osf.io/389w6/). I preregistered this experiment, and the preregistration can be found on the OSF (https://osf.io/e8bsy).

2.3.1.1 Participants

Following the preregistered plan to obtain a sample size sufficient to detect a small effect size with high power, I recruited 300 participants via Mechanical Turk. According to G*Power 3.1.9.7, a sample size of 266 participants is required to detect an interaction effect size of $f = 0.1$ (approximately equivalent to $d = 0.2$) with 90% power, assuming an alpha of 5% and a correlation among repeated measures of 0.5. I oversampled to account for potential exclusions. Participants were residents of the United States or Canada, had a HIT approval rating of at least 95%, were over the age of 18, and passed a prescreening questionnaire prior to beginning the experiment (see Table 1 for demographics).

2.3.1.2 Procedure

This experiment followed a similar procedure to what was used in Experiments 2 and 3, with modifications to the items being rated. Participants first assessed their understanding of four items: two mechanical devices (how a zipper works, how a sewing machine works) and two procedures

(how to cook scrambled eggs, how to drive a vehicle with a manual transmission). The mechanical devices were selected because they showed consistent effects in Experiments 1 - 3. The procedures were selected because most participants indicated they were familiar with them in the previous experiment. Participants rated their understanding on a scale from 1 (very little understanding) to 7 (very thorough understanding). They were instructed to rate their understanding based on their ability to describe all the details of how the item works or how to perform the procedure, from the first step to the last, ensuring each step connected in a continuous chain.

Next, participants were randomly assigned to either the explanation condition or the control condition. In the explanation condition, participants explained one of the four items they had rated in as much detail as possible. In the control condition, participants copied a block of text from an image. After the explanation phase, participants re-rated their understanding of all four items. Those in the explanation condition re-rated their understanding of the item they explained, while those in the control condition re-rated their understanding of a randomly selected item they had rated earlier. Finally, participants provided demographic information, including ethnicity, sex, age, and education level.

2.3.2 Results

I computed a mixed effects analyses treating Time and Condition as fixed effects and participant as a random effect with Understanding Rating as the dependent variable. Participants' self-reported understanding of mechanical devices and common procedures did not significantly reduce across time, $F(1, 2170) < 1$, $p = .782$. Additionally, there was no significant interaction between time and writing condition, $F(1, 2241) = 1.16$, $p = .282$. In other words, the difference between the Time 1 and Time 2 understanding ratings for participants who explained how a mechanical device or how a procedure works were comparable to participants who copied text from a box. This suggests the inclusion of procedures as stimuli inoculated people from exhibiting an illusion of understanding, even for items where people have exhibited the illusion before.

I then performed an item-level analysis to determine whether any item, especially the mechanical devices, showed an illusion exposure effect (significantly lower ratings at Time 2 after explanation). The understanding ratings for each item across time are displayed in Figure 6 below. As illustrated by the figure, no item exhibited a different pattern between the control and explanation conditions. People did not exhibit an illusion of understanding for any item in the experiment.

Alternatively, explanation was not sufficient to expose the illusion when procedures were included as stimuli.

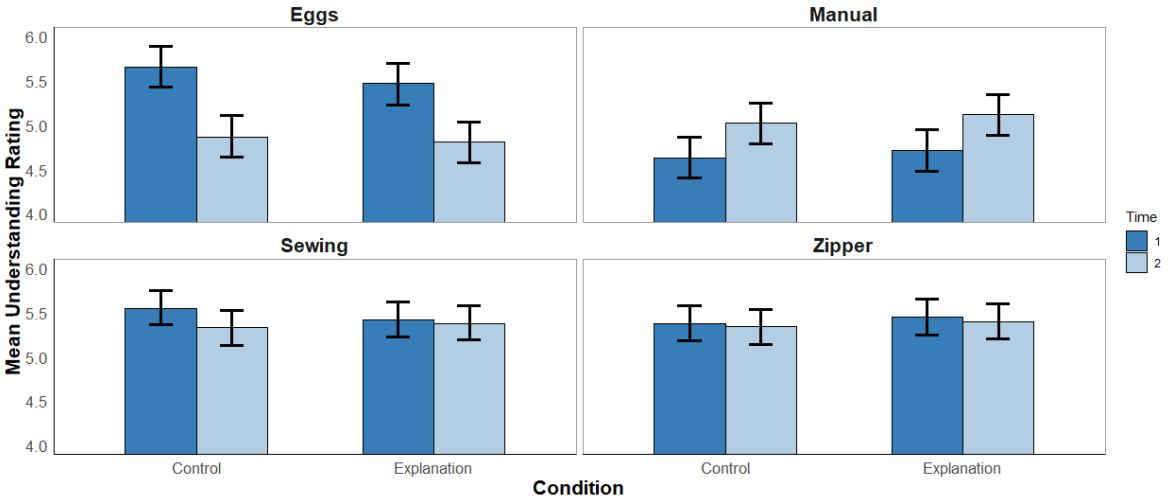


Figure 6. Mean understanding ratings by Condition and Time for the items of Experiment 6 (Eggs, Manual Transmission, Sewing, Zipper). The Y-axis represents participants' mean understanding ratings at Time 1 and Time 2, with error bars representing 95% confidence intervals. The Control and Explanation conditions are displayed on the X-axis, with Time 1 (solid fill) and Time 2 (lighter fill) ratings shown in adjacent bars. For all objects, participants showed minimal changes in understanding from Time 1 to Time 2, across both conditions, with the exception of the procedure “how to cook scrambled eggs”.

2.3.3 Discussion

In this experiment, I incorporated common procedures as stimuli into a breadth design similar to those used in experiments 1-3. The primary objective was to examine whether the breadth effect of explanation extends to items not previously demonstrated to exhibit the illusion. It did not. Participants did not report knowing less about procedures or mechanical devices after explaining one of them. Unexpectedly, there was not a target-specific effect of explanation, even for items that consistently demonstrated the effect in the earlier experiments. The inclusion of common procedures as stimuli not only inoculated people against the breadth effect of explanation but against exhibiting the illusion entirely.

These results are puzzling. They may be due to random error, or they may be genuine, and without replication, any speculation about the results must be approached with caution. Assuming the results are not random, several possible explanations arise. One idea is that including procedures led to more conservative estimates of understanding of the mechanical devices. Procedures might be easier to assess because they are easier to imagine (Rozenblit & Keil, 2002; Wilson & Marsh, 2023). This could result in them serving as anchors for mechanical device assessments. However, the mean ratings for mechanical devices fell within previously collected ranges, and no clear difference was observed between ratings for procedures and mechanical devices (see Figure 6). Thus, any effect on Time 1 ratings remains unclear. Additionally, the nature of the design limits my ability to assess the validity of other possible explanations, such as whether the inclusion of procedures affected how people reassessed their knowledge of the items, affected how people explain mechanical devices, or affected how people perceive their understanding of mechanical devices when presented alongside procedures. As these are not the primary focus of the thesis, I will devote no more attention to them.

I can draw some inferences from the lack of reduction in reported understanding observed in this experiment. For instance, concerns of demand effects producing the breadth effect are further undermined. There were no meaningful differences in the possibility of demand effects between this experiment and the experiments of similar designs previously reported (e.g., Experiments 2 & 3). If conversational norms were nudging participants toward changing their answer when asked to re-rate their understanding of the item, I should have observed that. Likewise, if participants became wise to the idea that they should admit they know less about the item than previously stated and the same applies to the other items, then I also should have observed that. The strongest argument for a demand

interpretation must be that these forces do exist, but that they are not strong enough to override the inoculating effect of including procedures as stimuli.

The findings of this experiment further confirm a limitation to the applicability of the illusion to common procedures (at least those tested). Moreover, my findings suggest a limit to the breadth effect of explanation such that, insofar that is distinct from the target-specific effect of explanation, the breadth effect does not reliably occur without the target-specific effect also occurring. While this experiment did not provide a direct test of this claim, it provides a clear direction for future work. However, that future work prerequisites establishing that the breadth effect of explanation is indeed separate from the target-specific illusion. The remaining research I present focuses on this question.

Chapter 3
Evidence of two distinct effects of explanation

3.1 Secondary Data Analysis of Reported Experiments

Why do people report knowing less about an item after explaining a different item? This thesis posits that the failure to explain an item not only undermines confidence in one's knowledge of that item but also casts doubt on one's broader knowledge base. Furthermore, it suggests that these represent distinct effects resulting from failing to explain a phenomenon. The experimental evidence presented thus far lends support to this framework and suggests it is stronger than other accounts, though it does not conclusively substantiate it. For instance, the Single Broad Effect Model can account for the variance observed between the effect sizes for the target-specific and breadth explanation effects observed in Experiments 1–3. The model of a single broad effect of explanation that applies to both the explained and unexplained items struggles to explain this variance in effect sizes but could suggest that this is due to item effects resulting from not studying enough items. The model that the illusion is not distinct from general overconfidence is not supported.

Another contending framework driven by the results of the first experiments is that people report knowing less about an item after explaining a different item because they become wise to the research hypothesis or that social norms cue them into responding differently—a demand interpretation. While demand was already an unlikely candidate, though a possible one, the recent experiments failing to find effects in the control conditions across the two designs serve as further grounds to reject the claim that the results are due to demand effects. This leaves the competing frameworks of whether failing to explain a phenomenon has two distinct effects, one target-specific and one broad effect on intellectual humility, or one broad effect. It is imperative to pursue this distinction.

Having already collected data from over one thousand participants across six experiments, I must consider whether additional insights can be gleaned from the existing data. One unexplored variable is the participants' actual explanations. Coding these explanations for knowledge evidenced by the content of them could help identify the nature of the effects. By comparing whether the discrepancy between initial judgment ratings and coded explanation quality predicts the drop in reported understanding for the item explained versus the items not explained, I can test the

distinctness of the effects. If the discrepancy is predictive of both or neither, it suggests a single effect. Conversely, if it is predictive of one but not the other, it suggests distinct effects.

Additionally, coding for explanation quality provides an 'objective' measure of knowledge for the items. This allows me to compare participants' self-assessed understanding with their demonstrated knowledge, offering insight into whether people are truly miscalibrated in their self-assessments. However, the value of this comparison is limited by the criteria used to code the explanations, and I will return to this limitation later.

One reason why the explanations given by participants have been unanalyzed thus far is because coding is challenging. It is resource and labor-intensive. Indeed, coding of explanations is uncommon in the literature, though it has been performed (e.g., Wilson & Marsh, 2024). One reason for its rarity is that it is not necessary to demonstrate the illusion (as the illusion is subjective in nature), despite the immense additional value it could provide. However, with the public availability of large language models, researchers have begun using these tools to code written responses in other areas of social science, including psychology. For example, studies have successfully employed models like ChatGPT to handle coding tasks in social science research (Hoyt et al., 2022; Lu & Stepchenkova, 2015; Yin et al., 2020; Cowls et al., 2021). What if I applied this approach here?

In this chapter I use ChatGPT to evaluate each participant’s explanation. I will have ChatGPT create a rubric explanation for each phenomenon and then compare each explanation against its respective rubric. This approach ensures reliability and efficiency. By aggregating these scores, I will create a coded explanation quality variable that can be used to analyze relationships across all experiments reported in the thesis. This will provide direct evidence to determine whether the target-specific effect and the breadth effect of explanation are indeed distinct phenomena.

3.1.1 Methods

The materials, data, and analysis script for this secondary data analysis are available on the OSF (https://osf.io/389w6/). This analysis was not preregistered.

3.1.1.1.1 Data

I used data from all six experiments reported in this thesis thus far. ChatGPT-4 was employed to code the explanation quality of each participant who provided an explanation.

3.1.1.2 Generating rubric explanations

To create rubric explanations, I provided ChatGPT (model 4.0.0) with the exact same prompts given to participants in the experiments. ChatGPT generated initial explanations, which were then refined to more closely resemble the concise responses typically given in online experiments. This process resulted in a list and short description of 8-10 key points for each item. Sample ChatGPT-generated explanations for some items can be found in the figure below.

Table 8. Sample ChatGPT-4 rubric explanations and evaluation instructions.

Item	Rubric Explanation
Zipper	**1. Slider Engagement:** The process begins when the slider, the central component of the zipper, is positioned at the bottom or one end of the zipper tracks. Each track consists of a series of teeth or coils, attached to separate pieces of fabric. **2. Teeth Interlocking:** As the slider is pulled up along the tracks, it brings together two opposing sides of zipper teeth. Inside the slider, there is a Y-shaped channel that forces these teeth to interlock with each other. **3. Continuous Alignment:** The slider moves along the zipper tracks, continuously aligning and meshing the teeth from each side. This action effectively joins the two separate pieces of fabric into one unified form. The precision of the teeth's design ensures that they mesh smoothly and securely. **4. Locking Mechanism:** At the top of the zipper, there is often a locking mechanism built into the slider that prevents the zipper from sliding back down once it is pulled up. This mechanism engages when the slider's pull tab is flipped down against the zipper, pinning an internal wedge against the teeth to lock them in place. **5. Pull Tab Activation:** The user operates the zipper by using the pull tab attached to the slider. Pulling the tab upward drags the slider, and hence the teeth into interlocking positions, while pushing the tab downward can release the locking mechanism to allow for the zipper to be opened. **6. Fabric Integration:** The teeth or coils are sewn onto fabric tapes at the edges of the fabric pieces. This sewing ensures that the zipper is securely attached to the fabric, allowing for repeated opening and closing without tearing the fabric. **7. End Stop Formation:** At the bottom of the zipper, a metal or plastic stop prevents the slider from coming off the tracks. This stop ensures that the slider and teeth remain aligned and functional throughout the life of the zipper. **8. Operational Efficiency:** As the zipper is used, the teeth or coils continue to align and interlock efficiently, thanks to the precision engineering of the slider and the teeth. This system provides a quick and reliable means of fastening two pieces of fabric together.
Helicopter	**1. Engine Power Up:** The process begins with the helicopter's engine(s) being started. These engines provide the necessary power to turn the rotor blades. Helicopters typically use either piston engines or gas turbines. **2. Rotor Engagement:** The engine's power is transmitted to the main rotor shaft, causing the rotor blades mounted on this shaft to spin. The main rotor is responsible for lifting the helicopter off the ground and controlling its movement in the air. **3. Blade Pitch Control:** Each rotor blade can change its angle of attack (pitch) through controls operated by the pilot. This is facilitated by a swashplate mechanism that allows the angle of the blades to be adjusted collectively (all blades changing pitch in unison) or cyclically (changing pitch as they rotate around the rotor disc). **4. Lift Generation:** As the rotor blades spin, they create lift by pushing air downwards. The angle of attack of each blade is critical; increasing the angle increases the amount of lift generated, allowing the helicopter to rise. This is known as the collective pitch control, which affects the overall pitch of the rotor blades. **5. Directional Control:** To move in a specific direction, the pilot uses the cyclic control, which changes the pitch of each rotor blade individually as it rotates around the rotor hub. This cyclic adjustment tilts the rotor disk in the direction the pilot wishes to move, causing the helicopter to fly in that direction. **6. Torque Management:** The action of spinning the main rotor blades generates an equal and opposite reaction in the body of the helicopter, causing it to rotate in the opposite direction. To counteract this, helicopters are equipped with a tail rotor or a secondary means of anti-torque control like NOTAR (no tail rotor system). The tail rotor provides sideways thrust that counters the torque effect of the main rotor. **7. Takeoff:** When the pilot is ready to take off, they increase the collective pitch to generate more lift than the helicopter's weight, allowing it to ascend. Simultaneously, the pilot must manage the tail rotor to keep the helicopter from spinning uncontrollably.

Evaluation Instructions	Given the rubric explanation below of how a helicopter takes flight, please evaluate the quality of the following explanation based on the presence or absence of key points in the rubric. Use a linearly graded scale ranging from 1 to 7, where 1 represents a very poor explanation and 7 represents an excellent explanation that is as good as or better than the rubric. Output a single number from 1 to 7 representing your rating.
Zipper example explanations	"A zipper has 3 parts. the left track and the right track and the part that connects or disconnects the two. as you put the middle part on both sides it locks the two sides together as you pull the middle part. as you pull in the other direction it unlocks the two sides" *Aggregate ChatGPT rating: 2/7* "Start with the two rows of teeth (called elements) separated from each other. Step 2: Position the two rows of elements so that their edges are aligned with one another. Step 3: Pull the slider, a small metal part of the zipper, down the zipper until it is fully closed. This will cause the two rows of teeth to interlock and become one. Step 4: To open the zipper, pull the slider in the opposite direction. This will loosen the interlocking of the two rows of teeth and cause them to separate. Step 5: Zip and unzip the zipper a few times to make sure its working properly. Step 6: Continue to use the zipper as normal and adjust the slider when necessary." *Aggregate ChatGPT rating*: *3.4/7* "A zipper has two sides with separate sets of "teeth" which, when connecting, lock into each other. There is a bottom male straight edge on one side which slides into the female fitting on the other side of the zipper, There is a slider which sits above where the male/female parts interconnect. As the slider moves up it grasps both sides of the zipper and it's sliding upwards motion unites the two sides of teeth so they grip each other tightly. This closes the zipper. To open the zipper the slide performs the same process going downward and at the bottom the male/female connectors are lifted apart so the garment is wide open. In pants the process is slightly different in that there is no male/female connector at the bottom. The slider simply sits at the bottom of the zipper and is started and stopped there rather than the garment opening fully." *Aggregate ChatGPT rating:* 5.5/7
Helicopter example explanations	"The person presses the pedal to move the middle it then goes up and comes back down making a stitch. Now there are more complex stitches. The adjustments can be made using the bopper." *Aggregate ChatGPT rating: 2/7* "From rest, the blades on top of the helicopter begin to rapidly spin to generate force, which causes the helicopter to lift off the ground and into the air. From here, the pilot uses the pedals and levers in the helicopter to control its speed and force as well as to maneuver its direction, such as to move forward in flight." *Aggregate ChatGPT rating: 3/7* "Pilot starts engine, engages propellors (large horizontal prop and tail-mounted prop), causing them to spin. When large prop is spinning quickly enough, pitch of blades causes air to be propelled downward, causing "lift" to be produced. At a certain point, lift exceeds weight of helicopter and causes it to rise. Pilot is able to control angle at which blade spins, tilts blade so that helicopter rises AND is pushed forward or backwards. Rear propellor's rotation prevents helicopter from spinning due to impetus of propellor. Once above the ground, helicopter begins forward flight." *Aggregate ChatGPT rating: 5.2/7*

Note. The table provides rubric explanations for the items "Zipper" and "Helicopter," generated by ChatGPT and used for evaluating participant explanations. The ChatGPT model rated each explanation on a scale from 1 (very poor) to 7 (excellent) based on the presence and quality of key points outlined in the rubric explanations. Sample explanations with their corresponding ChatGPT ratings are included to illustrate the evaluation process.

3.1.1.3 Evaluating Participant Explanations

Once the rubric explanations were created, I used the ChatGPT API to evaluate each participant's explanation relative to the appropriate rubric explanation five times. The evaluation criteria were based on the number of key points from the rubric present in the participant's explanation. ChatGPT was instructed to assign a score from 1 to 7, where 1 indicated no key points and 7 indicated all or most key points were present. Parameters for the API included specifying the model ("gpt-4"), setting the temperature (0.6), max tokens (5), top_p (1), frequency penalty (0.0), and presence penalty (0.0).

For the purposes of this coding scheme, frequency penalty and presence penalty are largely irrelevant as they penalize response outputs that would be overly repetitive or biased in a conversation (where the output is much longer than is being requested here). These parameters are set to zero and forgotten. The response output is limited by the max tokens parameter. Tokens are the currency of ChatGPT and are units smaller than characters. The limit of 5 tokens per response output that I set ensures a single numeric character is outputted by ChatGPT and no more. The "top_p" set to 1 means the model is not restricted in its possible response options (if this was set to a lower value, the model would be restricted from choosing less common words). Finally, the temperature determines how reliable versus creative ChatGPT is in its response output. A value closer to 1 makes the model more rigid, repetitive, and most importantly, more reliable. However, a lower value allows ChatGPT to more flexibly interpret inputs. I tried to strike a balance by selecting the lowest setting that would still afford a high degree of reliability, which preliminary tinkering led me to settle on 0.6. These parameters are further discussed in the General Discussion.

Large language models, such as ChatGPT, work by predicting the statistical likelihood of words appearing next to each other based on vast amounts of text data. While ChatGPT can generate coherent and contextually appropriate text, it lacks actual understanding of the content. Therefore, the evaluation of explanation quality is based solely on the presence of key points as determined by the model's training data and statistical patterns.

3.1.1.4 Analysis

With the explanation quality variable, I re-analyzed the data from each experiment. I assessed how the difference between judged understanding at Time 1 and the coded explanation quality predicted the change in judged understanding from Time 1 to Time 2. Specifically, I examined

whether the discrepancy between initial judgment ratings and coded explanation quality predicted the drop in reported understanding for the item explained versus the items not explained.

3.1.2 Results

3.1.2.1 Validation of ChatGPT-4 Coding

I began the analyses by assessing the reliability and validity of using ChatGPT to code explanations. First, I computed the Intraclass Correlation Coefficient (ICC) within each experiment and aggregated it across experiments, yielding an average ICC of 0.97. Indeed, no ICC of any experiment reached below .90. This is considered excellent reliability. More notably, this reliability was demonstrated using responses from over 1000 participants, supporting the generalizability of the findings. Although I did not analyze the impact on different demographic segments, the broad diversity of the sample bolsters the robustness of the results.

To assess validity, I compared ChatGPT's explanation quality ratings with participants' self-ratings of their understanding. There was a positive correlation between the two, as shown in Table 9 below, meaning that when participants rated their explanations higher, ChatGPT did too. While this is not a direct comparison with human coders, it still provides evidence that ChatGPT's ratings align with participants' own evaluations, supporting the validity of using it for explanation coding. Additionally, Figure 7 below presents the distribution of coded explanation quality for each item in Experiments 2 compared to the Time 1 and Time 2 ratings (see Appendix D for the same figures produced for Experiments 1, 3 & 6). These distributions show great consistency across studies and reflect a pattern expected if participants were accurately in touch with their understanding of the items. For instance, the modal coded explanation quality is about 2 out of 7, with participants performing comparatively worse on more complex items lacking causal structure (e.g., immigration and trade). This offers face validity for the measure, but it does not contribute to the more desirable construct validity. This limitation will be discussed further.

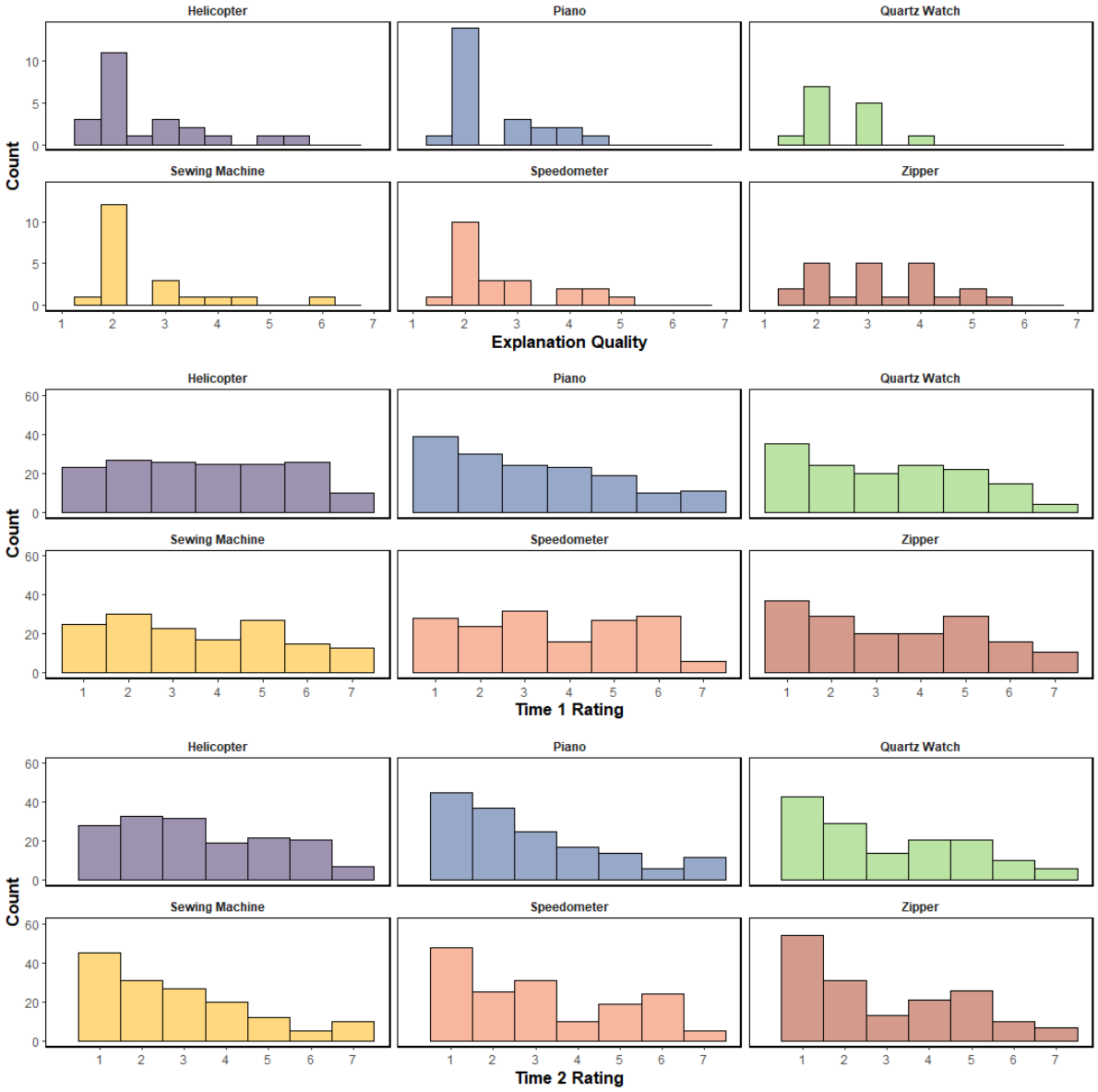


Figure 7. Distribution of explanation quality and understanding ratings over time, split by item from Experiment 2. The first row shows the distribution of ChatGPT coded explanation quality (aggregated across five ratings each scored between 1 to 7). The second and third rows show the distribution of understanding ratings at Time 1 and Time 2, respectively, for the same set of items.

3.1.2.2 Primary Analysis

I examined the relationship between explanation quality and participants' perceived understanding across all the studies presented in this thesis. The correlations between explanation quality and various measures of understanding are summarized in Table 9 below. Notably in the four breadth designs (Experiments 1-3 & 6), the difference between a person's Time 1 judgment of understanding and the coded quality of their explanation predicted the difference between their Time 1 judgment of understanding and their Time 2 judgment of understanding for the item they explained. In other words, the difference between how well someone thought they could explain an item and how well they actually explained the item (as coded by ChatGPT) predicted the reduction in reported knowledge after explanation. This is evidence for the target-specific effect of the illusion of understanding. In contrast, the difference between a person's Time 1 judgment of understanding and the coded quality of their explanation of that item only weakly predicted the difference between their Time 1 judgment of understanding and Time 2 judgment of understanding of the items they did not explain. In other words, the difference between how well someone thought they could explain an item and how well they explained the item weakly and inconsistently predicts the reduction in reported knowledge for the items they did not explain. This is evidence for the separation of the target-specific and breadth effects of explanation.

One objection to the analysis described above is that the target-specific correlations (i.e., between the difference of Time 1 judgment and explanation quality, and the change in understanding from Time 1 to Time 2 of that item) might be statistical artifacts. The key correlation I report is between two differences scores:

(Time 1 understanding rating – explanation quality as rated by ChatGPT)

and

(Time 1 understanding rating – Time 2 understanding rating)

Both difference scores have the Time 1 understanding rating as their first component, and we might consequently expect a positive correlation between the two difference scores as a statistical artifact. To evaluate how much the correlation might be artifactually inflated, I calculated a second correlation keeping the shared Time 1 understanding rating component, but replacing a different component:

(Time 1 understanding rating – explanation quality as rated by ChatGPT)

And

(Time 1 understanding rating – **Participant ID**)

As Participant ID should not correlate with the variables of interest, it serves as a useful foil to test whether the shared Time 1 understanding rating component is artificially driving the *target-specific* correlations discussed above. As Table 9 exhibits, the cross-experiment pattern of the *foil correlations* is markedly different from the cross-experiment pattern of the *target-specific* correlations. Whereas the *target*-specific correlations exhibit a consistent, moderate, positive association, the *foil correlations* are of inconsistent strength, significance, and direction. This provides evidence counter to the claim that the *target-specific* correlations are statistical artifacts. Therefore, it is reasonable to conclude that the results reflect genuine associations rather than noise caused by including Time 1 ratings in both components of the correlation.

Table 9. Correlations with ChatGPT coded explanation quality across each experiment.

Association	Experiment 1	Experiment 2	Experiment 3	Experiment 4 (narratives)	Experiment 5 (procedures)	Experiment 6
Target Specific	.38***	.40***	.36***	.03	.11	.38***
Broad effects	.06	.12	.19*	NA	NA	.21***
Foil for target specific correlations using participant ID	.19**	-.04	-.18*	NA	NA	.14
Explanation Quality and Time 1 Ratings	.32***	.24**	.25**	.07	-.14	.05
Explanation Quality and Time 2 Ratings	.41***	.32***	.26**	.15	-.16	.15
Explanation Quality and (Time 1 Ratings – Time 2 Ratings)	.15*	.09	.01	.05	.06	.11

Note. Correlation coefficients for key associations involving ChatGPT coded explanation quality across each experiment. The *Target Specific, Broad Effects, and Foil* correlations each contain the same X which is the difference between initial understanding rating of the explained item and explanation quality. However, each correlation contains a different Y component. For the *Target Specific* associations, Y is the difference across time in understanding rating of that item. For the *Broad Effects* associations, Y is the average difference across time in understanding ratings of items not explained. For the *Foil* correlations, Y is the difference between time 1 understanding ratings of the item explained and participant IDs. The foil correlations were not analyzed for Experiments 4 and 5. The breadth correlations could not be computed for those experiments (because they were not breadth designs). Significance levels are indicated as follows, * p <.05, ** p < .01, *** p < .001.

Explanation quality did not predict the reduction in reported understanding in any experiment (with the modest exception of Experiment 1). This suggests that explanation by itself is not sufficient to expose the illusion and that an initial rating might be necessary to prime the illusion for exposure. Explanation quality did predict reported understanding at Time 1 and Time 2 in the first three experiments. There is a small, non-significant trend indicating that Time 2 understanding ratings correlate better with explanation quality than Time 1 ratings. Assuming that ChatGPT has accurately captured participants' true knowledge of the item, this trend is consistent with the hypothesis that individuals are adjusting their reported understanding to be more reflective of their true understanding after explaining.

3.1.3 Discussion

This experiment employed a novel approach of using ChatGPT to code explanations given across the six experiments presented in this thesis. The main goal was to help determine whether the breadth effect of explanation is distinct from the target-specific effect. The results suggest a notable distinction between the two effects. Specifically, the correlations between the difference of Time 1 understanding rating and coded explanation quality and the changes in understanding across time were of moderate strength for items that were explained but weaker or non-significant for items that were not explained. This finding suggests that people adjust their understanding judgment of an item after explaining it based on the difference between how well they thought they could explain it and how well they actually explained it. Conversely, people do not adjust their ratings for unexplained items in this fashion. This is evidence for the separability of the effects of explanation.

The use of large language models for coding explanations proved efficient and reliable, offering a scalable solution to a traditionally resource-intensive task. However, while I provided evidence of content validity and face validity, further effort ought to be put toward establishing the use of ChatGPT to code explanation quality (or other variables). Furthermore, the lack of substantial validation of this method should be considered a limitation to this experiment. Caution is warranted in interpreting these results due to the nature of how ChatGPT and large language models function. ChatGPT does not possess actual understanding and does not make genuine judgments about the explanations. While this is unlikely to be an ideal measure of explanation quality, the degree to which it is correlated with self-assessments suggests it represents something meaningful. Previous research

using ChatGPT for coding supports its validity to some extent, but future work should further explore and validate using ChatGPT to code responses.

Consistently, I found that participants rated their understanding higher than ChatGPT rated their explanations. This provides confirmatory evidence that people are indeed overestimating their knowledge and adjusting in the correct direction, though likely not to the optimal degree. This miscalibration in self-assessment aligns with the concept of the illusion, where individuals believe they understand complex phenomena better than they actually do.

Chapter 4
Summary and Conclusions

4.1 General Discussion

After attempting to explain how a speedometer works, participants reported a decreased understanding of a speedometer. They also reported a decreased understanding of a speedometer after attempting to explain how helicopters fly (Experiments 1 & 2), how zippers work (Experiments 1-3), how snow forms (Experiment 3), and several other phenomena. They adjusted their understanding for phenomena other than speedometers in this manner too, including helicopters, zippers, and snow. The breadth of this effect led me to seek a boundary condition, informed by the previously reported boundary conditions of the illusion (e.g., Mills & Keil, 2004; Rozenblit & Keil, 2002). Contrary to the original research, participants exhibited an illusion of understanding of popular movie plots (Experiment 4). However, consistent with the original research, participants did not exhibit an illusion of understanding for common procedures (Experiment 5). Including these procedures as items in a breadth design, I found that the inoculating effects of procedures prevented participants from exhibiting an illusion of understanding for mechanical devices (Experiment 6). Finally, ChatGPT coding of the explanation quality across all reported experiments revealed a correlation between the difference in how well participants said they understood the item and how well they explained it and the difference in understanding ratings for that item. This correlation was not found for items that were not explained. These findings have important implications for understanding the illusion of understanding and the role of explanation in exposing it.

My findings, which align with those of Roeder and Nelson (2015), challenge a Narrow Model that posits the illusion only applies to a single phenomenon and can only be exposed by failing to explain how that particular phenomenon works. While this Narrow Model seems reasonable—we typically do not expect people to report knowing less about how snow forms after explaining how a speedometer works—my data do not support it. Hence, alternative models of the illusion warrant further consideration.

An alternative model suggests that the illusion of understanding applies specifically to individual phenomena but is exposed generally. In other words, while the illusion might apply to different phenomena individually, attempting to explain just one phenomenon exposes a broader set

of illusions held. According to this model, explanation has one broad effect at exposing illusions, meaning the observed target-specific and breadth effects are considered the same effect and could represent a temporary increase in intellectual humility (Porter et al., 2022). For example, failing to explain how a speedometer works might lead someone to question their understanding of how snow forms because they associate a general feeling of ignorance with the rest of their knowledge after failing to explain a speedometer, not because they specifically recognize the limits of their knowledge of speedometers. Similarly, failing to explain the speedometer might lead to someone questioning their understanding of how a speedometer works not because failing to explain how it works led them to realize their lack of knowledge of speedometers, but because failing to explain how it works made them feel less knowledgeable about things in general. Williams and Lombrozo (2010) demonstrated that explaining why an item belongs to a category promotes generalization, which supports the idea that explanation has a broad effect at exposing illusions. Lombrozo (2006) highlights how explanation promotes understanding across different domains, supporting the idea that explanation has a broader effect than the specificity of the illusion itself. I refer to this model as the Single Broad Effect Model.

Another model, in contrast to the Single Broad Effect Model, is the Two-Effects Model. This model posits that there are two distinct effects of failing to explain something: a target-specific effect, wherein people recognize the limits of their knowledge of the specific item they are explaining, and a breadth effect, wherein people feel less confident about their knowledge more generally. For example, failing to explain how a speedometer works might lead someone to question their understanding of how snow forms for similar reasons to the Single Broad Effect Model. However, the Two-Effects Model claims that failing to explain how a speedometer works might lead someone to question their understanding of how a speedometer works because explanation makes them aware of the gaps in their knowledge of a speedometer (e.g., Keil, 2003). This model also does not view the illusion as an expression of general overconfidence but suggests that one of the effects of failing to explain something might be to reduce confidence in one's overall knowledge in addition to reducing confidence in one’s knowledge of what they explained.

Another account is that the illusion is not separable from the broader phenomenon of general overconfidence (e.g., Merkle & Weber, 2011). This model posits that the illusion is not a unique phenomenon and that any reduction in understanding reported reflects explanation having a general debiasing effect on reducing overconfidence. According to this model, the illusion broadly applies,

meaning that no domain of knowledge is immune to it at baseline. This suggests that explanation should have a debiasing effect across all domains, without any specific exceptions. I refer to this model as the General Overconfidence model.

A final account considered, informed by the results of the initial experiments, was that my results may be attributed to demand effects. Participants who provided an explanation for a phenomenon might have felt compelled to adjust their understanding for phenomena they did not explain in a similar manner. This could be because re-evaluating their understanding of all six phenomena after explaining one may have made it particularly salient that they should make the same adjustments for each phenomenon (Schwarz, 1999; Orne, 1962). This may be due to conversational norms suggesting that being asked the same question twice implies a need to revise one's answer, or perhaps because the research hypothesis is easy to infer from the task (i.e., I'm supposed to adjust downward), or because people are exhibiting an initial elevation bias, a tendency to provide a higher initial rating than subsequent ratings (Frederick & Loewenstein, 1999; Fiedler, 2012). This list cannot be exhaustive but contains some of the most salient possible mechanisms behind a possible demand effect.

To evaluate the competing models, I first consider demand effects and the General Overconfidence model, as these are largely ruled out by the experimental evidence. The General Overconfidence model is undermined by the replication of the (lack of) effect for procedures. In Experiments 5 and 6, participants did not exhibit an illusion of understanding for common procedures, consistent with the original Experiment 9 of Rozenblit and Keil (2002). This suggests that the illusion of understanding is distinct from general overconfidence, as it does not apply to all domains of knowledge. Furthermore, the model cannot account for the ChatGPT coding results that indicate people adjust their reported understanding of the item they explained for different reasons than they adjust their reported understanding of the items they did not explain. I deem this model unsupported.

The demand effects interpretation suggests that participants' responses are influenced by the experimental design rather than a genuine cognitive phenomenon. This interpretation struggles to account for the null results observed in the control conditions of the narratives and procedures replication experiments. In Experiment 4, while participants exhibited an illusion of understanding for popular movie plots in the explanation condition, there was no significant change in the control

condition. Similarly, Experiment 5 showed no effect in the control condition for procedures. The conditions hypothesized to underwrite the demand interpretations in the breadth design (e.g., conversational norms, the hypothesis being easy to predict, people tending to overestimate their first judgment in a series of judgments) also apply to the more traditional design used in these experiments. Even if the breadth design made these factors especially salient to participants, the complete lack of results in the breadth design of Experiment 6 suggests that the demand interpretation of my results holds little weight.

Given the evidence against demand effects and the General Overconfidence model, the focus now shifts to the remaining competing models: the Single Broad Effect Model and the Two-Effects Model. These models propose that explanation has either one broad effect or two distinct effects (a target-specific effect and a breadth effect).

One of the key observations supporting the Two-Effects Model is the varying effect sizes observed between the items within the experiments. Although the breadth effect appeared generally stronger, a closer look at the distribution of understanding ratings across items reveals considerable variation: some items show no difference between the size of the effects, some items show larger target-specific effects, and others tend to show more pronounced breadth effects. If explanation had one broad effect, it would be difficult to account for this variation in the magnitude of target-specific and breadth effects. A stronger breadth effect relative to the target-specific effect would suggest that there is an additional mechanism or effect responsible for the larger effect size. While my data reveal the variability and suggest a trend toward the breadth effect being stronger, it cannot establish the trend by itself. However, when combined with the work of Roeder and Nelson (2015), who found the same trend, the weight of the evidence suggests it is indeed the case that the breadth-effect tends to be larger than the target-specific effect. The Single Broad Effect Model is not supported by this evidence.

The ChatGPT coding results further support the Two-Effects Model. These data revealed a moderate correlation between the difference in how well participants said they understood an item and how well they explained it, and the difference in understanding ratings for that item. In other words, the difference between how well someone thought they understood how a zipper works and the coded quality of their explanation of how a zipper works predicted their reduction in reported understanding of how a zipper works. This serves as direct evidence of the long-assumed target-

specific effect of explanation on the illusion of understanding: people believe they can explain an item well, they are given the chance and do not explain it well, and then they report knowing less about the item in a manner somewhat proportionate to the discrepancy between how well they thought they understood and how well they explained it. However, this correlation was not found for items that were not explained. That is, the difference between how well a person thought they understood how a zipper works and how well they explained it did not predict the reduction in reported knowledge across the understanding ratings for the rest of the items. This suggests that there is a separate mechanism at work producing the breadth effect, which could be a temporary inducement of intellectual humility (Meyers et al., 2020).

Additionally, the Two-Effects Model is consistent with the possibility that there are even more distinct effects of explanation on debiasing the illusion of understanding. For instance, people exhibit an even larger reduction in reported knowledge for more socially desirable topics such as national history (e.g., the history of their own country) compared to less socially desirable topics such as mechanical devices or everyday household tasks (Gaviria et al., 2017; Gaviria & Corredor, 2021). This suggests a further separate mechanism distinct from the target-specific and breadth effects responsible for the debiasing effect of explanation. Therefore, a model positing the existence of at least two distinct debiasing effects, such as the target-specific and breadth effects, provides a more comprehensive framework for understanding the illusion of understanding and its implications across different domains.

A potential issue with the Two-Effects Model is that if both the target-specific and breadth effects are applying, one might expect the item targeted with explanation to consistently show the largest decrease in understanding. This is because the combined influence of both effects should result in a greater reduction in knowledge for the explained item than for the unexplained items. However, this expectation may be flawed if one considers the target-specific effect a proximal effect and the breadth effect a distal effect. When re-evaluating their knowledge of the explained item, individuals are confronted with their specific lack of understanding, making the proximal effect more salient. They do not need the distal effect, which possibly influences their general sense of intellectual humility, to further reduce their reported understanding of that item. Despite it being more logical for people to assume that they have not been made aware of the full extent of their lack of knowledge of the item (i.e., “unknown unknowns”; Smithson, 1989; Waters et al., 2017), if people were capable and motivated to exhibit such behavior, it is unlikely there would be an illusion to study.

Another potential issue is the inconsistency in the size of the breadth effect across different items. If the breadth effect induces general intellectual humility, its impact should be relatively uniform, but this was not the case. I suggest the observed variability across items likely results from item-specific factors. For instance, item familiarity (Gaviria et al., 2017), complexity (Rozenblit & Keil, 2002), and social desirability (Gaviria & Corredor, 2021) have each been demonstrated to impact the illusion of understanding. As my research did not attempt to control these factors beyond ensuring there were at least two items in each domain, I cannot offer direct evidence of their impact or lack thereof. While this is intellectually unsatisfying, the core idea that there are additional uncontrolled factors that might have contributed to the variance in effect sizes across items is sound in principle and random error might be a further source. But the soundness of principles when applied to specific cases must be borne out by evidence, and random error can be ever reduced with more data. Thus, it would be highly beneficial for future work to systematically probe various item-level factors, such as those discussed above, to better understand their influence on the target-specific and breadth effects of explanation on the illusion of understanding.

4.2 Relevance of the Two-Effects Model of the illusion of understanding

My work suggests that explanation might prove to be a robust, real-world debiasing technique. Research on science communication, misinformation, and pseudo-profound bullshit collectively highlight the importance that scientists place on what people understand, and critically, what people think they understand. This makes sense, as overestimating how much one knows is at the heart of pervasive cognitive biases (Fischhoff et al., 1977; Dunning et al., 2003; Kruger & Dunning, 1999; Moore & Healy, 2008). As such, there is no shortage of attempts to reduce people's overconfidence (Arkes, 1987; Croskerry, 2013; Ferretti, 2016; Larrick, 2004; Lichtenstein & Fischhoff, 1980). However, few techniques are widely effective. Explanation as a debiasing method can be indirect and nonconfrontational. This has implications for various domains, including educational settings and the broader public sphere.

Recognizing the existence of both a target-specific and a breadth effect of explanation has significant implications for cognitive psychology and practical applications. The target-specific effect reinforces the idea that individuals can recalibrate their understanding of specific phenomena when confronted with their lack of detailed knowledge (Rozenblit & Keil, 2002). This aligns with existing educational practices where students are encouraged to articulate their understanding as a means of

identifying knowledge gaps (Chi et al., 1994; Lombrozo, 2006). By systematically identifying and addressing these gaps, learners can achieve a more accurate and comprehensive understanding of the material.

The breadth effect, on the other hand, suggests a more generalized form of intellectual humility that can have far-reaching consequences. If explaining one phenomenon can induce a broader sense of humility about one's knowledge base, it implies that the act of explanation can be used to mitigate overconfidence across various domains. This is particularly relevant in contexts where overconfidence can lead to poor decision-making, such as in medical, legal, or financial settings (Croskerry, 2013; Larrick, 2004). Implementing strategies that incorporate explanation tasks could enhance decision-making accuracy and reduce the likelihood of errors stemming from overconfidence. More broadly, people might be more willing to engage with information that contradicts their controversial or strongly held beliefs after failing to explain something unrelated. While people do not generally entrench further in their positions after being confronted (Nyhan, 2021; Swire-Thompson et al., 2020), they often fail to move after being confronted. Thus, explanation offers a way to avoid confrontation while still encouraging someone to question whether they know something.

Moreover, understanding these dual effects of explanation can inform interventions aimed at combating misinformation (e.g., Littrell, Meyers, & Fugelsang, 2024). In an era where misinformation is rampant, particularly through social media, fostering a general sense of intellectual humility could make individuals more critical of the information they encounter and less likely to accept falsehoods at face value (Lewandowsky et al., 2012; Vraga & Bode, 2017). This could be achieved through educational programs that emphasize the importance of explaining and critically evaluating information. By promoting these strategies, it becomes possible to reduce overconfidence and enhance the public's ability to discern credible information from misinformation. This approach has broad implications not only for individual learning and decision-making but also for societal resilience against the spread of misinformation and the cognitive biases that fuel it. Additionally, the dual effects model can enhance collaborative and interdisciplinary work. When individuals recognize the limits of their own knowledge and remain open to learning from others, collaboration becomes more effective. This is crucial in fields that require integrated approaches, such as climate science, public health, and engineering, where complex problems cannot be solved by a single discipline alone

(Fiore, 2008). Encouraging explanation and intellectual humility can thus facilitate better communication and problem-solving in these contexts.

4.2.1 Additional insights of this research

One additional insight from this research is the importance of including control conditions when investigating the illusion of understanding. Control conditions allow researchers to isolate the effects of explanatory tasks from other factors that may influence participants' self-assessed understanding. However, including control conditions in this research is not standard, albeit it is becoming increasingly common (e.g., Johnson et al., 2016; Kirk, 2020; Virtiol & Marsh, 2018; Wilson & Marsh, 2024). In my experiments, control conditions were essential in distinguishing between genuine cognitive phenomena and potential demand effects. For instance, the lack of reduction in understanding in control conditions for narratives and procedures indicated that the observed reduction in understanding in the experimental conditions were not simply artifacts of the experiment design (Lewandowsky et al., 2012). This was important because in the control conditions of Experiments 2 and 3, I found a significant but small reduction in understanding. This suggested that some portion of the reduction in understanding was not due to explaining anything. However, this effect was not consistent, even in the breadth design as evidenced when procedures were included as stimuli in Experiment 6. This suggests that there is variability in whether the control condition will show an effect. Granted, I have not collected enough evidence to completely rule out the claim that the breadth design could be artificially creating the effect. However, the possibility that repeated measurement might affect the results suggests it should be controlled for.

Despite being a small effect (i.e., $d = 0.10$), repeated measurement might account for some explanatory depth effects published to date. For instance, it is possible that the effects of reflecting on one's explanatory ability (stopping just before the actual explanation would be generated) can be largely explained by the act of providing two ratings (e.g., Johnson et al., 2016). This could be because people tend to take advantage of the opportunity to change their answer after having additional time for reflection (Pennycook et al., 2014; Thompson et al., 2011). It may also be because people's understanding of the response scale (their understanding of the word "understanding") shifts across ratings (see Ostrom & Upshaw, 1968; Petty & Wegener, 1993). However, neither mechanism can fully account for why the process of explanation reduces self-reported understanding. Thus,

future research should explore to what extent reductions in self-reported understanding are caused by processes other than explanation and whether this can be identified with any consistency.

Another contribution of this research is the demonstration of the utility of ChatGPT for coding qualitative responses. Traditionally, coding qualitative data is labor-intensive, time-consuming, and prone to human error and bias (Saldaña, 2015). By leveraging ChatGPT, I was able to efficiently and reliably code participants' explanations, providing quantitative insights into the quality of their responses. This approach not only saved considerable time but also allowed for a more standardized evaluation of explanation quality. The moderate correlation found between participants' self-assessed understanding and ChatGPT-coded explanation quality underscores the validity of this method. Recent studies have shown that AI, including ChatGPT, can effectively code qualitative data, enhancing consistency and reducing biases (Hoyt et al., 2022; Lu & Stepchenkova, 2015). This innovative use of AI in psychological research opens up new possibilities for large-scale qualitative data analysis, making it feasible to process and analyze extensive datasets that were previously impractical to handle manually. Future research can build on this by further refining AI coding algorithms and exploring their applications across various domains of psychological inquiry (Yin et al., 2020; Cowls et al., 2021) as well as compare the validity of its coding to human coders.

Lastly, this research has successfully replicated and extended a boundary condition of the illusion of understanding: the finding that common procedures do not exhibit the illusion and can inoculate other stimuli against it. This discovery is pivotal because it highlights that not all types of knowledge are susceptible to the illusion of understanding. Understanding why procedures are immune to this effect can provide deeper insights into the cognitive mechanisms underlying the illusion. One theoretical suggestion is that procedures involve a series of well-defined steps that are more concrete and less abstract than other types of knowledge, making them less prone to overestimation (Alter et al., 2010). Additionally, the hands-on, experiential nature of learning procedures might lead to a more accurate self-assessment of understanding. Future work should aim to delineate these unique characteristics of procedures and explore whether similar inoculating effects can be observed in other domains of concrete, step-by-step knowledge. This line of inquiry could significantly enhance our theoretical understanding of the illusion of understanding and its boundaries (Keil, 2003; Chi, 2009).

4.3 Limitations and Conclusion

There are at least three limitations of the present work that deserve consideration. One limitation is that the current effects might only occur when questions of understanding are asked by an experimenter, even if the experimenter was not physically present. Whether an educator, a parent, or a peer in the same role would observe the same effect deserves further study. The role of social context and authority in eliciting these responses could significantly influence the outcomes, suggesting that the observed effects might be unique to the experimental setting. Future research should investigate whether these effects are replicable in more naturalistic settings to determine their generalizability (Hawthorne, 2008).

A second limitation is that the debiasing effect might not apply to new phenomena introduced only after explanation, not before. In this work, people reported knowing less about the phenomenon they had previously rated but did not explain. However, had they only rated the phenomenon after explanation (their first rating as the post-explanation rating), they may not have rated it any differently than they would have in a pre-explanation (first) rating. This raises questions about the timing and sequencing of the explanatory tasks and their influence on perceived understanding. Future research should explore whether the order in which phenomena are explained and rated affects the debiasing impact.

The third limitation is that my data do not indicate the duration of the debiasing effect. My experience suggests the effect is more likely to be short-lived than long-lasting. Understanding the temporal dynamics of this effect is crucial for determining its practical applications. For instance, if the effect dissipates quickly, its utility in educational or professional settings might be limited. Future research should focus on longitudinal studies to estimate the duration of the debiasing effect, thus providing a clearer picture of its long-term benefits and potential for sustained impact (Wilson & Brekke, 1994).

In conclusion, while this research has demonstrated two effects of explanation on the illusion of understanding, it is essential to acknowledge these limitations, especially when considering its use as a real-world debiasing technique. Addressing these gaps in future studies will not only enhance our understanding of the phenomenon but also improve our practical application of it. By delving deeper into these aspects, researchers can transform our grasp of this cognitive bias from a mere illusion to a well-grounded understanding.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451. https://doi.org/10.1037/a0020218

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*(1), 133– 144. https://doi.org/10.1016/0749-5978(87)90049-5

Bart, W. M. (2014). On the effect of chess training on scholastic achievement. *Frontiers in Psychology, 5.* doi:10.3389/fpsyg.2014.00762

Cadario, R., Longoni, C. & Morewedge, C.K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behavior, 5*, 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. https://doi.org/https://doi.org/10.1016/0364-0213(94)90016-7

Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: Three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman's (2013) findings. *Psychological Science, 32*(4), 611–621. https://doi.org/10.1177/0956797620972367

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety, 22*(2), 58-64. https://doi.org/10.1136/bmjqs-2012-001712

Detterman, D. K. (1993). *The case for the prosecution: Transfer as an epiphenomenon.* In D. K. Detterman & R. J. Sternberg (Eds.), Transfer on trial: Intelligence, cognition, and instruction (pp. 1–24). Ablex Publishing.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. Psychological Science, 24(6), 939–946. https://doi.org/10.1177/0956797612464058

Ferretti, V., Guney, S., Montibeller, G., & von Winterfeldt, D. (2016). Testing best practices to reduce the overconfidence bias in multi-criteria decision analysis. Proceedings of the 49th Annual Hawaii International Conference on System Sciences.1547-1555. DOI: 10.1109/HICSS.2016.195

Gaviria, C., Corredor, J. (2021). Illusion of explanatory depth and social desirability of historical knowledge. Metacognition Learning, 16, 801–832. https://doi.org/10.1007/s11409-02109267-7

Gaviria, C., Corredor, J., & Rendon, Z. Z. (2017). "If it matters, I can explain it": Social desirability of knowledge increases the illusion of explanatory depth. Proceedings of the 39th Annual Meeting of the Cognitive Science Society.

Gretton, J. D. (2017). Perceived breadth of bias as a determinant of bias correction. [Unpublished doctoral dissertation]. Ohio State University.

Johnson, D. R., Murphy, M. P., & Messer, R. M. (2016). Reflecting on explanatory ability: A mechanism for detecting gaps in causal knowledge. Journal of Experimental Psychology. General, 145(5), 573–588. https://doi.org/10.1037/xge0000161

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. Psychological Bulletin, 145(2), 165–188. https://doi.org/10.1037/bul0000180

Littrell, S., Risko, E. F., & Fugelsang, J. A. (2021). The bullshitting frequency scale: Development and psychometric properties. British Journal of Social Psychology, 60. 248-270. https://doi.org/10.1111/bjso.12379

Littrell, S., Meyers, E. A., & Fugelsang, J. A. (2022). Not all bullshit pondered is tossed: Reflection decreases receptivity to some types of misinformation but not others. PsyArXiv. https://doi.org/10.31234/osf.io/4bstf

Kleinberg, S., & Marsh, J. K. (2020). Tell me something I don't know: How perceived knowledge influences the use of information during decision making. Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.

https://cogsci.mindmodeling.org/2020/papers/0410/index.html

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Meyers, E. A., Turpin, M. H., Bialek, M., Fugelsang, J. A., & Koehler, D. J. (2020). Inducing feelings of ignorance makes people more receptive to expert (economist) opinion. *Judgment and Decision Making, 15*(6), 909–925.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology, 87*(1), 1–32. https://doi.org/10.1016/j.jecp.2003.09.003

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences, 118*(15), e1912440117. https://doi.org/10.1073/pnas.1912440117

Ostrom, T. M. & Upshaw, H . S. (1968*). Psychological perspective and attitude change*. In A G. Greenwald, T. C. Brock & T. M. Ostrom (Eds.), Psychological foundations of attitudes (pp. 217-242). San Diego, CA: Academic Press.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 544–554. https://doi.org/10.1037/a0034887

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology, 1*(9), 524–536. https://doi.org/10.1038/s44159-022-00081-9

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual representation in a community of knowledge. *Trends in Cognitive Sciences, 23*(10), 891–902. https://doi.org/10.1016/j.tics.2019.07.011

Roeder, S., & Nelson, L. (2015). Folk theories are corrupted by cross-domain explanations [Unpublished manuscript]. Haas School of Business, University of California, Berkeley. https://dx.doi.org/10.2139/ssrn.2622301

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science, 26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Schwarz, N., & Bless, H. (1992). *Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment*. In L. L. Martin & A. Tesser (Eds.), The construction of social judgments (pp. 217–245). Lawrence Erlbaum Associates, Inc.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences, 115*(1), 15–23. https://doi.org/10.1073/pnas.1712277115

Sloman, S. A., & Vives, M.-L. (2022). Is political extremism supported by an illusion of understanding? *Cognition, 225*, 105-146. https://doi.org/https://doi.org/10.1016/j.cognition.2022.105146

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition, 9*(3), 286–299. https://doi.org/https://doi.org/10.1016/j.jarmac.2020.06.006

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Vitriol, J. A. and Marsh, J. K. (2018), The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology, 48*, 955-969. https://doi.org/10.1002/ejsp.2504

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science, 19*(7), 645–647.

Williams, J. J., & Lombrozo, T. (2010a). *Explanation constrains learning, and prior knowledge constrains explanation*. In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Cognitive Science Society.

Williams, J. J., & Lombrozo, T. (2010b). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science, 34*(5), 776-806.

Wong, R. M. F., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction, 12*(2), 233–262. https://doi.org/https://doi.org/10.1016/S0959-4752(01)00027-5

Zeveney, A., & Marsh, J. K. (2016). *The illusion of explanatory depth in a misunderstood field: The IOED in mental disorders*. In A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive Science Society. Cognitive Science Society.

Appendix A
Experiment 1 Supplemental Analyses

4.3.1 Confirmatory Analyses

I deviated somewhat from the preregistration protocol.[5] I analyzed the effect of (attempted) explanation on self-reported understanding in a linear mixed model with fixed effects of Phenomenon Judged, Time, and Phenomenon Explained, and random effects of Phenomenon Judged nested within participants. I effect-coded Time. I also effect-coded Phenomenon Explained by contrasting whenever the Phenomenon Judged matched the Phenomenon Explained (e.g., rated zipper and explained zipper) with whenever the Phenomenon Judged did not match the Phenomenon Explained (e.g., rated zipper and explained sewing machine). All fixed effect predictors were allowed to interact, and the random effect predictors were allowed to interact.

I found a main effect of Time: understanding ratings significantly decreased from pre- to post-explanation, $F(1, 1428) = 278.48$, $p < .001$. Notably the Time by Phenomenon Explained interaction was not significant, $F(1, 1097)$, $p = .067$. This suggests that the reduction in understanding of, for example, a zipper (device) was similar when a zipper was explained and when something other than a zipper was explained. However, as can be seen in Figure 1 in the main text, it could be that explaining something other than a zipper leads to an even greater reduction in understanding of a zipper. I caution against such an interpretation. This is because understanding ratings for some items appear to decrease more when the same object is explained, and others decrease more when something else is explained. This can be observed in Figure 2 in the main text, which displays the mean understanding ratings for each Phenomenon Rated by Phenomenon Explained.

4.3.2 Similarity Analyses

In Study 1, I collected similarity ratings on a 1 (Not at all similar) to 5 (Extremely similar) scale for each possible pair of devices. This summed to 15 ratings in total. Figures S1a-f below display 6 plots that show for each device, the average similarity (with accompanying 95% confidence

[5] I incorrectly preregistered a model that contained Time as a random effect. Time should only be entered as a fixed effect and so I report a model not including Time as a random effect. I also preregistered a model that effect-coded Phenomenon Judged. This was incorrect and so I report a model where Phenomenon Judged is not effect-coded. This applies to all linear mixed model analyses reported in this Appendix.

intervals) with each other device. These figures demonstrate that similarity does not mediate the effect of reducing the estimate of understanding of one device after explaining a different one. For instance, figure S1a shows that the sewing machine was rated as more similar to a zipper than any other device compared. Yet, consulting Figure 2 of the main manuscript, the reduction in knowledge of a zipper when a sewing machine was explained was highly similar to the reduction in knowledge of a zipper when any other item was explained. Similarly, the reduction in knowledge of a sewing machine when a zipper was explained was highly similar to the reduction in knowledge of a sewing machine when any other item was explained. This should illustrate that similarity between devices cannot predict the extent to which a person reduces their estimate of one device after explaining a different one.

4.3.2.1 Similarity Analyses

In Study 1, I collected similarity ratings on a 1 (Not at all similar) to 5 (Extremely similar) scale for each possible pair of devices. This summed to 15 ratings in total. Figures S1a-f below display 6 plots that show for each device, the average similarity (with accompanying 95% confidence intervals) with each other device. These figures demonstrate that similarity does not mediate the effect of reducing the estimate of understanding of one device after explaining a different one. For instance, figure S1a shows that the sewing machine was rated as more similar to a zipper than any other device compared. Yet, consulting Figure 2 of the main manuscript, the reduction in knowledge of a zipper when a sewing machine was explained was highly similar to the reduction in knowledge of a zipper when any other item was explained. Similarly, the reduction in knowledge of a sewing machine when a zipper was explained was highly similar to the reduction in knowledge of a sewing machine when any other item was explained. This should illustrate that similarity between devices cannot predict the extent to which a person reduces their estimate of one device after explaining a different one.

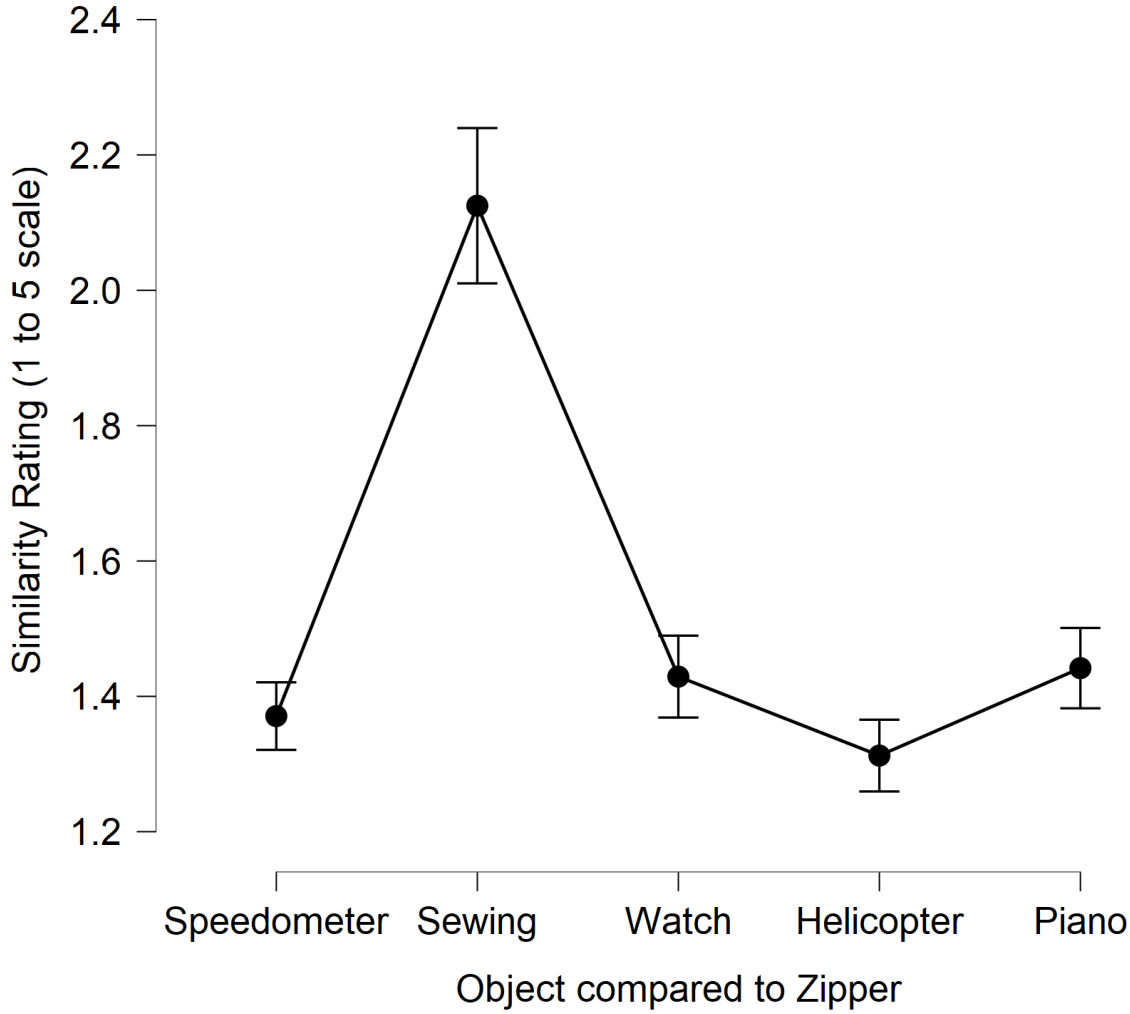


Figure A.1. Average similarity ratings between a zipper and a given object. Error bars are 95% within-subject confidence intervals.

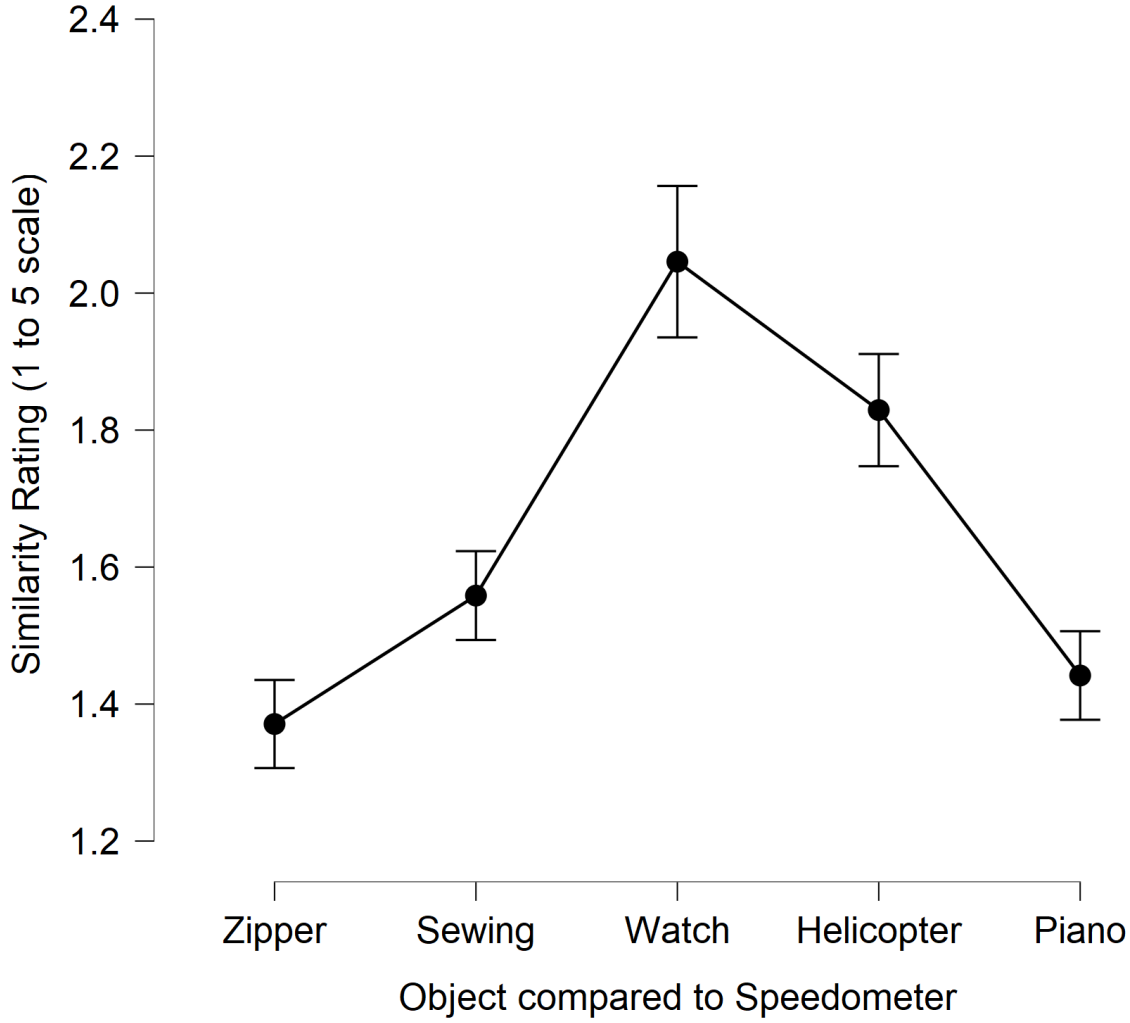


Figure A.2. Average similarity ratings between a speedometer and a given object. Error bars are 95% within-subject confidence intervals.

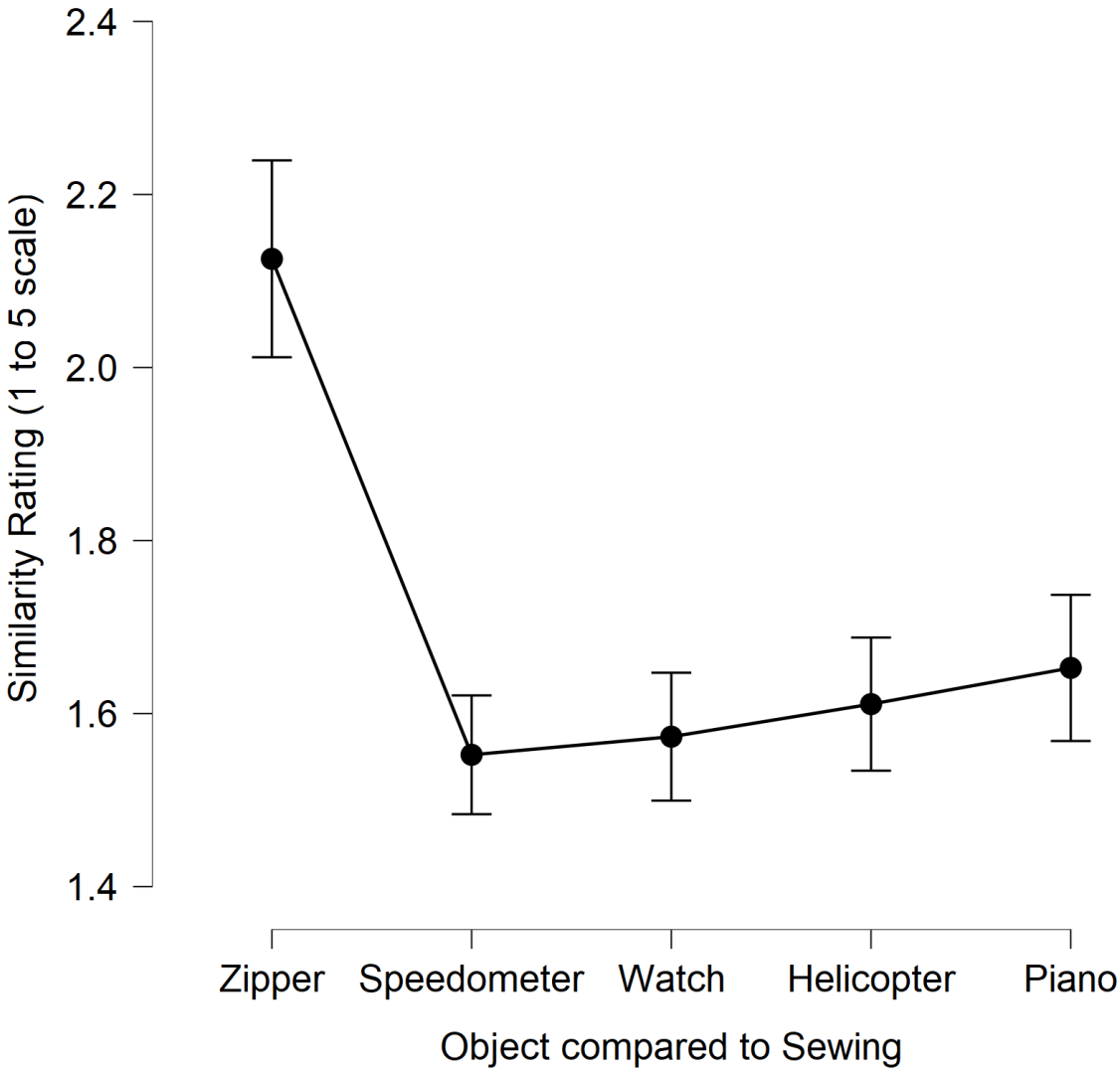


Figure A.3. Average similarity ratings between a sewing machine and a given object. Error bars are 95% within-subject confidence intervals.

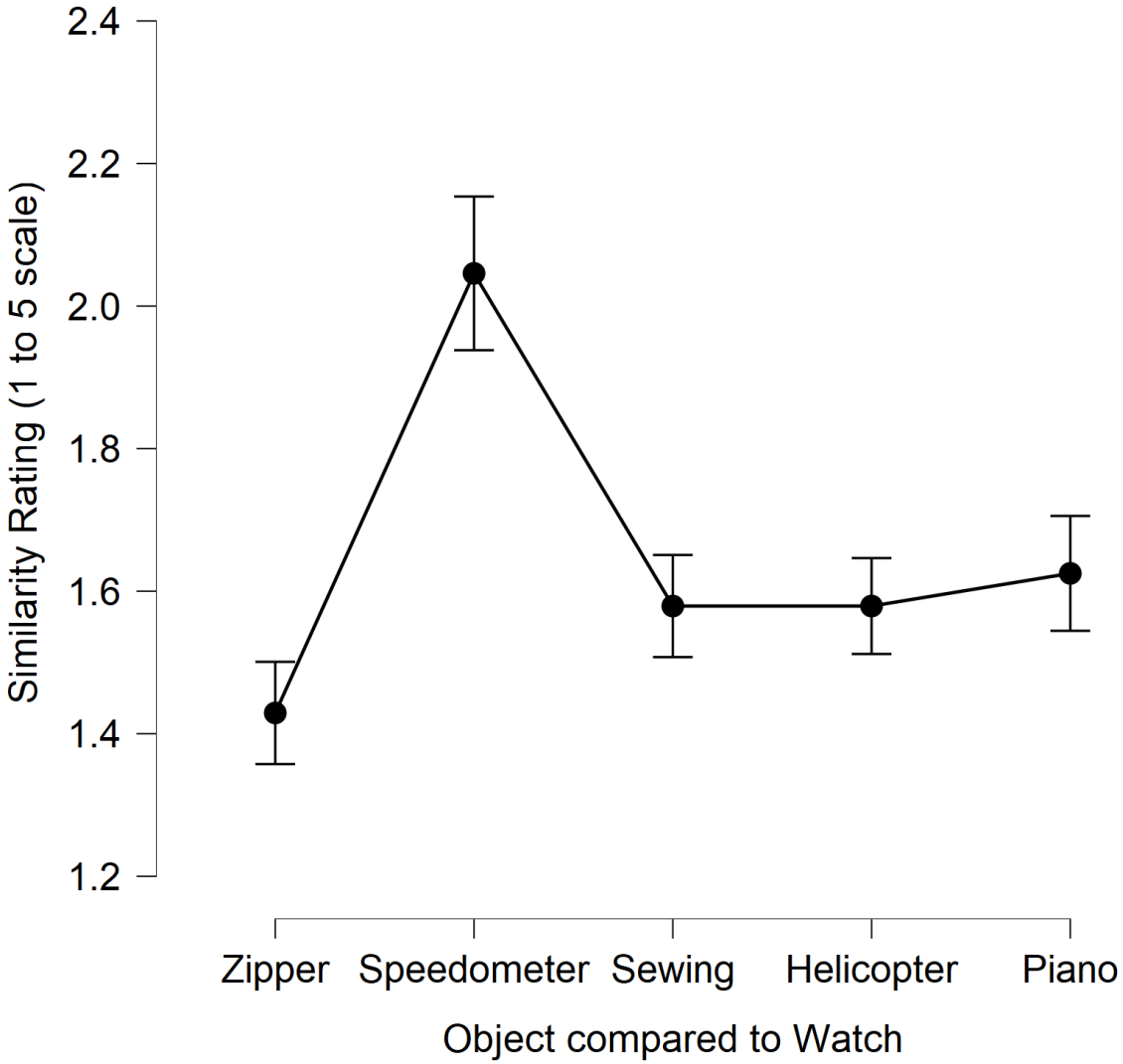


Figure A.4. Average similarity ratings between a quartz watch and a given object. Error bars are 95% within-subject confidence intervals.

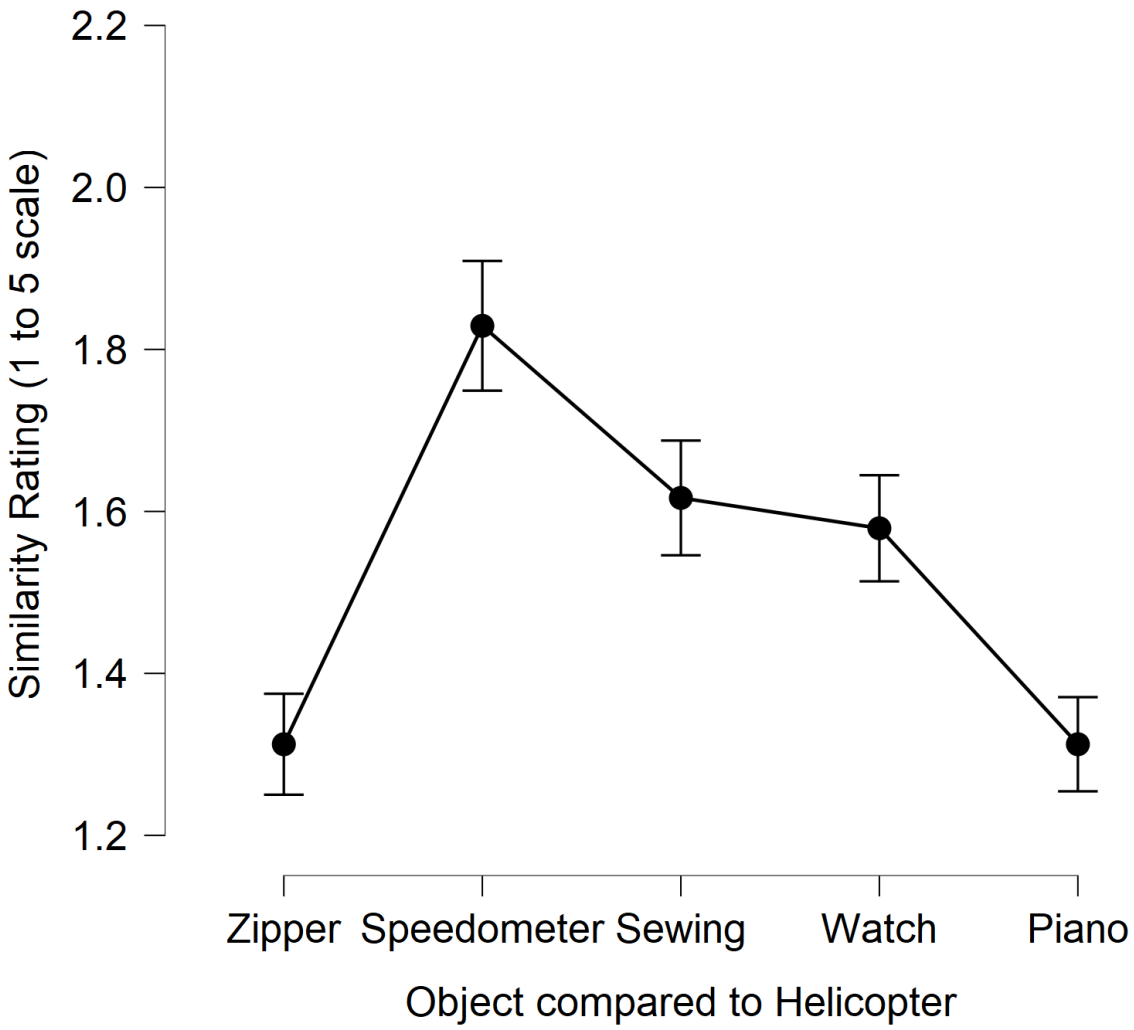


Figure A.5. Average similarity ratings between a helicopter and a given object. Error bars are 95% within-subject confidence intervals.

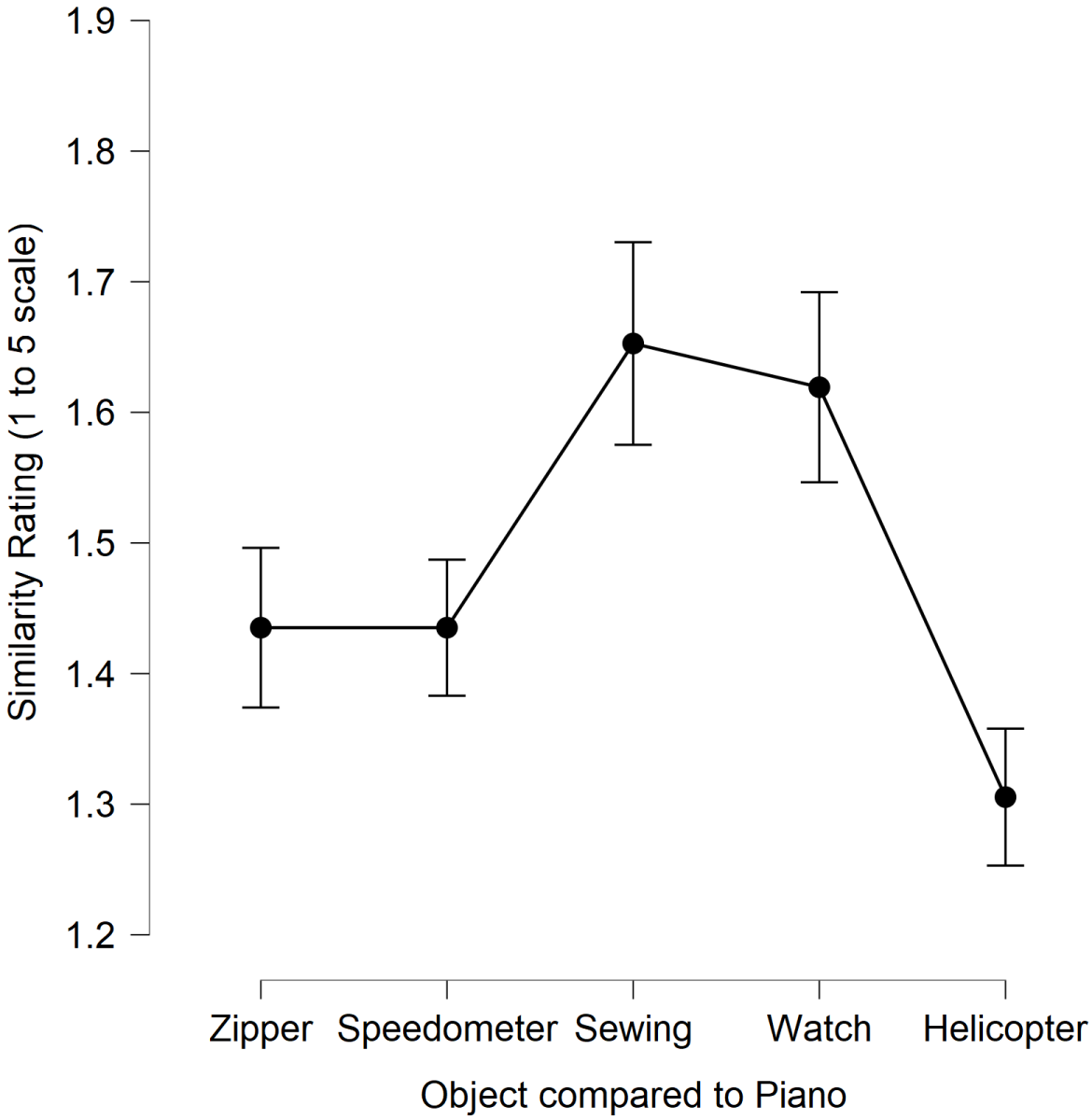


Figure A.6. Average similarity ratings between a piano and a given object. Error bars are 95% within-subject confidence intervals.

This is further supported by exploratory analyses. I computed a linear mixed model with Similarity Rating as the dependent variable and Pair as the fixed effect predictor variable nested within a random participant intercept. Pairs that were significantly more (or less) similar relative to the similarity ratings of the pair Speedometer-Zipper are indicated in Table S1 below. These correlations confirm the pattern noted in the above paragraph. For instance, the sewing machine is considered much more similar to a Zipper than it is to any of the other devices yet the reduction in a Sewing Machine after having explained a zipper looks is similar to the when the rest of the items are explained. Similarly, the reduction in zipper after explaining a Sewing Machine is similar to the when the rest of the items are explained.

Table S1. Correlations of Similarity ratings between device pairs, statistically compared to the Speedometer-Zipper pair.

	Helicopter	Piano	Quartz Watch	Sewing Machine	Speedometer	Zipper
Helicopter	—	-.06	.21*	.25*	.46*	-.06
Piano		—	.25*	.29*	.07	.07
Quartz Watch			—	.21*	.67*	.06
Sewing Machine				—	.19*	.75*
Speedometer					—	NA
Zipper						—

Note. An asterix indicates that the correlation is significantly different (at the .001 level) from the correlation between the similarity ratings of the Zipper and Speedometer pair.

Appendix B
Experiment 2 Supplemental Analyses

4.3.4 Confirmatory Analyses

I analyzed the effect of explaining on self-reported understanding in a linear mixed model with fixed effects of Phenomenon Judged, Time, and Phenomenon Explained, and random effects of Phenomenon Judged nested within participants. I effect-coded Time. I also effect-coded Phenomenon Explained by contrasting whenever the Phenomenon Judged matched the Phenomenon Explained with whenever the Phenomenon Judged did not match the Phenomenon Explained. All fixed effect predictors were allowed to interact, and the random effect predictors were allowed to interact.

I replicated the findings of Experiment 1. I found a main effect of Time: understanding ratings significantly decreased from pre- to post-explanation, $F(1, 924) = 145.83$, $p < .001$. Notably the Time by Phenomenon Explained interaction was not significant, $F(1, 924) = 0.93$, $p = .335$. This suggests that the reduction in understanding of, for example, a zipper did not differ between whether a zipper was explained or something other than a zipper was explained.

I then compared the reduction in understanding between providing an explanation (experimental condition) and not providing an explanation (control condition) using a similar mixed effect model in which Writing Condition replaced Phenomenon Explained as a fixed effect. I found a significant Time by Writing Condition interaction, $F(1, 1956) = 29.94$, $p < .001$.[6] As Figure 1 shows, people reported knowing less about devices after attempting to explain how a device works compared to not attempting to explain how a device works. However, post-hoc tests revealed that even in the no-explanation control condition, people reported knowing less about devices when understanding was re-assessed, $z = 6.41$, $p < .001$. This suggests that some portion of the reduction in understanding is not attributable to having explained something.

[6] A less complex model with Understanding Rating entered as dependent variable and Time and Writing Condition entered as predictors (in an Analysis of Variance) revealed a non-significant Time by Writing Condition interaction, $F(1, 3932) = 3.72$, $p = .054$.

Appendix C
Experiment 3 Supplemental Analyses

4.3.5 Confirmatory Analyses

I analyzed the effect of explaining on self-reported understanding in a linear mixed model with fixed effects of Phenomenon Judged, Time, and Phenomenon Explained, and random effects of Phenomenon Judged nested within participants. I effect-coded Time. I also effect-coded Phenomenon Explained by contrasting whenever the Phenomenon Judged matched the Phenomenon Explained with whenever the Phenomenon Judged did not match the Phenomenon Explained. All fixed effect predictors were allowed to interact, and the random effect predictors were allowed to interact.

I first analyzed the reduction in understanding when an explanation was provided (not including participants in the control condition). I found a main effect of Time: understanding ratings significantly decreased from pre- to post-explanation, $F(1, 690) = 87.33$, $p < .001$. I also found a significant Time by Phenomenon Explained interaction, $F(1, 690) = 4.52$, $p = .049$. As Figure 1 suggests, explaining something else may lead to a greater reduction in understanding, however, as in Experiment 1 I suggest this finding should be interpreted with caution. This is reflected in the non-significant (what some might term "marginal") three-way interaction between Time, Phenomenon Explained, and Phenomenon rated, $F(1, 690) = 2.09$, $p = .065$. I computed exploratory post-hoc contrasts for each Phenomenon Rated by Phenomenon Explained across Time, correcting for multiple comparisons. I found the greatest difference in contrast estimates for the phenomena "Snow" and "Speedometer". In these cases, when the phenomenon itself was explained, people did not report understanding the phenomenon significantly less (Snow: $B = -0.22$, $p = .315$; Speedometer: $B = -0.10$, $p = .673$) and the trend of the data suggests somewhat greater understanding of the phenomenon. However, when a different phenomenon was explained, people reported understanding Snow and Speedometer significantly more (Snow: $B = 0.47$, $p < .001$; Speedometer: $B = 0.44$, $p < .001$). People behaved a bit differently for the other phenomena (Earthquake, Immigration, Trade, Zipper). In these cases, participants always reported a nominally lower understanding of the phenomenon regardless of whether they explained the same phenomenon or a different one.

I then compared the reduction in understanding between providing an explanation (experimental condition) and not providing an explanation (control condition) using a similar mixed

effect model in which Writing Condition replaced Phenomenon Explained as a fixed effect. I found a significant Time by Writing Condition interaction, $F(1, 1572) = 24.22$, $p < .001$. As Figure 1 shows, people reported knowing less about how phenomena occur after attempting to explain how they work, and this reduction was greater than after not attempting to explain how they work. However, post-hoc tests revealed that people still reported knowing less about phenomena even without attempting to explain how they work, $z = 4.74$, $p < .001$. Consistent with Experiment 2, this suggests that some portion of the reduction in understanding is not attributable to having explained something.

Appendix D

Experiment 7 Supplemental Analyses

4.4

4.5 Distribution Figures

Below are the supplemental figures for Experiment 1, 3, and 6 displaying the Explanation Quality, Time 1 Rating, and Time 2 Rating of each item.

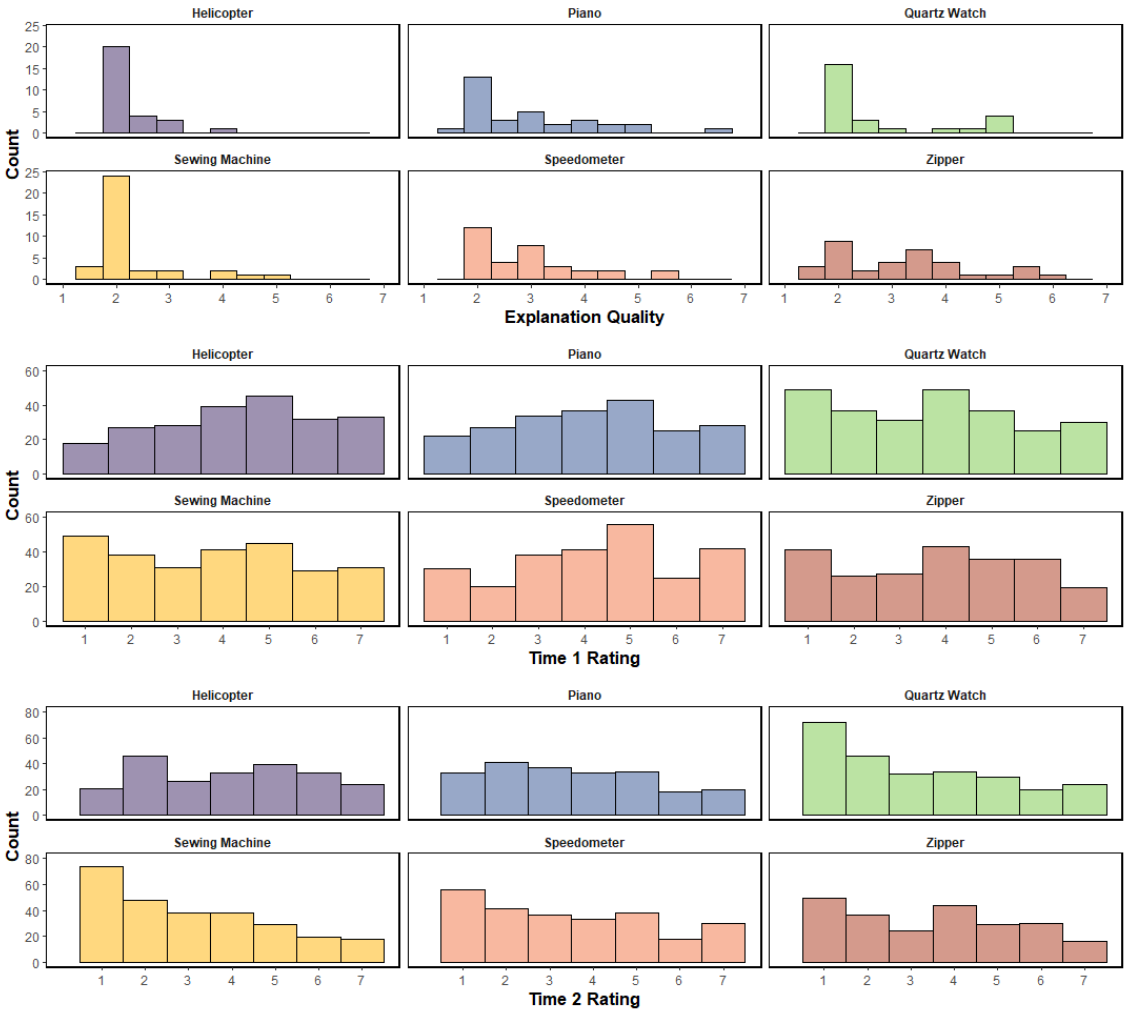


Figure D.1. Distribution of Explanation Quality, Time 1 Ratings, and Time 2 Ratings for Experiment 1 by Item Explained. The top row represents the distribution of explanation quality ratings for six items: Helicopter, Piano, Quartz Watch, Sewing Machine, Speedometer, and Zipper. The middle row presents the distribution of Time 1 understanding ratings, and the bottom row shows the distribution of Time 2 understanding ratings for the same items. Y-axes denote the count of participants, and X-axes indicate explanation quality or understanding ratings on a 1–7 scale.

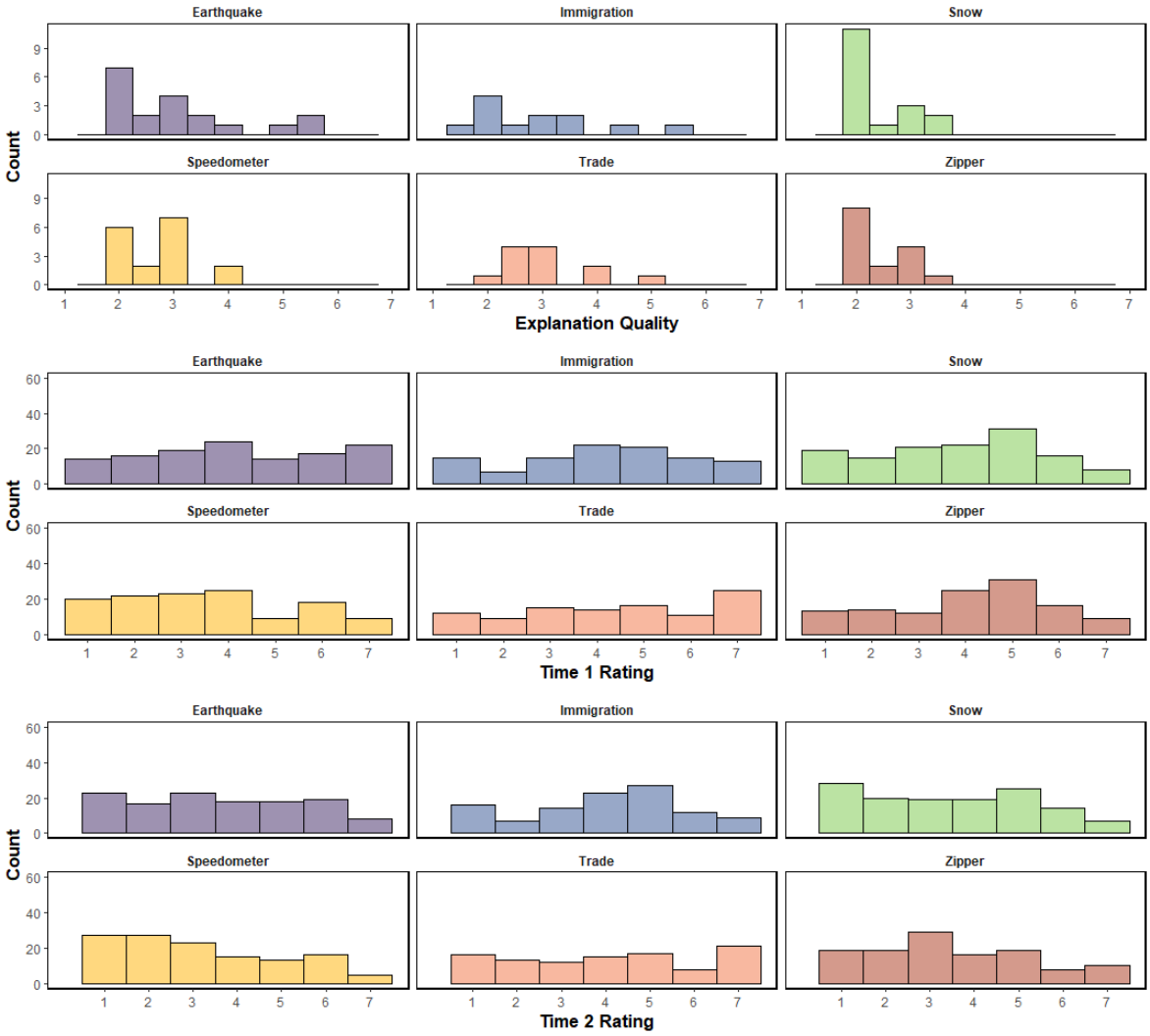


Figure D.2. Distribution of Explanation Quality, Time 1 Ratings, and Time 2 Ratings for Experiment 3 by Item Explained. The top row illustrates the distribution of explanation quality ratings across six items: Earthquake, Immigration, Snow, Speedometer, Trade, and Zipper. The middle row shows the distribution of Time 1 understanding ratings across the same items. The bottom row displays the distribution of Time 2 understanding ratings. Y-axes represent the count of participants, and X-axes represent explanation quality or understanding ratings on a 1–7 scale.

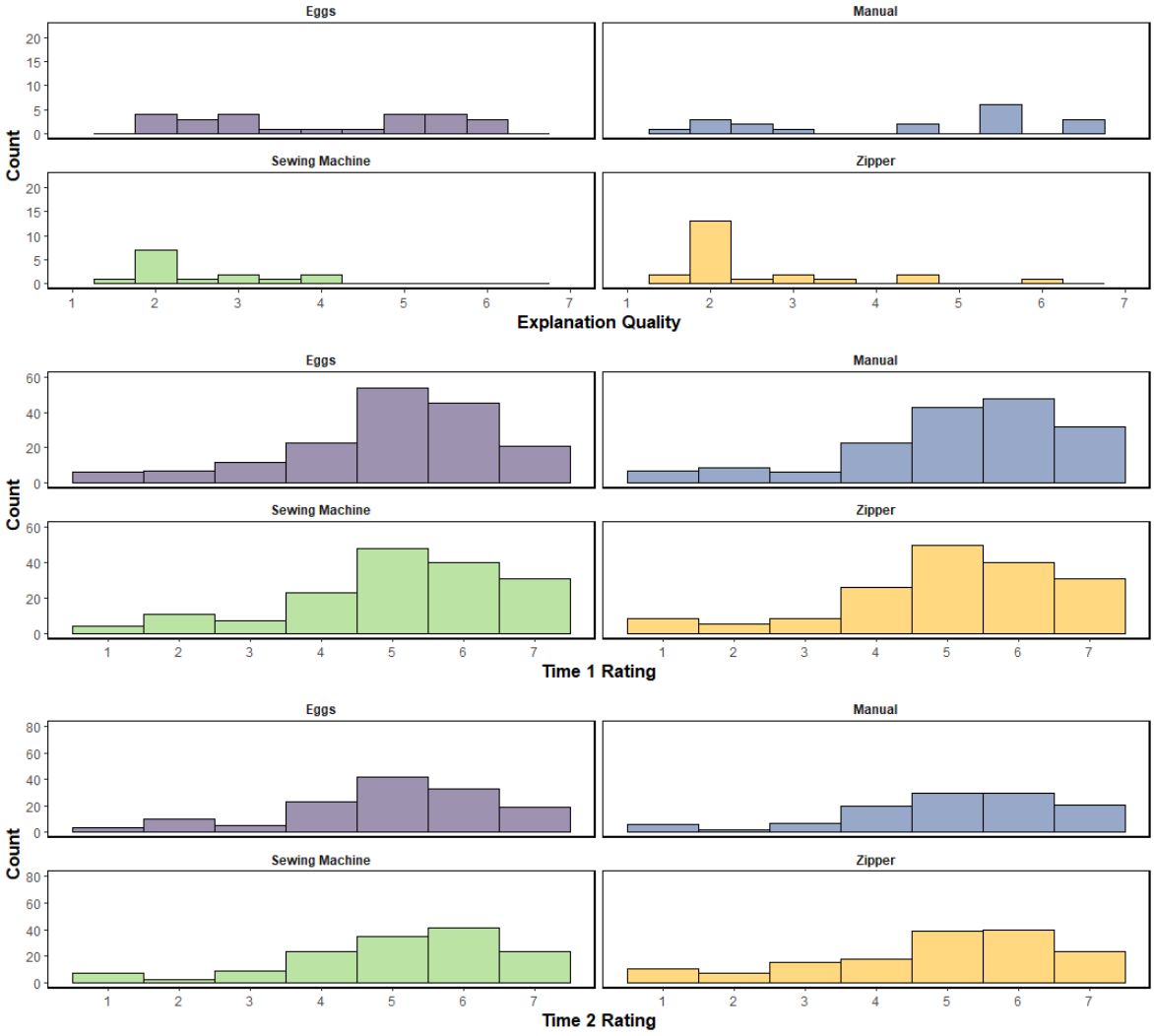


Figure D.3. Distribution of Explanation Quality, Time 1 Ratings, and Time 2 Ratings by Item Explained in Study. This figure presents three sets of histograms across four items: Eggs, Manual, Sewing Machine, and Zipper. The top row details the distribution of explanation quality within the context of a writing condition. Each subsequent row displays the understanding ratings at two different time points—Time 1 and Time 2, respectively. Each facet shows the count of ratings across a 1-7 scale, with the x-axis representing the rating scale and the y-axis denoting the number of participants. The colors across the histograms signify different items, illustrating the variability in response distributions among the items tested.