Judgment and Decision Making (2023), **Vol. 18**:e24 1–18
doi:10.1017/jdm.2023.24

EMPIRICAL ARTICLE

Broad effects of shallow understanding: Explaining an unrelated phenomenon exposes the illusion of explanatory depth

Ethan A. Meyers, Jeremy D. Gretton, Joshua R. C. Budge, Jonathan A. Fugelsang and Derek J. Koehler

Department of Psychology, University of Waterloo, Waterloo ON, Canada

Corresponding author: Ethan A. Meyers; Email: emeyers@uwaterloo.ca

Received: 21 March 2023; **Revised:** 26 June 2023; **Accepted:** 26 June 2023

Keywords: explanation; illusion of explanatory depth; overconfidence; intellectual humility

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 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 three preregistered experiments, we tested whether the process of explaining one phenomenon (e.g., how a zipper works) would lead someone to report knowing less about a completely different phenomenon (e.g., how snow forms). In each experiment, we found that attempting to explain one phenomenon led participants to report knowing less about various phenomena. For example, participants reported knowing less about how snow forms after attempting to explain how a zipper works. We discuss alternative accounts of the illusion of explanatory depth that might better fit our results. We also consider the utility of explanation as an indirect, non-confrontational debiasing method in which a person generalizes a feeling of ignorance about one phenomenon to their knowledge base more generally.

1. Introduction

People are motivated to understand the world around them, and for good reason. Having a functional model of how 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 you could be allergic to 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 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 manuscript, is the illusion of explanatory depth.

The illusion of explanatory depth occurs when a person overestimates their understanding of a topic (Rozenblit & Keil, 2002). This applies to various 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 (Mills & Keil, 2004; Rozenblit & Keil, 2002). These exceptions distinguish the illusion from general overconfidence (see Moore & Healy, 2008).

When a person holds an illusion of explanatory depth, 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 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 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.

There are several proposed mechanisms for why people tend to overestimate their knowledge of a phenomenon. We 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; but see Rozenblit & Keil, 2002, for an opposing perspective). 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 focus 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). The illusion is complex, and the variety of mechanisms that may contribute to it underscore its intricacy.

Despite not definitively understanding how it works mechanistically, researchers continue to examine the consequences of exposing an illusion of explanatory depth (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 that they are explaining, and this exposes the illusion of explanatory depth. 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. We 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 struck us 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 is also

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

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 our lab suggests that people may indeed transfer their feelings of ignorance to a phenomenon other than the one explained. We demonstrated that people revise their opinions on economic issues when exposed to professional economists' perspectives rather than random members of the public (Meyers et al., 2020). This effect was observed specifically after participants failed to explain how the issue worked, in contrast to situations where 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 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 participants to revise more to experts than the public. This suggests the possible existence of a 'breadth principle' wherein the illusion of explanatory depth 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; Williams & Lombrozo, 2010a; Wong et al., 2002). This is in part because explanation promotes generalization, as it motivates a person to identify the underlying structure that unifies phenomena (Williams & Lombrozo, 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 explanatory depth 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 explanatory depth is based on the breadth principle rather than the specificity principle, the implications go beyond the illusion of explanatory depth literature. For example, if a student fails to explain a concept in one subject (e.g., Chemistry) and generalizes their feelings of ignorance to what they think they know overall, they may be more open to engaging with unrelated subjects (e.g., philosophy readings), at least temporarily, to the extent that they were previously overconfident (Meyers et al., 2020). This effect may exist more broadly, such that explanation could be an effective, nonconfrontational technique for challenging someone's beliefs and reducing their confidence in those beliefs.

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, we aimed to systematically test whether similar upstream consequences (e.g., reporting less understanding of a phenomenon than originally believed) would also occur. One way to test this is to have participants re-evaluate their understanding of items they previously rated but did not explain. To the best of our knowledge, this had not been attempted before.

After completing this work, we became aware of an unpublished manuscript by Roeder and Nelson (2015), who explored the same question as ours. They conducted three experiments with methods similar to ours, 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

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

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 merely providing 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). Our work, primarily Experiments 2 and 3, which included a no-explanation comparison condition, thus extends the work of Roeder and Nelson (2015). We therefore believe our work represents the strongest test of whether attempting to explain a phenomenon can lead to reporting knowing less about another. We will further discuss Roeder and Nelson (2015) in the general discussion.

2. Experiment 1

2.1. Methods

The materials, data, and analysis script for each experiment are available on the Open Science Framework (OSF) website (https://osf.io/6huwq/). We preregistered each experiment. The preregistration for this experiment 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).

2.1.1. Participants

Following our preregistered plan to obtain 90% power to detect a within-subjects effect of $d = 0.2$, we recruited 240 participants via Mechanical Turk. The participants were residents of the United States 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). All experiments reported in this paper followed these inclusion criteria. See Table 1 for descriptions of participant demographics in all three experiments.

2.1.2. Procedure

We modified a standard illusion of explanatory depth task to test the specificity principle. We will first explain the standard task and then highlight our modifications.

The standard task comprises three phases. In Understanding Rating 1, participants assess their understanding of how various mechanical devices[1] work on a scale from 1 (*very little understanding*) to 7 (*very thorough understanding*). Prior to rating, participants receive instructions on using the scale.[2] Subsequently, participants rate their understanding of each provided device. After rating, participants proceed to 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 enter Understanding Rating 2, where they re-rate their understanding of the explained item(s) only. For instance, if a participant initially rated their understanding of how a zipper, a can opener, and a speedometer work, but was only prompted to explain the can opener, Understanding Rating 2 would focus solely on their understanding of the can opener. After re-rating understanding for the relevant item, the task concludes. Our expanded task followed the same three-phase process, with minor modifications to Understanding Rating 1 and Understanding Rating 2. In Understanding

[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, p. 7), for an example of a scale orientation using a crossbow.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

Table 1. *Participant demographics in each experiment.*

Demographic variable	Experiment 1	Experiment 2	Experiment 3
N	240	327	264
Age_{Mean}(*SD*)	40.07 *(13.36)*	39.18 *(13.16)*	38.50 *(11.11)*
Gender (%)			
Male	59.5	50.3	53.4
Female	39.3	48.8	45.5
Other	0.1	0.3	0.4
Education (%)			
High school diploma	22.7	20.9	25.9
College diploma	11.6	10.0	8.6
Bachelor's degree	45.0	45.5	44.0
Master's	16.9	19.1	16.2
Professional degree	1.7	3.0	2.6
PhD	1.2	0.9	1.9
Ethnicity (%)			
American Indian or Alaska Native	0.0	0.6	0.0
Asian	14.5	10.9	6
Black or African American	10.7	12.1	14.7
Hispanic or Latino/a	4.5	4.8	3.0
Native Hawaiian or other pacific islander	0.0	0.3	0.0
White	68.2	70.3	69.5
Other	1.2	0.3	0.4

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 nonresponses and a small amount of rounding error.

Rating 1, we 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 each experiment and Table 3 for the list of stimuli used in each experiment). The Explanation phase remained unchanged, where participants explain the functioning of one of the devices they had just rated. In Understanding Rating 2, participants re-rated their understanding of all the devices they had previously rated in Understanding Rating 1, rather than only the device they explained. The rating of each device took place on a single experiment page, with the devices presented in a random order. These were the sole distinctions between our expanded task and the standard task.

Following the expanded 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.

2.2. Results

We deviated from[3] the preregistration protocol to present a simpler analysis.[4] After attempting to explain how a device works, participants reported knowing less about the device, $t(239) = -4.62$,

[3] The imbalanced conditions (Explanation and Control) of Experiment 2 and Experiment 3 suggest selective attrition (e.g., Zhou & Fishbach, 2016). A concerned reader might claim that the IOED task led less introspective participants to drop out, and only introspective people show the IOED effect, thus the effects are due to this attrition. However, the effects were smaller in Experiment 3 compared to Experiment 2, despite having more selective attrition.

[4] We preregistered computing mixed-effects models for our analysis. However, following a suggestion from the Editor, we instead report a simplified analysis that is more descriptive than inferential in nature. The results of our modeling are consistent

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

Table 2. *Primary task instructions.*

Task	Instructions
Understanding Rating 1	'We are going to present you with six items. For each item, we want you to rate on a seven-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 and 3)* *'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 and 3)* *'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 and 3 only*)	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 chari 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 and 3. 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 our prior research. Experiment 1 did not contain a control condition.

$p < .001$.[5] However, participants also reported knowing less about the device after attempting to explain how a different device works, $t(239) = -10.53$, $p < .001$. In other words, participants reported more knowledge of a zipper (and other devices) before explanation compared to after. Table 4 presents the means and standard deviations of each Understanding Rating across conditions, including the effect size of the within-subjects comparison of the understanding ratings. The larger effect size for the General

with the results reported here. Interested readers can consult our mixed effect-modeling in the Supplementary Material. This applies to the analyses of each experiment in this manuscript.

[5]We report one-tailed tests throughout the manuscript as our hypotheses are directional.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

Table 3. *Phenomena used in Experiments 1–3.*

Experiment	Domain	Item
Experiments 1 and 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

Note: We obtained all 'mechanical device' and 'natural phenomena' items from Rozenblit and Keil (2002, Experiments 1–6 and Experiment 10, respectively). We obtained 'political/economic system' from Meyers et al. (2020).

Table 4. *Summary statistics of understanding ratings by condition.*

Experiment	Condition	*n*	Rating 1 Mean *(SD)*	Rating 2 Mean *(SD)*	Cohen's *d* of difference
Experiment 1					
	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 2					
	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 3					
	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

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, as 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 scale from 1 (little understanding) to 7 (very thorough understanding).

condition suggests that the reduction in understanding of, for example, a zipper (device) was somewhat greater when something other than a zipper was explained, which is an interpretation consistent with Figure 1. We caution against such an interpretation. This is because understanding ratings for some items appear to decrease more when the same device is explained, and others decrease more when a different device is explained. This can be observed in Figure 2, which displays the average difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

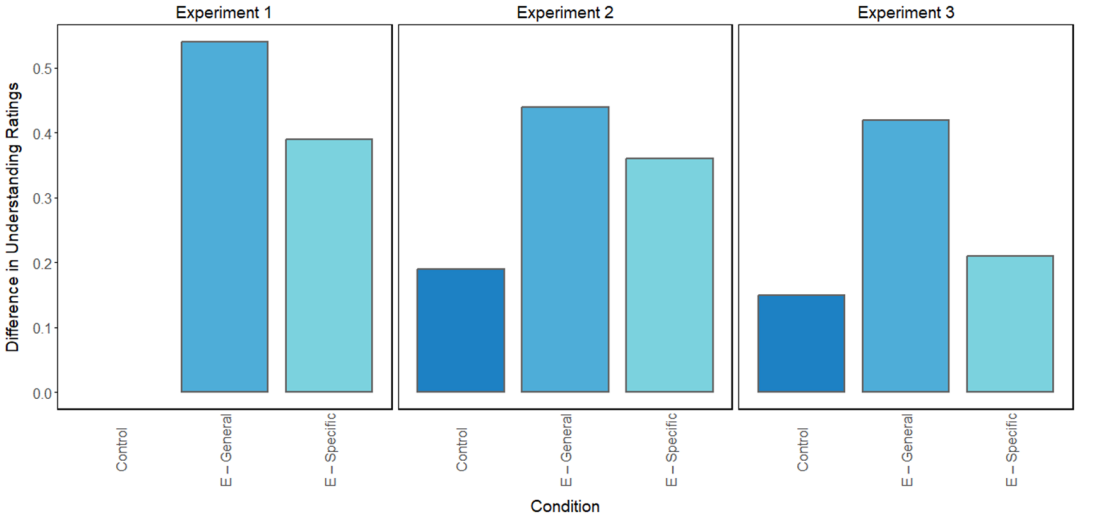

Figure 1. *Mean difference in understanding ratings aggregated across phenomena for each condition (Control = dark, green-colored bars; E – General, which stands for 'Explanation – General' = teal-colored bars; E – Specific, which stands for 'Explanation – Specific' = light, blue-colored bars) across each experiment. This figure displays the differences in understanding ratings between two time points (Understanding Rating 1 – Understanding Rating 2) for each condition, with each bar representing the mean difference. A positive value indicates a reduction in understanding ratings over time. There was no control condition in Experiment 1.*

Table 5 displays the mean and standard deviations of the similarity ratings of each phenomenon. We did not find an effect of similarity on the reduction in understanding across explanation types (see Supplementary Material). This might be due to the devices being rated as generally similar. However, 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), $F(14, 3346) = 39.46$, $p < .001$.

2.3. 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. We found no evidence that the similarity between devices predicted the reduction of reported knowledge.

Despite item variations in similarity, we found no relationship between device pair similarity and understanding rating differences. This unexpected finding challenges the assumption that explaining a device's operation exposes knowledge gaps. For example, while a zipper and a sewing machine share properties likely revealed during explanation, the same is not true for a zipper and a piano key. Our similarity judgments focused on general resemblance rather than functional similarity. Exploring the latter may yield different outcomes. Given the secondary focus on item similarity and cost considerations, subsequent experiments excluded similarity judgments.

Experiment 1 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

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

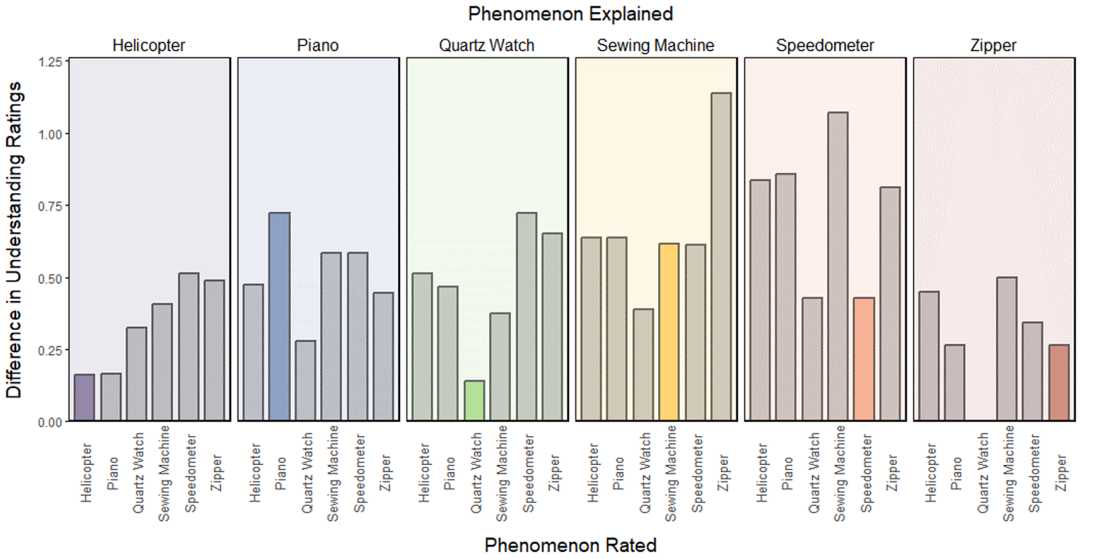

Figure 2. *The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1 minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. The panels (and respective participants) are separated by the randomly assigned Phenomenon Explained. Each gently colored panel represents one of the phenomena which is correspondingly labeled 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 nonmatching phenomena are gray colored). The difference score for the phenomenon quartz watch when zipper was explained was .00 and so is not visible in the plot.*

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

	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.

our observed effects. One such explanation is that providing a second judgment could inherently lead to a reduction in reported understanding. In making a second judgment, a participant may contemplate the device more deeply. This deeper reflection could reveal knowledge gaps, causing a drop in self-perceived understanding due to reconsideration. This draws parallels with 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

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

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; Fernbach et al., 2013; Kleinberg & Marsh, 2020; Zeveney & Marsh, 2016). Because our design is meaningfully different from the designs that previously implemented a control condition, we cannot meaningfully rely on these previous findings. Therefore, we 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 our results could reflect a change in participants' interpretation of the scale between the two ratings. This is especially plausible, as we largely replaced the scale-rating instructions that are normally included, which could have otherwise prevented this potential issue. Therefore, in a preregistered Experiment 2, we attempt to replicate the results of Experiment 1 while ruling out the two aforementioned alternative explanations.

3. Experiment 2

3.1. Methods

Experiment 2 used 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 OSF (https://osf.io/h3trk).

3.1.1. Participants

Following our preregistered plan to obtain 95% power to detect a within-subjects effect of $d = 0.2$, we recruited 327 participants via Mechanical Turk.

3.1.2. Procedure

We made three changes to the procedure of Experiment 1. First, we 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).[6] 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, we 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, we removed the similarity ratings. This produced a 2 (writing condition: control/experimental; between-subjects) × 2 (time: pre-explanation/post-explanation; within-subjects) mixed design.

3.2. Results

Consistent with Experiment 1, we deviated from the preregistration protocol to present a simpler analysis. After attempting to explain how a device works, participants reported knowing less about

[6]The control condition was adopted from Meyers et al. (2020).

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

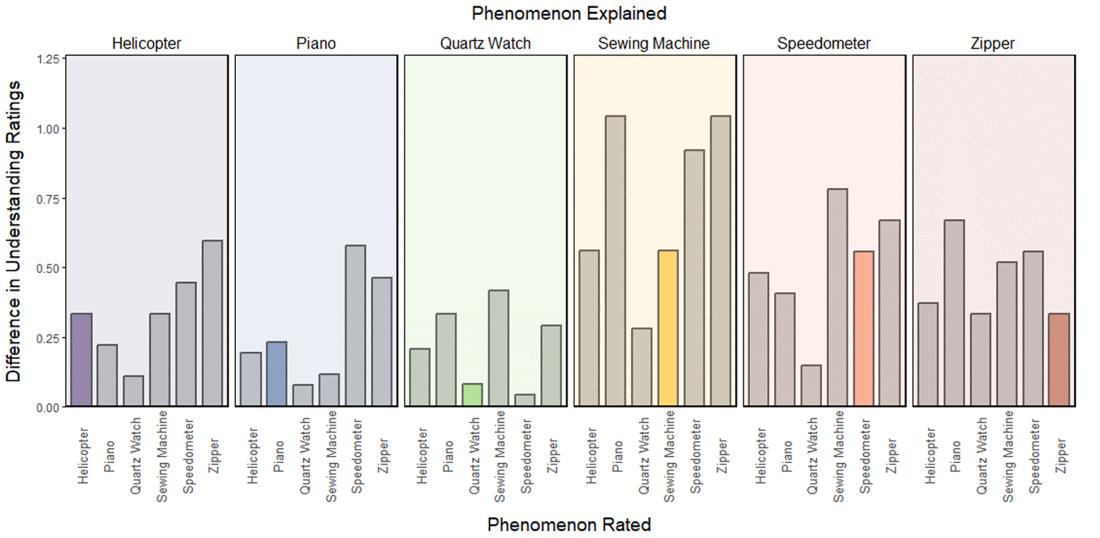

Figure 3. *The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1 minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. The panels (and respective participants) are separated by the randomly assigned Phenomenon Explained. Each gently colored panel represents one of the phenomena which is correspondingly labeled 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 nonmatching phenomena are gray colored).*

the device, $t(155) = -3.92$, $p < .001$. However, participants also reported knowing less about the device after attempting to explain how a different device works, $t(155) = -7.08$, $p < .001$. In other words, participants reported more knowledge of a zipper (and other devices) at time 1 (pre-explanation) compared to time 2 (post-explanation). This is observed in Figure 3. 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 understanding was significantly greater both when the specific phenomenon was explained, $t(198.18) = 1.69$, $p = .046$,[7] and when a different phenomenon was explained, $t(240.95) = 3.53$, $p < .001$, compared to when nothing was explained. The reduction in reported knowledge in the control condition was half the size of the reduction in the explanation conditions. The results are also consistent across phenomena. Compared to when no explanation was required, understanding of a phenomenon dropped more both when the specific phenomenon was explained, $t(5) = -2.40$, $p = .031$, and when a different phenomenon was explained, $t(5) = -7.07$, $p < .001$.

3.3. 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

[7]Between-subjects *t*-tests in this manuscript are Welch's *t*-test.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

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. Participants who copied a block of irrelevant text instead of explaining a device also 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. We examined whether the reduction in understanding was due to participants reinterpreting the understanding scale after explanation. Despite adding further detail to the instructions to minimize this possibility, participants still reported lower understanding of the devices after attempting explanation. We further assessed if the universal drop in understanding was just an effect of rendering a second judgment. While double judgment led to decreased reported understanding, it didn't wholly account for the observed effect.

So far, our findings hint that the subject of explanation might be insignificant in revealing an illusion of explanatory depth, based on our studies with mechanical devices. We hypothesized that explaining unrelated phenomena, such as snow formation, could prompt someone to reassess their knowledge about something distinct, like a zipper's mechanism. Alternatively, this illusion could be domain-specific, suggesting that its revelation occurs at a domain level. This implies that trying to explain any mechanical device (e.g., a sewing machine's function) might expose an illusion about another mechanical device (e.g., a zipper), but not for unrelated phenomena (e.g., snow formation). Likewise, attempting to clarify one natural phenomenon (e.g., earthquakes) might dispel illusions about other natural phenomena (e.g., snow formation), but not about different domains (e.g., zippers). We explored these conflicting predictions in our preregistered Experiment 3.

4. Experiment 3

4.1. Methods

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

4.1.1. Participants

Following our preregistered plan to obtain approximately 90% power to detect a within-subjects effect of $f = 0.1$ (approximately equivalent to $d = 0.2$), we recruited 264 participants via Mechanical Turk.

4.1.2. Procedure

We made a single modification to the procedure of Experiment 2. We replaced four of the six device items with four items from two different domains of knowledge (see Table 3). The domains and items were natural phenomena (items: 'How snow forms', 'How an earthquake occurs') 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 et al. (2020).

4.2. Results

Consistent with the prior experiments, we deviated from the preregistration protocol to present a simpler analysis. Participants claimed lesser knowledge about a phenomenon post-explanation, $t(118) = -1.76$, $p = .040$, and even after they explained a different phenomenon, $t(118) = -6.11$, $p < .001$. Generally, explaining a different phenomenon led to lesser self-reported understanding (refer to Table 4), yet Figure 4 shows this effect is not consistent. Interestingly, as in Experiment 2, even non-explainers reported diminished understanding across the two ratings, $t(144) = -3.62$, $p < .001$, but this drop was steeper among those who attempted an explanation, $t(195.43) = -2.85$, $p = .002$. The results are also consistent across phenomena. Understanding of a phenomenon dropped more when a phenomenon was explained compared to when no explanation was required, $t(5) = -6.26$, $p < .001$.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

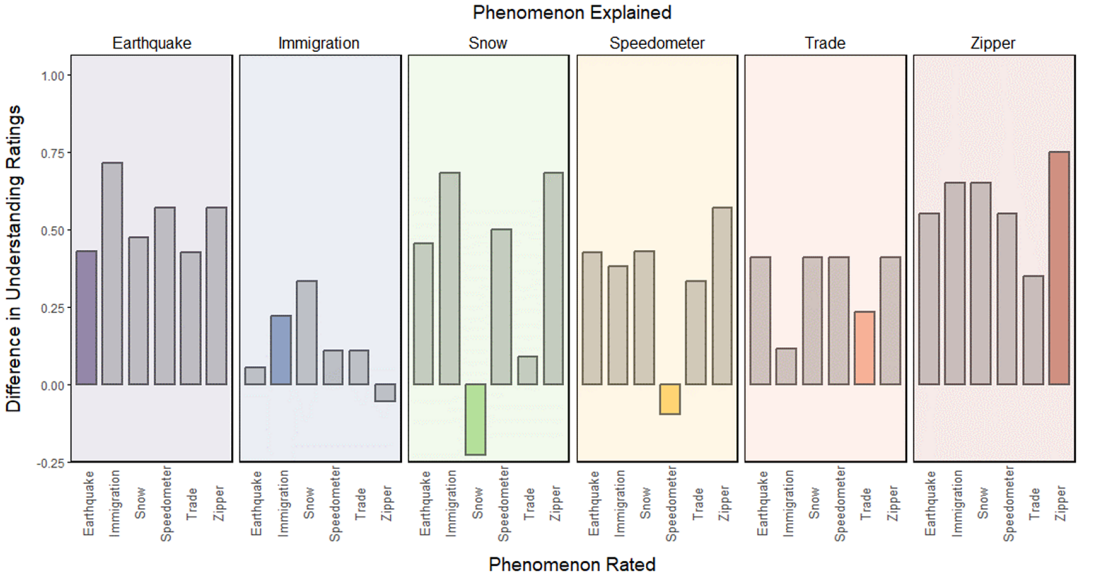

Figure 4. *The difference in understanding ratings for each Phenomenon Rated by Phenomenon Explained. The difference in understanding ratings reflects Understanding Rating 1 minus Understanding Rating 2. A larger value represents a greater reduction in reported understanding of the phenomenon. A negative 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 labeled 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 nonmatching phenomena are gray colored).*

4.3. Discussion

We replicated the results of Experiments 1 and 2 using a set of items spanning three knowledge domains. Participants claimed lesser understanding of a phenomenon after trying to explaining that phenomenon—or a different one. For instance, knowledge claims about zippers dropped after explaining either how they work or how unrelated phenomena like speedometers, snow formation, or earthquakes work. Similarly, copying irrelevant text instead of explaining also lowered self-reported understanding, albeit the reduction was larger when explanation was involved.

5. General discussion

After attempting to explain how a zipper works, participants reported a decreased understanding of a zipper. They also reported a decreased understanding of a zipper after attempting to explain how helicopters fly (Experiments 1 and 2), how speedometers work (Experiments 1–3), how snow forms (Experiment 3), and several other phenomena. They adjusted their understanding for phenomena other than zippers in this manner too, including helicopters, sewing machines, and snow. It was unclear if explaining a specific phenomenon resulted in a larger knowledge claim reduction than explaining a different one. These results are in line with the findings of Roeder and Nelson (2015). Interestingly, participants also reported a decreased understanding of the zipper when they transcribed a block of descriptive text instead of attempting to explain it. However, the reduction in understanding was even greater when participants attempted to explain any phenomenon, regardless of whether it was

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

the same or different than the zipper. These findings have important implications for understanding the functioning of the illusion of explanatory depth and the role of explanation in exposing it. Moreover, this study suggests that explanation may be an effective non-confrontational debiasing technique.

Our 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—our data do not support it. Hence, we must consider alternative models.

One alternative model suggests that the illusion of explanatory depth 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 zipper 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. We refer to this model as the 'linked illusions model'.

Another model suggests that the illusion 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. This model also assumes that failing to explain just one phenomenon should expose the entire illusion. We refer to this model as the 'general overconfidence' model.

The linked illusions model and general overconfidence 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; Mills & Keil, 2004; Rozenblit & Keil, 2002), 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 the linked illusions model and the general overconfidence model are indistinguishable given how the illusion is usually assessed. Future research could use mechanical devices and procedures as stimuli to investigate this further, as procedures have appeared robust to the illusion (Mills & Keil, 2004; Rozenblit & Keil, 2002, Experiment 8). For example, people do not report a lower understanding of how to make chocolate chip cookies after explaining the procedure. So, people could be asked to explain procedures, such as how to make chocolate chip cookies, to test the boundary conditions of our findings.[8]

Our findings also highlight the need for control conditions in examinations of the illusion of explanatory depth. In the control conditions of Experiments 2 and 3, we found a significant but small reduction in judged understanding. This suggests that some portion of the observed reduction

[8]Roeder and Nelson (2015) tested this question using the procedure of 'boiling an egg' (in addition to the items, 'How a helicopter flies' and 'How an official is elected to the Nigerian House of Representatives') in their Experiment 3. They found that explaining how to boil an egg can produce a similar reduction in reported knowledge of other phenomena.

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

in understanding is not due to explaining anything. Despite being a very 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 (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. 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 initially overestimate their knowledge, then a second guess may be lower than the first. This does not suggest that the second guess is more accurate, but rather that the person's true knowledge of the phenomenon lies around the average of their two guesses.

One concern is that our results may be attributed to demand effects. While the no-explanation comparison condition eliminates several potential demand explanations, it is plausible that the demands 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 phenomenon 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, we 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 remains unclear which demand characteristic could account for the observed variability in the magnitude and direction of understanding changes due to the explanation of a phenomenon. Hence, we do not ascribe our results to demand effects.

Our 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 (Kruger & Dunning, 1999; Moore & Healy, 2008). As such, there is no shortage of attempts to reduce people's overconfidence (Arkes et al., 1987; Croskerry et al., 2013; Ferretti et al., 2016). However, few techniques are widely effective. Explanation as a debiasing method can be indirect and nonconfrontational. This has implications for educational settings, where some students may find that explaining concepts in one subject, and experiencing difficulty, might help them engage with concepts and subjects for which they overestimated their knowledge. 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. However, there are at least three limitations that

https://doi.org/10.1017/jdm.2023.24 Published online by Cambridge University Press

should be considered. One is that the reported effects might only occur when questions of understanding are asked by an experimenter (although 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. A second limitation is that the debiasing effect might not apply to new phenomena introduced only after explanation, not before. In our work, participants reported knowing less about the phenomenon they had previously rated but did not explain. However, had they only rated the phenomenon after explanation, they may not have rated it any differently than they would have before explanation (i.e., like the first rating of the two made in the present studies). Thus, one promising avenue for future research is assessing the value of explanation as a debiasing technique outside of the lab. The third limitation is that our data do not indicate the duration of the debiasing effect. Our experience suggests the effect is more likely to be short-lived than long-lasting. Estimating the effect's duration should be a focus of future research to better understand its usefulness as a debiasing technique. Fortunately, this future research can lead to deeper understanding of this illusion—rather than the mere illusion of understanding.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/jdm.2023.24.

Data availability statement. The datasets generated by the survey research during and analyzed during the current study are available in the Open Science Framework, https://osf.io/6huwq/.

Funding statement. This work was financially supported through grants awarded by the Natural Sciences and Engineering Research Council of Canada to authors E.M., D.K., and J.F.

Competing interest. The authors have no competing interests to declare.

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*, 762. https://doi.org/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/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*(Suppl 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). New York: Ablex Publishing.

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. In *Proceedings of the 49th Annual Hawaii International Conference on System Sciences*, pp. 1547–1555. Koloa, HI: IEEE.

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. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, UK: 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

Kleinberg, S., & Marsh, J. K. (2020). Tell me something I don't know: How perceived knowledge influences the use of information during decision making. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. https://cogsci.mindmodeling.org/2020/papers/0410/index.html

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

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. [Unpublished manuscript] PsyArXiv*. https://doi.org/10.31234/osf.io/4bstf

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

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

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

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

Petty, R. E. & Wegener, D. T. (1993). Flexible Correction Processes in Social Judgment: Correcting for Context-Induced Contrast. *Journal of Experimental Social Psychology*, *29*(2), 137–165. https://doi.org/10.1006/jesp.1993.1007

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://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). Mahwah, NJ: 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/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/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. & 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*. Austin, TX: 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/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* (pp. 1020–1025). Philadelphia, PA: Cognitive Science Society.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Cite this article: Meyers, E. A., Gretton, J. D., Budge, J. R. C., Fugelsang , J. A., and Koehler, D. J. (2023). Broad effects of shallow understanding: Explaining an unrelated phenomenon exposes the illusion of explanatory depth. *Judgment and Decision Making*, e24. https://doi.org/10.1017/jdm.2023.24