


Strength and Persistence of the Illusion of Explanatory Depth

Julianne Wilson

Department of Psychology, Lehigh University

Author Note

Julianne Wilson  <https://orcid.org/0000-0002-9433-7012>

I have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Julianne Wilson, 17

Memorial Drive East, Bethlehem PA 18015. Email: juw521@lehigh.edu

Abstract

People's tendency to overestimate their understanding has been shown to be pervasive, especially in relation to causal knowledge. While this phenomenon, termed the illusion of Explanatory Depth (IOED), can be broken through the act of generating a causal explanation, there are still aspects about the boundaries of the IOED that have yet to be explored. Across two experiments, I investigated the strength and persistence of the IOED. In Experiment 1, I determined whether breaking an IOED has the ability to transfer, breaking the illusions of similar items that were not explained. In Experiment 2, I used a 2-Session design to determine if a broken IOED remains broken over a period of one week. Additionally, I assessed the quality of the explanations given by participants to see if it had an influence on either the strength and/or persistence of the IOED. Results for Experiment 1 showed that a decrease in understanding ratings for explained devices lead to an even greater decrease in ratings for unexplained devices, although explanation quality was not found to impact this transfer of knowledge reassessment. Results of Experiment 2 showed the IOED is at least partially retained over a period of one week, for both devices that were previously explained and for completely new devices, and that the IOED can be broken a second time to a greater degree than the first IOED session. These studies provide a better understanding of the IOED and its limits, which is imperative for successfully combating this metacognitive error.

Keywords: Illusion of explanatory depth, causal relationships, causal knowledge, explanation

Strength and Persistence of the Illusion of Explanatory Depth

Throughout their lives, humans are constantly learning how different items and events relate to one another through cause and effect. This causal knowledge serves as a guide for making inferences (Matute et al., 2015) and categorizing the world around us (Keil, 2003). In addition, causal reasoning is often central in making judgements (Bes et al., 2012) and real-world decisions about important aspects of daily life including politics (Alter et al., 2010), finances, and healthcare (Zheng et al., 2020). Although this frequent use of causal knowledge may lead people to believe they have a vast understanding of the world, research has shown that people's causal understanding is often shallow and rife with omissions (Matute et al., 2015; Wilson & Keil, 1998). Worse still, people tend to be unaware of their own ignorance (Rozenblit & Keil, 2002). According to Rozenblit and Keil (2002), this lack of complete understanding and poor introspection leads to a phenomenon known as the illusion of explanatory depth (IOED).

The IOED may seem relatively harmless for examples such as whether a person understands the causal mechanism of how a can opener works. However, misunderstanding one's knowledge can be particularly detrimental when making important decisions such as whether to receive medical treatment or which politician to vote for. The IOED can mistakenly lead an individual to believe they already contain all the necessary knowledge to confidently make well-educated choices (Alter et al., 2010) and may even cause them to decline expert knowledge or advice (Scharrer et al., 2014). Therefore, when people believe they understand something to a higher extent than they actually do, they may make decisions that are not in the best interest of themselves or others.

One way to break this illusion of understanding is to have the individual write out an explanation. Rozenblit and Keil (2002) developed a paradigm using this strategy, which is often

used as both a test of the illusion's existence and a technique for breaking it. In this manuscript, I explore how exposing one's lack of understanding influences their knowledge reassessment more globally. Specifically, I attempt to inform whether breaking an illusion of causal understanding for a specific concept can cause a more generalized reassessment of knowledge and whether this knowledge reassessment persists over time, both using the IOED paradigm. In the following, I first describe the IOED paradigm in detail, explain the areas of knowledge for which it has been tested, and discuss possible explanations for the existence of the IOED. Next, I connect the IOED to aspects of the metacognitive literature that suggest how illusions of understanding may differ under different conditions. I then describe recent research that investigated the generalizability of knowledge reassessment using methods similar to my first experiment. Finally, I discuss two experiments with which I explored the transferability of explanation-induced knowledge reassessment across concepts and over time.

The IOED Paradigm

The IOED paradigm, first employed by Rozenblit and Keil (2002), begins with the presentation of detailed instructions and training on how to assess one's own understanding of a causal mechanism using a 1 to 7 scale. The instructions include an example explanation of how a crossbow works with pictures that represent what a low (1), middle-of-the-road (4), and high (7) understanding explanation would look like. Participants are then asked to use the scale they just learned to rate their understanding of a list of items (see Table 1A for example items). These ratings are considered the pre-explanation or Time 1 (T1) understanding ratings.

Next, participants are asked to write out a step-by-step causal explanation of how one of the items from the original list operates, after which they immediately re-rate their understanding of that item using the same understanding scale. These ratings are considered the post-

explanation or Time 2 (T2) understanding ratings. The cycle of generating an explanation and then re-rating the explained item continues until all items selected by the experimenter are complete. The results of the IOED task are typically determined by comparing the average of the T1 ratings to the average of the T2 ratings across items. A decrease in understanding rating from T1 to T2 shows that the IOED has been broken, and the subject is now aware of their lack of understanding of the items they explained.

Rozenblit and Keil (2002) tested the prevalence of the IOED in different types of knowledge to determine when the phenomenon was more likely to occur. They found that people showed little to no decrease in understanding ratings for items such as facts, procedures, and narratives, but consistently overestimated their knowledge of causal mechanisms. Illusions of understanding for the causal mechanisms of devices and natural phenomena were successfully able to be broken through the IOED paradigm. These findings led Rozenblit and Keil (2002) to conclude that (1) it is not simple overconfidence that leads to the IOED, and (2) it is the act of writing a causal explanation that successfully breaks the IOED. Other work has shown that deeply reflecting on a causal explanation can reduce understanding ratings, but to a smaller degree than the act of writing out the explanation (Johnson et al., 2016). Further research on the IOED has confirmed its existence in devices (Lawson, 2006; Mills & Keil, 2004), and additionally found people overestimate their causal understanding of political policies (Alter et al., 2010; Fernbach et al., 2013), historical and economic issues (Gaviria & Corredor, 2021), and mental health disorders (Zeveney & Marsh, 2016).

What Causes the IOED?

A popular current explanation for the IOED is the community-of-knowledge hypothesis (Sloman & Rabb, 2016). Under this hypothesis, individuals have an inflated sense of knowledge

due to being unable to distinguish their own knowledge from the knowledge of others (Fernbach & Light, 2020; Sloman et al., 2021). For example, almost everyone knows that humans require oxygen to live; however, only a small fraction of those people can actually explain why. Although both groups are considered to “know” this information (Sloman et al., 2021), the depth of their understanding is widely varied. Humans’ ability to outsource information in this way increases their feeling of understanding, even when they have no true knowledge on the topic (Fernbach & Light, 2020; Rabb et al., 2019; Sloman et al., 2021).

It has also been shown that people often think they know more about something if they are told that experts have knowledge about it or if they think they can readily find the information on the internet (Fisher et al., 2015; Rabb et al., 2019). For example, people were found to significantly increase their understanding ratings when they were presented with made-up phenomena that were said to be understood by scientists and easily accessible to them (Sloman & Rabb, 2016). This shows that simply believing others hold relevant knowledge can cause people to feel they know more about that subject.

An alternative explanation for the IOED is regression to the mean. In the general statistical literature, regression to the mean refers to the phenomenon that more extreme ratings (i.e., ratings toward either end of a scale) will trend toward the group mean with repeated testing (Mazor & Fleming, 2021). Applied to the IOED, this hypothesis contends that participants who rate their initial understanding closer to either end of the 1 to 7 scale do so not because they suffer from self-misperception, but because of a simple estimation error. Therefore, if those participants were to be tested again, their subsequent responses would show scores closer to that of the group average (Krueger & Muller, 2002). However, because the IOED paradigm has never been repeated within the same experiment, this hypothesis has not been tested directly.

The Role of Metacognition in the IOED

The IOED is a specific example of a metacognitive error. Metacognition, defined as knowledge about one's own knowledge (Flavell, 1979), and metacognitive monitoring, or assessing one's own knowledge (Rhodes, 2019), are both essential to the IOED. A person must have awareness of their current knowledge and be able to assess the adequacy of what they know to make a metacognitive judgement, or a judgement of their performance on a particular task (Rhodes, 2019), such as rating their understanding of a phenomenon.

The IOED paradigm uses metacognitive monitoring and introspective feedback to break the illusion of knowledge. During the IOED paradigm, the participant is asked to give both prospective (T1) and retrospective (T2) metacognitive judgements. Prospective and retrospective metacognitive judgments are thought to be based on both declarative knowledge and subjective experience (Siedlecka et al., 2016). New information acquired during post-decisional processing may cause retrospective judgements to differ from prospective judgements (Koriat & Levy-Sadot, 2000). In the case of the typical IOED paradigm, this means that T2 ratings may be affected by information gathered in between T1 and T2 ratings, which is when participants are asked to generate a causal explanation. Because the generation of a causal explanation is the only task that occurs during this period of the paradigm, any information obtained that can affect T2 ratings would have to be in the form of introspective feedback. This suggests that it is this introspective feedback that causes the participant to reassess their knowledge and alter their retrospective metacognitive judgement (T2; Schwarz, 2004).

Metacognitive Evidence for Knowledge Transfer

There is evidence from the metacognitive literature that an IOED task may have the potential to cause knowledge reassessment more broadly. In metacognition experiments,

participants who received feedback on their metacognitive accuracy, or how much their understanding ratings agreed with their objective accuracy, were able to improve their metacognitive accuracy on both similar and divergent tasks more than participants who received feedback on only their objective accuracy (Carpenter et al., 2019). This shows that the addition of metacognitive feedback, much like the introspective feedback that occurs during the IOED paradigm, allowed participants to make more generally accurate metacognitive judgements in the future.

People have also been found to use a domain-general mechanism for making metacognitive judgements in certain situations to improve metacognitive accuracy (Schraw, 1996). For example, individuals may use previously learned information about their performance on a math test to assess how well they feel they will perform on an English test. In addition, within an educational setting, the incorporation of metacognitive instruction was found to increase students' ability to transfer knowledge to different contexts (Georghiades et al., 2000).

There is also evidence for the potential retention of metacognitive judgements over time. Participants asked to make metacognitive judgements while responding to test questions made similar metacognitive judgments one week later about the previously answered questions (Barenberg & Dutke, 2019). Additionally, the regular practice of metacognitive activities in general was found to improve the amount of information retained and the ability to transfer knowledge to new contexts over time (Georghiades et al., 2006).

Previous Work on Knowledge Transfer in the IOED

One doctoral dissertation attempted to investigate domain transfer in the IOED (Roeder, 2016). In his first experiment, change in understanding ratings was compared between one group asked to rate the same items they explained (the typical IOED procedure) and a different group

asked to rate items they did not explain during the paradigm. Results showed a significant decrease from T1 to T2 for the group that rated unexplained items, although it was significantly smaller than the decrease in ratings for group that rated the items they explained. Unfortunately, the results of the other experiments did not provide good corroboration for this finding due to the stimuli either being inappropriate or indeterminate. For example, one of the items used in the IOED paradigm in a number of Roeder's (2016) follow-up experiments was actually a procedure, which the original set of IOED experiments by Rozenblit and Keil (2002) had already determined to be unaffected by explanation. In addition, the results section for the last experiment discusses the difference in understanding ratings for a device that was not mentioned in the methods section for that experiment, making it difficult for the reader to determine which items were actually rated for this experiment. Given that this is not a peer reviewed document, the findings remain unclear.

Similar to Roeder (2016), my first experiment explored the idea of domain transfer in the IOED, however I also looked at the quality of the explanations and their potential influence on transfer within a particular domain. Previous work in our lab showed that both perceived completeness of an explanation and the number of causal links an explanation contained were predictive of the magnitude of decrease from T1 to T2 (Wilson & Marsh, 2023). In addition, unlike Roeder's (2016) between-subjects design, my first experiment was a within-subject comparison with participants making post-explanation ratings for both explained and unexplained items. This allows for a more direct comparison when assessing both differences in understanding ratings and explanation quality.

After these experiments had been preregistered (<https://osf.io/8h2k5>) and conducted, I was directed to a recently published work by Meyers et al. (2023) where the investigators

explored the transferability of the break in the illusion of causal knowledge both within (as in Experiment 1 of this manuscript) and across domains. While this paper investigated the same question relative to my first experiment, there are significant differences related to methodology and analyses used.

In relation to methodology, Meyers et al. (2023) had participants explain one item and rate a total of six items (five unexplained items), while I had participants explain six items and rate a total of twelve items (six unexplained items). Also, the instructions used in their experiments, as well as their post-explanation ratings for both the explained item and the unexplained items, varied from Rozenblit and Keil (2002). In the original IOED paradigm, participants were prompted to re-rate their understanding of a particular item (and that item only) directly after writing out an explanation for that item. In Meyers et al. (2023), all post-explanation ratings were made at once on the same page, after participants wrote out the single explanation. While this may seem like a minor difference, performing the paradigm in this fashion arguably sets up the participant to compare secondary ratings among explained and unexplained items, as opposed to simply focusing on the one item that they had just explained.

Meyers et al. (2023) also included a control condition in two of their experiments where participants did not have to generate an explanation but simply transcribe one¹. This addition is interesting and important in assessing the overall reason for the decrease in understanding ratings specific to the IOED paradigm. However, it is irrelevant for the question at hand - whether the decrease that is seen through explanation generation in this paradigm can be generalized to items that were not explained. If my question had been related to whether explanation generation is the

¹ Similar to Johnson's (2016) REA group, where participants simply reflected on their ability to provide an explanation, the control group in Meyers et al. (2023) had a significant decrease in understanding ratings from T1 to T2, but this decrease was significantly less than the other groups.

only thing leading to a decrease in understanding ratings, then this type of control condition would definitely serve to enhance the experimental design. In the case of my experiments, the explained items (representative of the typical IOED paradigm) serve as the control condition.

Second, the statistical methods performed in Meyers et al. (2023) differed from the standard analyses used for the IOED paradigm as well as the analyses used in the following experiments. Meyers et al. (2023) used *t*-tests to compare the average initial ratings to the average post-explanation ratings for both explained and unexplained items separately. From this, they were able to determine that (1) post-explanation ratings were significantly lower than pre-explanation ratings for explained items, and (2) post-explanation ratings were significantly lower than pre-explanation ratings for unexplained items. While these results support the idea of knowledge transfer, they do not consider the variability in ratings among explained and unexplained items, as discussed directly by the authors (Meyers et al., 2003, Figures 2 & 3) and in Wilson and Marsh (2023). In addition, these analyses do not directly compare ratings for explained and unexplained items, with the authors cautioning against any such interpretation of their results (Meyers et al., 2023, p. 7). This also leads one to caution comparing their calculated effect sizes between groups. Meyers et al. (2023) did use a linear mixed models (LMM) analysis in their supplementary materials, which accounts for these concerns, but they do not make any conclusions based on these results.

Overall, the work presented in Meyers et al. (2023) is a sound preliminary investigation into the generalization of knowledge transfer within the IOED paradigm that my first experiment serves to replicate using alternative methods and to extend with more advanced statistical analyses.

Overview of Experiments and Hypotheses

Current research on the IOED suggests that explanations can expose an individual's misperception of their causal knowledge. Previous work in metacognition found that engaging in metacognitive judgements can lead to a broader update of metacognitive beliefs that hold over time. A less-explored question is whether explanations hold this same power. That is, are the reassessments people make to their causal knowledge when asked to explain a phenomenon able to be carried over to other phenomena and over time?

The goal of the research presented in this manuscript is to investigate the strength and persistence of the knowledge reassessment that occurs during an IOED paradigm. First, I explored whether the knowledge reassessment of explained items is generalizable to similar but unexplained items. Second, I investigated whether this change in causal knowledge is able to be retained over an extended period of time. Lastly, I looked at whether explanation quality plays a role in the degree of strength and/or persistence observed in both experiments.

Given the previous literature on metacognition and that of Johnson et al. (2016) and Roeder (2016), I hypothesized that knowledge reassessment would transfer to unexplained items, although not to the same degree as explained items. To test this, I examined whether the decrease in understanding ratings that occurs for explained items during the IOED paradigm also occurs for unexplained items within the same domain, during the same task. I predicted that the knowledge reassessment for the unexplained items would be similar to the results of Johnson et al.'s (2016, Experiment 5) cognitive reflection group and Experiment 1 of Roeder (2016), showing a significant decrease in understanding ratings for unexplained items, but one that is still significantly smaller than the decrease in ratings for explained items. Therefore, while I expected the knowledge reassessment to be generalizable, I did not expect the strength of the carryover to reach the same magnitude as what is seen for explained items.

In relation to time, the metacognition literature suggests that introspective reflection induces a deeper understanding of an experience (Georghiades et al., 2000), making it more likely to be remembered over time. Since the generation of a causal explanation causes an introspective reassessment of understanding, I hypothesized that the experience, and in turn the reassessment of knowledge, would be retained over time. To test this, I had participants complete the IOED paradigm on two separate lab visits, one week apart, and assessed whether changes in understanding ratings that occurred during the first visit carried over to the second. I predicted that the T2 ratings of items explained during the initial IOED session would be similar to the T1 ratings of the second session for the same items, suggesting that the change in knowledge assessment was retained during the time between paradigms.

Lastly, previous work in our lab has shown that the quality of the explanations given in an IOED paradigm predicts the strength of knowledge reassessment (Wilson & Marsh, 2023). In Wilson and Marsh (2023), perceived completeness, perceived inclusion of important details, and the number of causal connections within an explanation were found to have a significantly positive relationship with the magnitude of change in understanding ratings (i.e., higher levels of completeness, important details, and number of connections predicted less of a decrease in understanding ratings from T1 to T2). Because this was the first work to explore explanation quality, I attempted to replicate the results seen for explained devices in both Experiment 1 and Experiment 2 by measuring participants' perceived completeness of explanations and objectively coding all explanations to determine the number of causal connections in each. I also extended these findings by investigating whether explanation quality may influence the transfer of knowledge reassessment. Based on the results for explained devices in Wilson and Marsh (2023), I predicted that participants who have lower quality explanations in relation to

completeness and causal connections would be more likely to successfully transfer their knowledge reassessment across items and over time.

Experiment 1

To determine whether the knowledge reassessment that occurs after an IOED paradigm carries over to similar but unexplained devices, participants completed a typical IOED paradigm for a set of devices with the addition of giving a secondary rating for the devices they did not explain.

Method

Participants

I used effect sizes from a similar study previously conducted in our lab (Wilson & Marsh, 2023) to determine the number of participants needed for adequate power in this study. Specifically, in a sample of 51 participants who completed the IOED task for 5 devices, I calculated Cohen's d for the paired-samples t -tests² used to determine whether the post-explanation understanding ratings were significantly different from the pre-explanation understanding ratings. Ratings for four out of the five devices significantly decreased, with effect sizes ranging from medium-small to medium-large in size (Can opener $d = 0.37$, Piano keys $d = 0.65$, Zipper $d = 0.67$, Car ignition system $d = 0.83$). To remain conservative, I used the smallest effect size of the devices found to have a significant drop ($d = 0.37$) to calculate the sample size needed to achieve 0.80 power using the G*Power software. The power analyses found a need for 60 participants. However, because I would be asking participants to rate a wider range of items in this study, some of which could potentially have smaller effects, I chose to oversample and recruit 74 participants.

² Results of paired-sample t -tests were used instead of LMM (which is the actual analyses that will be used in these studies) because, at present, there is not a consensus on how to determine sample size from LMM results.

Participants were recruited from the introductory psychology classes at Lehigh University. Students received credit toward the course. Participants were included who had normal or corrected-to-normal vision. Of the 74 participants (age $M = 18.97$, range 18 – 21), most identified as men (86%; women = 5%; nonbinary = 4%; preferred to self-describe = 1%; preferred not to respond = 3%), as White (63%; African American = 6%; Asian = 15%; preferred to self-describe = 9%; preferred not to respond = 8%), and not Hispanic (82%; Hispanic = 14%; preferred not to respond = 4%). Participants who did not provide ratings for all items were excluded from analysis ($n = 3$). In addition, two participants were excluded from analysis due to their responses to screening questions – one participant explicitly expressed their lack of fluency in English while the other noted previous knowledge about the nature of the experiment.

Materials

I used twelve devices as stimuli. Because many of the devices on the list used by Rozenblit and Keil (2002) for their original IOED experiment are outdated (i.e., “How a VCR works”), I chose or adapted eight devices that the undergraduate sample were likely to be familiar with from this list (can opener, piano keys, flush toilet, zipper, spray bottle, ballpoint pen, water faucet, and “cylinder lock” was changed to just “lock” for clarity). Four other common, household devices (toaster, freezer, printer, and electric blanket) were chosen to complete the list. All devices can be found in Appendix A, Table 1A along with the wording used for each device.

All participants were given the same set of items. I counterbalanced which devices were explained across two groups. The exact materials explained by group 1 ($n = 33$) can be seen in Appendix A, Table 1A. The second counterbalanced group ($n = 36$) had the explained and

unexplained devices reversed. Initial instructions given to participants were based on those used in Rozenblit and Keil (2002; Experiments 1-4, Phase 1) and can be found in Appendix B.

Measures

IOED Paradigm. Participants completed the same T1 and T2 measures described in Rozenblit and Keil (2002), specifically Experiments 1-4, Phases 1-3. The T1 prompt stated “For each of the following, please rate your understanding using the 1 to 7 scale that you just learned about.” For T2 ratings of explained items, participants were asked “Now please rate how well you feel you understand X” (where “X” was one of the device phrases). For T2 ratings of unexplained items, participants were first reoriented with the statement, “Now, you are going to rate some more items that you rated before,” and then presented with the prompt, “Please rate how well you feel you understand X.” Responses for both time ratings were made on a numerical scale from 1 (Very vague understanding) to 7 (Very thorough understanding).

Participants were given the following prompt before writing each explanation: “Now, we’d like to probe your knowledge in a little more detail on some of the items. As best you can, please describe all the details you know about X, going from the first step to the last, and providing the causal connection between the steps. That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish. In other words, try to tell as complete a story as you can, with no gaps. Please take your time, as we expect your best explanation.”

Explanation Quality. To assess participants’ perceived completeness of their explanations, participants were asked to rate how much (completeness percentage question) and what (components questions) they included in their explanations. To measure how complete of an explanation they felt they generated overall, participants were given the prompt “Think about

everything a person could have produced in generating an explanation of X.” (where “X” was one of the device phrases), and then asked “What percent of the possible information do you think you produced?” Responses were made on a 0% to 100% sliding scale. To measure the level of detail participants felt they produced, they were asked to “Think about what you *did* produce in your explanation of X” and then to rate how much they agree or disagree with the following statements based on the completeness of their explanation for that device: “I included all of the big, important parts that would need to be in an explanation of X” and “I included all of the small, less important details that could be in an explanation of X.”³ Responses to these questions were made on a 1 (Strongly Disagree) to 7 (Strongly Agree) scale.

Lookup Questions. Participants were asked if they had previously looked up the mechanism for all devices rated. The question was phrased in the following manner to encompass a range of definitions for the phrase “looking up” and to elicit as honest a response as possible: “Sometimes people look up how things work because they have to fix something or because they’re interested in how something works. ‘Looking up’ could include watching a YouTube video about how a device works, reading a website, talking to a family member or friend, talking to an expert, or any other place where you could get information about how a device works. Which of the following best fits the description of how often you have looked up information about the following items? (Please be as honest as possible. Your response will not affect your credit for this study in any way.)” Participants were then given a list of all devices mentioned during the study and asked to select from the following responses: Never, Once or twice before, Within the last month, Within the last week.

³ Although the question about small details was not found to have a significant effect on rating decrease in (Wilson & Marsh, 2023), it is included in the measure for context and consistency.

Qualitative Questions about Understanding Ratings. To gain insight into participants' perceived reasoning for making their ratings, I included two additional questions. The first question asked participants to explain their reasoning for ratings made during T1 versus T2 more generally. The second question asked participants to explain their reasoning for how they chose the T2 rating for a device they explained versus a device they did not explain. Full text for both of these questions can be found in Appendix C under "Experiment 1". These data were not analyzed for this particular investigation.

Demographic and Screening Questions. Participants were asked general demographic questions, along with two screening questions to test their understanding of the experiment. These questions were: "What was the current study about?" and "Please describe what you did during this study."

Procedure

Participants performed the experiment in-person on a lab computer. After electronic informed consent, participants were shown the instructions for rating their understanding during the IOED paradigm (see Appendix B). They were then asked to use the newly-learned scale to rate all devices (T1 ratings). Next, participants were asked to explain a device, then re-rate their understanding of that device (T2 rating for explained item). They then repeated this process for five more devices. After all explanations and subsequent ratings were complete, participants were asked to re-rate their understanding of the rest of the devices they had previously rated but did not explain, using the same one-by-one format in which they re-rated the explained items (T2 ratings for unexplained items). The device order for T1 ratings, explanations, and T2 unexplained ratings was randomized for each participant.

After all T2 ratings were completed, participants were asked questions pertaining to the completeness of their explanations. The completeness percentage question was always presented before the components question. Participants completed these questions as a set – for one device at a time – although the device being asked about was randomized. Participants were only asked about the devices for which they wrote explanations. Finally, participants completed the look up question, the qualitative questions about understanding ratings, the demographic measures, and the screening questions. Before leaving, participants were debriefed on the purpose of the study and offered a copy of the consent form.

Results

Device-specific data for participants who claimed they had looked up that device “Within the last week” on the look up question was removed from analysis.⁴ This occurred on one device each for three participants, amounting to a total exclusion of six device ratings over the 1656 total device ratings (69 participants x 12 T1 ratings + 69 participants x 12 T2 ratings).

IOED Analyses

Analyses of knowledge reassessment using the IOED paradigm have been traditionally approached by averaging ratings across stimuli and comparing the mean T1 and T2 ratings. Recent work has highlighted the large variability among ratings between different devices and how using this method of analysis can provide a less-than-accurate interpretation of the data (Wilson & Marsh, 2023). Because of this, I employed a LMM approach, including participant and device as random intercepts and time as a repeated measure. Including the random intercepts accounts for both variability among participants (i.e., who are more conservative as opposed to those who strongly overestimate their ratings) as well as variability among the different devices

⁴ The significance of results for analyses specific to the IOED and carryover to unexplained items were not altered when this data was included.

being asked about (i.e., some devices may be less understood than others). For all following analyses, I compared the AIC for the LMMs with Compound Symmetry (CS) and Unstructured (UN) covariance structures. There was minimal or no improvement in the AIC for UN, so CS was used throughout. Variance Components (VC) was used for estimation of random intercepts. I focused on the ANOVA (*F*-style) output to determine whether there is a significant main effect(s) and/or interaction effects. Significant interactions were followed up with Sidak-corrected comparisons.

IOED for Explained Devices. The main goal of this study was to see if the knowledge reassessment that occurs during the IOED paradigm for explained devices is able to be transferred to similar but unexplained devices. Therefore, I first confirmed that the participants' illusion of understanding was successfully broken for the explained devices (the typical IOED). A LMM approach, as described above, was employed using Time (T1 vs. T2) as the independent variable and understanding ratings as the dependent variable. Data for this analysis was restricted to explained items only. There was a significant main effect of Time, $F(1,412) = 45.63, p < .001$, such that understanding ratings at T2 were found to be significantly less than at T1 for explained devices (Figure 1 left). These results are consistent with the IOED literature in that it shows that participants' illusions of understanding were successfully broken for explained items.

Carryover Effects to Unexplained Items. To determine whether the break in illusion seen for explained devices carried over to unexplained devices, I used a LMM analysis (as described above) but with Time (T1 vs. T2) and Item Type (explained vs. not explained) as factors and understanding rating for all items as the dependent variable.

There was, again, a main effect of Time, $F(1,823) = 200.3, p < .001$, where T2 ratings ($M = 3.40, SE = 0.23$) were significantly lower than T1 ratings ($M = 4.12, SE = 0.23$). A main effect

of Item Type was also found, $F(1,744.2) = 5.34, p = .021$, where average understanding ratings for explained devices ($M = 3.85, SE = 0.23$) were significantly higher than average understanding ratings for unexplained devices ($M = 3.67, SE = 0.23$). Lastly, a significant interaction was seen between Time and Item Type, $F(1,823) = 14.56, p < .001$. Follow-up tests showed that understanding ratings were significantly lower at T2, when compared to T1, for both explained and unexplained items ($ps < .001$). Importantly, when comparing across item types, understanding ratings did not differ for explained and unexplained items at T1 ($p = .913$) but were significantly lower at T2 for unexplained items ($p < .001$; Figure 1).

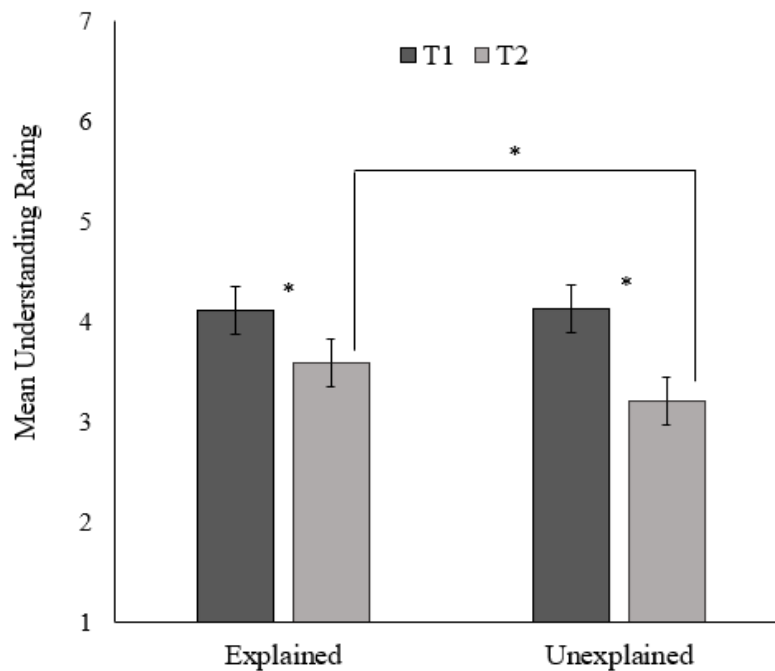


Figure 1: Experiment 1 results for explained (left) and unexplained (right) items Error bars indicate standard error. * $p < .001$

Explanation Quality Analyses

Explanation Coding. To examine the objective quality of participants' explanations, all explanations were coded by research assistants to determine the number of causal links present. A causal link was defined by the presence or inference of a part acting on another part. In other words, a complete causal connection would include three things: a part doing the acting, the action, and a part being acted on. Each explanation was coded separately by two different research assistants, who then met to settle disagreements. The few unresolved disagreements were determined by a third party.

Explanation Quality and Explained Items. To examine whether explanation quality influences understanding ratings for explained items, I used a LMM analysis with a focus on the regression output. For this analysis, change score, calculated by subtracting T1 ratings from T2 ratings for explained devices, was the dependent measure. Predictor variables included three perceived measures of explanation completeness (Percent Complete, Big Parts Inclusion, and Small Parts Inclusion – all person-centered), number of Causal Links (determined by explanation coding), and average of T1 and T2 ratings for explained items per participant centered by its grand-mean (to account for global variability in participants' ratings that is lost when calculating the change score). All measures of completeness were added to the same model to reduce the possibility of Type I errors based on correlations and/or multicollinearity. Additionally, person-level means for participants' explanation completeness ratings were added as predictors to the model to account for the person-centering of each variable (Field, 2018). Only data for explained devices was used for this analysis. Finally, to account for the likely variability among participant ratings and ratings per device, participant-level and device-level random intercepts were included in the model. When the initial model was run, including device as a random intercept was

determined not to significantly improve the model (determined by the Wald statistic, $z = 1.11$, $p = .267$), so the model was run again without device as a random intercept.

The resulting LMM can be seen in Table 1 under “Four Predictor Model” for predictors of relevance. Within the model, 2 of the 4 predictors were found to be significant: Percent Complete and Causal Links. The model was re-run including only significant predictors and their relevant averages (if applicable) with the rest of the model kept the same (including the average T1 and T2 ratings). Both predictor variables were found to be significant in the new Final Model⁵ (Table 1).

Figure 2 provides a visualization of how the two significant predictors varied with change in understanding scores based on the Final Model. The LMM results were used to calculate how much change in understanding (change score) was predicted for a participant whose values for each predictor were at one standard deviation below and above each predictor mean. Lower points on the y-axis of the graph show a greater decrease in understanding post-explanation. As shown in Figure 2, both variables positively predict the magnitude of difference in understanding ratings. Specifically, participants with lower Percent Complete ratings and less Causal Links in their generated explanations had a larger decrease in understanding ratings post-explanation.

⁵ Two additional models that included the random slopes for each of the predictor variables were also generated in an attempt to see if they would improve upon the model and, if so, could be used for additional analyses. With the addition of a random slope for Percent Complete, the AIC actually increased (from 1398.1 to 1400.0) and the Wald statistic for the random slope was not significant, $z = 1.182$, $p = .237$. Results were similar with the addition of random slopes for Big Parts Inclusion and Small Part Inclusion. When a random slope was added for Causal Links, the model could not reach convergence. This is likely due to the sample being too small to be able to estimate the appropriate parameters. As such, the addition of random slopes did not significantly improve the model.

Table 1*Experiment 1: Predictive Ability of Explanation Quality for Explained Items*

Four Predictor Model							
Predictor	Estimate	SE	95% CI		df	<i>t</i>	<i>p</i>
			Lower	Upper			
% Complete	0.026	0.006	0.014	0.038	369.5	4.26	< .001
Big Parts Inclusion	0.123	0.070	-0.015	0.260	343.8	1.76	.080
Small Parts Inclusion	0.108	0.070	-0.029	0.245	341.7	1.56	.120
Causal Links	0.073	0.021	0.032	0.113	400.8	3.50	< .001
Final Model							
% Complete	0.036	0.004	0.027	0.044	408.0	8.23	< .001
Causal Links	0.079	0.021	0.039	0.120	406.2	3.81	< .001

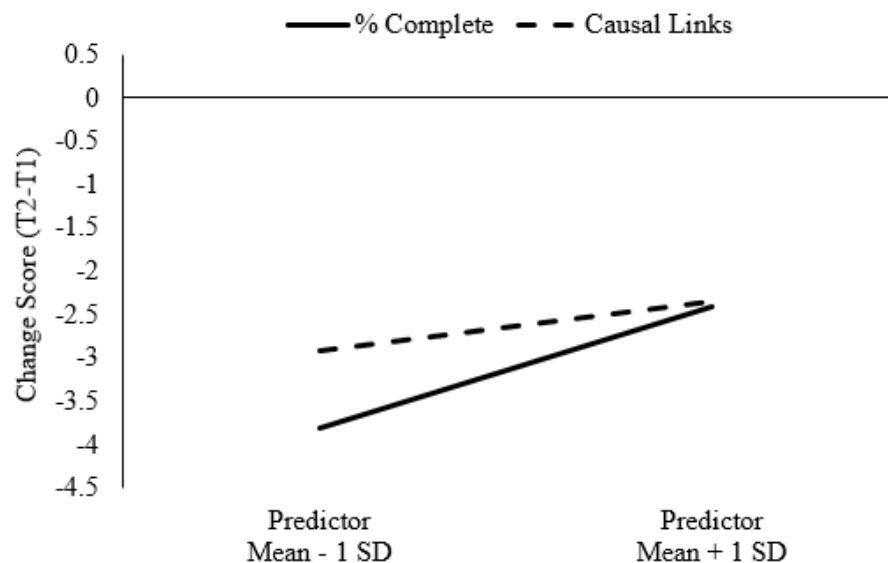


Figure 2: The effect of significant predictors on the difference in understanding ratings from T1 to T2 for explained devices in Experiment 1.

Explanation Quality and Carryover Effects to Unexplained Items. To examine whether the quality of the explanations for explained devices affects the transfer of knowledge reassessment to unexplained items, I used a LMM similar to the analysis above. An important difference is that, for this analysis, change score of unexplained items, calculated by subtracting

T1 ratings from T2 ratings for unexplained devices, was the dependent measure. Predictor variables were calculated by taking the average for each of the completeness items for each participant for all explained items (variables were then centered by their grand-means before being added to the model). In addition, the average of T1 and T2 ratings for unexplained items per participant centered by its grand-mean was also added to the model (to account for global variability in participants' ratings that is lost when calculating the change score). All data pertaining to explanation quality was taken from explained devices while all data pertaining to understanding ratings was taken from unexplained devices for this analysis. Because the model was unable to converge with random intercepts for both participants and devices, only a random intercept for participants was added to the final model. The model was also unable to converge with the addition of random slopes for any of the predictors. The final LMM can be seen in Table 2 for predictors of relevance. Within the model, none of the predictors were found to be significant.

Table 2

Experiment 1: Predictive Ability of Explanation Quality for Unexplained Items

Predictor	Estimate	SE	95% CI		df	<i>t</i>	<i>p</i>
			Lower	Upper			
Avg % Complete	0.018	0.010	-0.002	0.039	63.73	1.81	.075
Avg Big Parts Inclusion	-0.035	0.113	-0.260	0.190	62.76	-0.31	.758
Avg Small Parts Inclusion	0.148	0.132	-0.115	0.411	62.93	1.12	.265
Avg Causal Links	0.016	0.055	-0.094	0.126	63.03	0.29	.772

Discussion

My data aligns with recent data suggesting that carryover does happen in an IOED paradigm and extends these findings by confirming that the decrease in understanding ratings for unexplained devices is indeed larger than the decrease for explained devices.

In relation to explanation quality, my results for explained devices partially replicated what was seen in Wilson and Marsh (2023) in that lower ratings of perceived percent completeness and less causal links within an explanation were predictive of a greater decrease in understanding ratings from T1 to T2. Because Big Parts Inclusion almost reached significance for this experiment ($p = .080$), it is hard to determine whether the difference in procedure (mainly asking participants to make secondary ratings for unexplained devices) or the increase in sample size (69 vs. 50) caused this discrepancy. In addition, none of the aspects of explanation quality previously found to be significant for explained items were found to influence the change in understanding ratings for unexplained items. However, because the results of this experiment did not fully replicate the results from Wilson and Marsh (2023), I interpret these results with caution.

The results of this study suggest a powerful transfer of knowledge reassessment to unexplained items in the moment. My next experiment addresses whether this knowledge reassessment is retained over time.

Experiment 2

The second experiment was designed to assess whether the knowledge reassessment that occurs after an IOED paradigm is retained over time. I ran a two-session study that included an IOED task in both sessions with the knowledge that comparing the Session 1 T2 ratings and the Session 2 T1 ratings would allow me to determine whether the broken IOED for explained devices in Session 1 remained broken at Session 2.

Additionally, during the second session's IOED paradigm, I had participants explain some of the same devices they explained in Session 1. This allowed me to investigate whether the broken illusion, if it lasted over time, could be broken even further.

Method

Participants

Sample size was determined using the same method outlined in Experiment 1, however, to account for both the increase in range of items being asked about and the likely difficulty of retaining participants for both experimental sessions, I oversampled and recruited 143 participants. All recruitment practices were the same as Experiment 1 with the additional restriction that participants who were part of Experiment 1 were not allowed to participate in Experiment 2.

Of the 143 participants, those who did not complete both sessions were excluded from analysis ($n = 10$). Of the remaining 133 (age $M = 18.44$, range 18 – 22), most identified as female (80%; male = 19%; nonbinary = 1%; preferred not to respond = 1%), as White (63%; African American = 8%; American Indian or Alaska Native = 2%; Asian = 23%; Native Hawaiian or Other Pacific Islander = 2%; preferred to self-describe = 4%; preferred not to respond = 5%), and not Hispanic (82%; Hispanic = 17%; preferred not to respond = 1%).

Materials

Both sessions in Experiment 2 used the same twelve devices used for Experiment 1. Session 2 also included six additional devices, one taken from the list used by Rozenblit and Keil (2002; fireplace), one taken from Lawson's (2006) IOED experiment (bicycle), and four other common, household devices (washing machine, humidifier, air conditioner, and vacuum). All items used in Experiment 2 can be found in Appendix A: Table 1A.

All participants were given the same set of 12 items for Session 1 and 18 items for Session 2. Participants were randomly assigned to one of four versions that counterbalanced which items were explained for both sessions. Table 1A shows one counterbalancing order for

Sessions 1 and 2. The other counterbalancing orders explained the opposite items in one or both sessions. Table 3 provides the abbreviations used for the items based on their role in both Session 1 and Session 2, and how each device type will be distinguished moving forward.

Table 3

Experiment 2 Device Roles in Both Sessions

Session 1	Session 2	Abbreviation/Label
Explained	Explained	Exp-Exp
Explained	Unexplained	Exp-UE
Unexplained	Explained	UE-Exp
Unexplained	Unexplained	UE-UE
Not Tested	Explained	NT-Exp
Not Tested	Unexplained	NT-UE

Measures

Experiment 2 used the same measures as Experiment 1 with the following additions. For the ratings at different time points, it was specified to participants that they should be making a rating for their understanding in that moment, not trying to recall what they reported earlier. Qualitative questions were also edited to provide insight into how participants determined understanding ratings for each session. See Appendix C (Experiment 2) and D for specific wording.

In addition, participants were asked to report if they looked up the devices between session 1 and 2. The question stated: “When you came in a couple of weeks ago to the lab, you made ratings on some of these devices. Some people get curious and look up how these things work. ‘Looking up’ could include watching a YouTube video about how it works, reading a website, talking to a family member or friend, talking to an expert about how it works, etc. Since the last time you were in lab, did you look up any of these devices? (Please be as honest as possible. Your response will not affect your credit for this study in any way.)” Participants were

then presented with a list of the devices seen in Session 1 (both explained and unexplained) with the following answer choices next to each device: “Yes,” “No,” and “Not this specifically, but something similar.” Lastly, participants were presented with the prompt: “If you selected “Not this specifically, but something similar,” for any of the items above, please use the text box to indicate the specific item that you DID look up.” under which there was an open text box to record their response.

Procedure

Participants began Session 1 and completed the same basic procedure as in Experiment 1 except (1) participants were not asked to re-rate unexplained items (to mimic the basic IOED procedure) and (2) lookup questions, qualitative questions, and demographic questions were postponed to Session 2. Additionally, participants completed a question that generated a unique code for matching to Session 2.

Participants returned 7 to 10 days ($M = 8$ days) later to complete Session 2. In Session 2, participants did the same tasks as Session 1 with the addition of six devices (see Table 1A; addition of 3 explained devices and 3 unexplained devices). As in Experiment 1, participants made ratings before and after explanation for all items. After all completeness ratings were finished, participants were asked the following post-test measures: questions about whether they looked up devices in between Session 1 and Session 2 and whether they looked up devices before the beginning of the experiment, qualitative questions about understanding ratings, demographic measures, unique code question, and screening questions. Before leaving, participants were debriefed on the purpose of the study and offered a copy of the consent form.

Results and Discussion

Because of the length and complication of analyses, I will begin this section with a summary of what I found. Session 1 successfully replicated the standard IOED and the persistence of the knowledge reassessment was retained, although not at 100%, after the one-week delay period. There was also evidence of the carryover of this persistence to unexplained items and an additional decrease in understanding ratings in Session 2 for some device groups. Explanation quality was found to predict the amount of persistence over time. I will first present data cleaning and then each of these results in-turn.

Data cleaning

Device-specific ratings for participants were removed from analysis if they chose to skip writing an explanation for that device. This occurred for one device of one participant in Session 1, leading to the exclusion of two device ratings over the 2394 total device ratings from Session 1 (133 participants x 12 T1 ratings + 133 participants x 6 T2 ratings). This also occurred for three devices of one participant in Session 2, leading to the exclusion of six device ratings over the 4788 total device ratings from Session 2 (133 participants x 18 T1 ratings + 133 participants x 18 T2 ratings).

As in Experiment 1, device-specific data for participants who claimed they had looked up that device “Within the last week” before beginning the study were removed from analysis for both Session 1 and Session 2. This led to the additional exclusion of six device ratings from a total of three participants for both Session 1 and Session 2. In addition, device-specific data for participants who claimed they looked up that device between sessions were also removed from analysis for Session 2 only. This amounted to the exclusion of an additional 90 device ratings from a total of 22 participants for Session 2.

IOED Analyses

For ease of explanation, the four timepoints during which participants gave understanding ratings – Session 1 T1, Session 1 T2, Session 2 T1, and Session 2 T2 – will be referred to as T1, T2, T3, and T4, respectively, in the following analyses.

All IOED analyses were performed using the LMM approach outlined in Experiment 1, including participant and device as random intercepts and time as a repeated measure. CS covariance structure was attempted first for all models and then followed up with UN. The structure chosen based on AIC differences and/or statistical testing. VC was used for estimation of random intercepts and slopes. The ANOVA-style output was used to determine the presence of significant main effect(s) and/or interaction effects. As in Experiment 1, significant interactions were followed up with Sidak-corrected comparisons.

Figure 3 shows all understanding ratings for each item type at each timepoint during Experiment 2. The following sections include figures to highlight specific comparisons of interest.

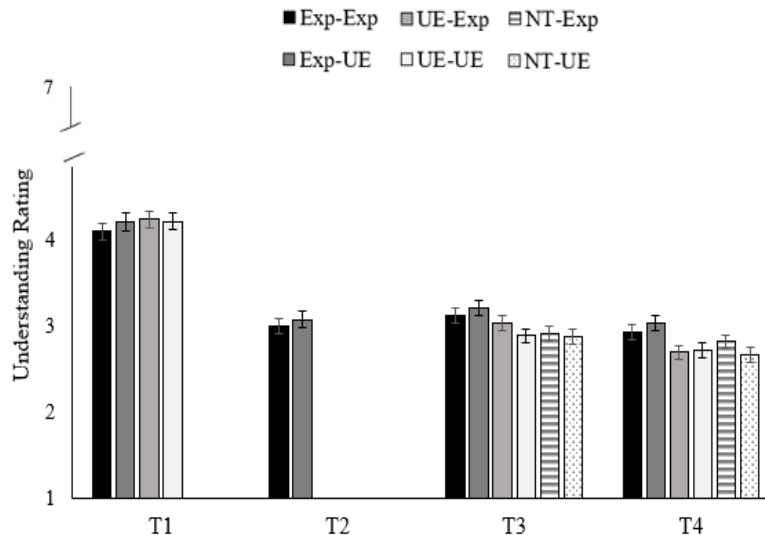


Figure 3: Understanding ratings for all item types at all timepoints in Experiment 2. The y-axis has been split to include the full scale while better displaying the results.

Session 1 IOED. To confirm that the participants' illusion of understanding was successfully broken in Session 1, I ran a LMM with Time (T1 vs. T2) as the independent variable and understanding rating as the dependent variable. Only data for explained devices in Session 1 was used in this analysis. The UN model was chosen as the best fitting model.

As in Experiment 1, there was a significant main effect of Time, $F(1, 794.2) = 344.27$, $p < .001$, such that understanding ratings at T2 were found to be significantly less than at T1 for explained devices (Figure 4A). Again, these results are consistent with the IOED literature in that it shows that participants' illusions of understanding were successfully broken for explained items in Session 1.

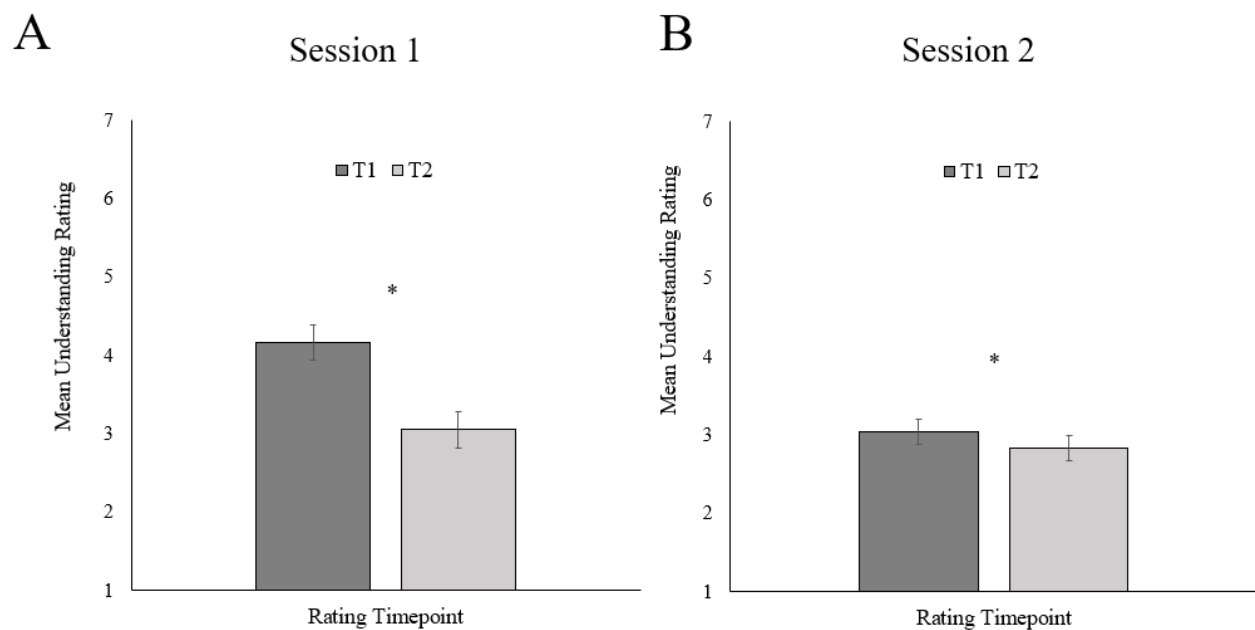


Figure 4: Average understanding ratings at T1 and T2 for explained devices at (A) Session 1 and (B) Session 2 of Experiment 2. Error bars indicate standard error. * $p < .001$

Carryover of Break in Illusion Over Time. Now that I have determined that the Session 1 IOED was successful in breaking participants' illusion of knowledge for explained devices, I move to the main goal of this experiment: looking at whether this break in illusion remained consistent one week after the original IOED had occurred (i.e., at the start of Session 2). To investigate this, I used a LMM analysis with Time (T1, T2, T3) as the independent variable and understanding rating as the dependent measure. Sidak-corrected comparisons were used to explore significant effects. Only data for devices explained in Session 1 were used for this analysis. Based on the AIC, UN was determined to be a significant improvement to the model.⁶

There was a significant overall main effect of Time, $F(2, 789.4) = 174.09, p < .001$. As previously seen in the Session 1 IOED results, there was a significant decrease in understanding rating for explained devices from T1 to T2 ($p < .001$; Figure 4 & Figure 5). After one week of time passed from the initial IOED, understanding ratings for previously explained devices increased significantly from T2 to T3 ($p = .010$). However, T3 ratings did not increase to the same level as initial T1 ratings from before any explanation was generated, with T3 ratings being significantly lower than T1 ratings ($p < .001$; Figure 5). These results show that while there was some loss in the degree of knowledge reassessment over the one-week period, the broken illusion

⁶ Determining model fit for this analysis was especially important because it had an effect on the significance of the data – specifically the difference in understanding ratings between T2 and T3 after Sidak-corrected follow-up tests. Using the CS model, the difference in understanding ratings from T2 to T3 was not found to be significant, $p = .067$, while the same comparison was significant using the UN model ($p = .010$). All other time comparisons did not change significance with change in covariance model type. The models were re-run using the maximum likelihood (ML) estimation method so that a chi-square ratio test could be applied to the change in deviance between the models reported by minus twice the log-likelihood (-2LL) to determine if UN covariance type leads to a significant improvement to the model (Field, 2018, pp. 1199-1200). Run using the ML estimation method, the -2LL for the CS model was 8294.98 with 7 degrees of freedom (df), while the -2LL for the UN model was 8122.79 with 11 df. The -2LL difference was 172.19 while the df difference was -4. When comparing to the chi-square distribution table, at a significance level of .05, the critical value for $df = 4$ is 9.488. Therefore, the UN covariance structure does significantly improve the model.

from the original IOED did not make a full recovery (i.e., a significant portion of the initial knowledge reassessment does persist over a period of one week).

Previous work in metacognition suggests that metacognitive judgements can be retained over time (Barenberg & Dutke, 2019) and knowledge can be transferred to new contexts over time (Georghiades et al., 2006). Taken together, this information led me to predict that the knowledge reassessment induced by generating an explanation in Session 1 would be retained over the week between the two sessions. Alternatively, if the illusion returned in some capacity ratings would be expected to rise at T3. This result would suggest that the IOED paradigm is not necessarily breaking an illusion of knowledge, but that ratings may be regressing toward the mean (Krueger & Mueller, 2002). My results provide support for my initial prediction. Ratings were significantly lower at T3 than T1, suggesting understanding ratings did not return to original levels. However, ratings at T3 were significantly higher than T2, suggesting that the broken illusion does return after a period of one week, even if only slightly.

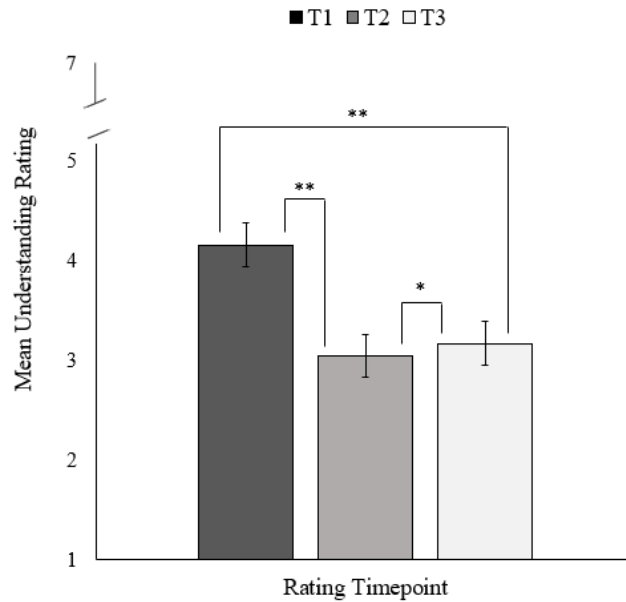


Figure 5: Understanding ratings for explained devices in Session 1 of Experiment 2. T1 was the initial rating of understanding before any type of explanation took place. T2 occurred after the devices had been explained (i.e., after the IOED paradigm). T3 occurred one week after the IOED paradigm had occurred. Error bars indicate standard error and the y-axis has been split to include the full scale while better displaying the results. * $p = .010$; ** $p < .001$

Session 2 IOED. Next, I tested whether the illusion of knowledge could be further broken in a secondary IOED session. Because there were three types of items asked to be explained in Session 2 – items previously explained in Session 1 (Exp-Exp), items previously rated but not explained in Session 1 (UE-Exp), and items not seen at all in Session 1 (NT-Exp) – rating differences among item types were also explored. I used the LMM discussed above with Time (T3 vs. T4) and Item Type (Exp-Exp, UE-Exp, vs. NT-Exp) as the independent variables and understanding rating as the dependent variable. Only data for devices explained in Session 2 were used in this analysis. There was no improvement in the AIC for UN, so CS was used. There

was a significant main effect of Time, $F(1,1159) = 31.02, p < .001$, such that understanding ratings at T4 were found to be significantly less than at T3 (Figure 4B). These results confirm that there was an additional break in participants' illusion of knowledge during Session 2. There was no main effect for Item Type ($p = .081$), however, a significant interaction was seen between Time and Item type, $F(2, 1157) = 3.56, p = .029$.

Follow-up tests focusing on the time course for each individual item type showed that both items that had been seen in Session 1 (whether previously explained or not) had a significant decrease in understanding ratings from T3 to T4 ($ps < .002$). However, the items new to Session 2 that were asked to be explained did not significantly decrease in understanding ratings during Session 2 ($p = .153$; Figure 6). To further compare these results, effect sizes were calculated using paired t-tests⁷ and are shown in Table 4. Previously explained and unexplained devices both had larger effect sizes than items that were not previously tested. One explanation for this finding is that knowledge reassessment over time is slightly over-generalized to completely new items (similar to what was seen for unexplained items in Experiment 1) to the point that understanding ratings reach a floor effect at T3 and are unable to decrease further at T4.

Additional follow-up tests showed that there was no significant difference between the three item types at T3 ($ps > .573$; Figure 6). This confirms that initial ratings for Session 2 were the same regardless of whether the device had been explained before. Additionally, the lack of difference across item types suggests that the degree of knowledge reassessment that persisted over time from Session 1 is not only retained for previously explained devices, but also generalizes to items within the same domain that have never been seen before. At T4, UE-Exp

⁷ While this type of comparison does not account for device differences like LMM analyses, it can give effect sizes for general comparison.

items were rated as significantly less understood than Exp-Exp items ($p = .017$), such that ratings for items explained for the second time in Session 2 were higher than items previously rated in Session 1 but actually explained in Session 2 (Figure 6). There was no difference at T4 for NT-Exp and Exp-Exp ($p = .970$) and NT-Exp and UE-Exp ($p = .968$).

Table 4

Experiment 2: IOED Paradigm Effect Sizes for Items Types Explained in Session 2

Item Type – Session 1	T3		T4		p	Cohen's d
	M	SD	M	SD		
Explained	3.13	1.67	2.93	1.68	< .001	.18
Unexplained	3.04	1.74	2.70	1.60	< .001	.25
Not Tested	2.91	1.69	2.82	1.62	.175	.07

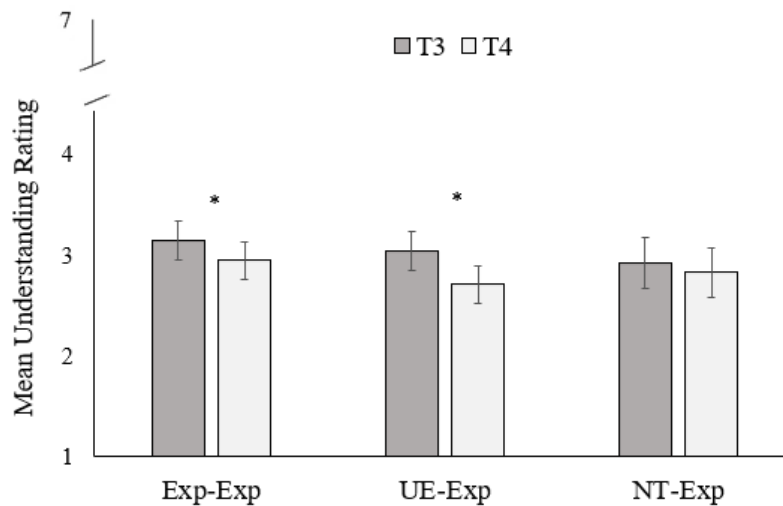


Figure 6: Understanding ratings for explained devices in Session 2 of Experiment 2 with item types separated out. Error bars indicate standard error and the y-axis has been split to include the full scale while better displaying the results. * $p < .002$

Explanation Quality Analyses

Explanation Coding. To examine the objective quality of participants' explanations, all explanations from both sessions were coded separately by different research assistants, who then met to settle disagreements, as in Experiment 1. The few unresolved disagreements were determined by a third party.

Explanation Quality and Explained Items in Session 1. To examine whether explanation quality influences understanding ratings for explained items in Session 1, I conducted the same LMM analyses as outlined in the Explanation Quality and Explained Items section of Experiment 1. Change score for this analysis was calculated by subtracting T1 ratings from T2 ratings for explained devices only in Session 1. All values were specific to explained items from Session 1 only. When the initial model was run, including device as a random intercept was determined not be to significantly improving the model (Wald statistic, $z = 1.69$, $p = .091$), so the model was run again without device as a random intercept.

The resulting LMM can be seen in Table 5 under "Base Model" for predictors of relevance. Within the model, all 4 predictors were found to be significant. Random slopes for each of the four significant predictor variables were also added to the model in an attempt to see if they would improve it. Each random slope addition significantly improved the model based on the Wald statistic ($ps < .001$). Therefore, this was taken as a better-fitting model. Results for the significance of relevant predictors in the updated version of the model can be found in Table 5 under "Random Slopes Model".

Of the four predictor variables, two were found to be significant: Percent Complete and Big Parts Inclusion. The model was re-run including only predictors and their relevant averages

with the rest of the model kept the same (including the average T1 and T2 ratings). Both predictor variables were found to be significant in the new “Final Model” (Table 5).

Figure 7 provides a visualization of how the Percent Completeness and Big Parts Inclusion related to change in understanding scores for explained items of Experiment 2 Session 1. As done in the analyses for Experiment 1, the LMM results were used to calculate the degree of change in understanding ratings predicted for a participant whose values for each predictor were at one standard deviation below and above each predictor mean. Lower points on the y-axis of the graph show a greater decrease in understanding post-explanation. As shown in Figure 7, both ratings positively predict the magnitude of difference in understanding ratings. Specifically, participants with lower Percent Complete and Big Part Inclusion ratings had a larger decrease in understanding ratings post-explanation.

Based on previous data, I had predicted that percent completeness, completeness of big details, and number of causal connections would all be significant positive predictors of change score for explained items for Session 1, while completeness of small details will not have a significant effect. The results are partially consistent with my prediction in that ratings of perceived completeness and perceived inclusion of big details were found to be significant predictors of change score. However, the number of causal links in the explanations was not found to be significant for explained items in Session 1.

Table 5*Experiment 2 Session 1: Predictive Ability of Explanation Quality for Explained Items*

Base Model							
Predictor	Estimate	SE	95% CI		df	<i>t</i>	<i>p</i>
			Lower	Upper			
% Complete	0.041	0.003	0.014	0.038	1501.0	14.2	< .001
Big Parts Inclusion	0.112	0.035	-0.015	0.260	1461.3	3.24	.001
Small Parts Inclusion	0.085	0.031	-0.029	0.245	1460.3	2.76	.006
Causal Links	0.030	0.013	0.032	0.113	1541.4	2.30	.021
Random Slopes Model							
% Complete	0.047	0.006	0.035	0.058	105.12	8.05	< .001
Big Parts Inclusion	0.145	0.064	0.019	0.271	100.44	2.29	.024
Small Parts Inclusion	0.029	0.086	-0.144	0.202	68.183	0.34	.738
Causal Links	0.036	0.022	-0.007	0.078	99.599	1.65	.102
Final Model							
% Complete	0.043	0.004	0.034	0.051	116.69	9.65	< .001
Big Parts Inclusion	0.145	0.051	0.044	0.246	116.22	2.84	.005

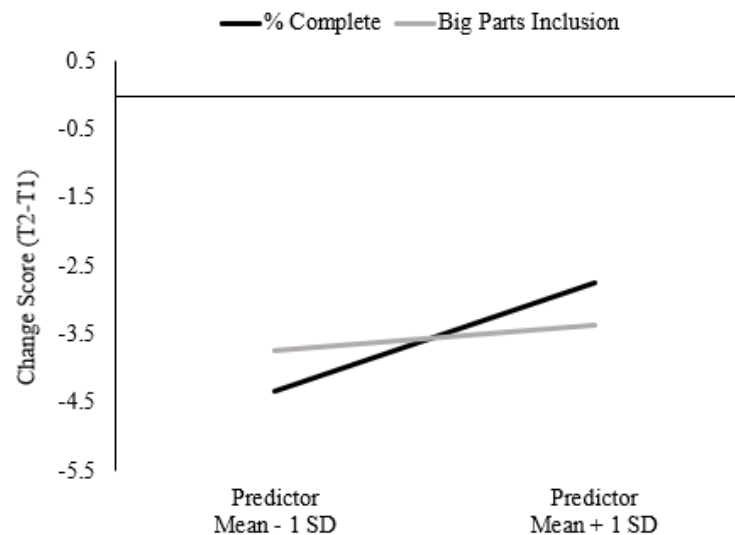


Figure 7: The effect of significant predictors on the difference in understanding ratings from T1 to T2 for explained devices in Experiment 2 Session 1.

Effect of Explanation Quality on Knowledge Reassessment Over Time. To examine whether the quality of the explanations for explained devices in Session 1 had an effect on the degree of retention of knowledge reassessment over time, I had proposed to run a LMM similar to the one above, but with the change score being calculated by subtracting T1 ratings from T3 ratings. The reasoning for this analysis was that, if knowledge reassessment was fully retained over time, then the T2 rating should be the same as the T3 rating. Because results showed a significant increase in understanding ratings from T2 to T3, instead of the hypothesized full retention of the IOED from T2 to T3, a comparison of T2 and T3 ratings (as opposed to T1 vs. T3) was determined to be the more informative analysis for determining whether explanation quality had an effect on whatever transfer over time is seen in Experiment 2.

For the updated analysis, I used the same LMM analysis as proposed (and as the previous analysis) except that change score was calculated by subtracting T3 ratings from T2 ratings. This allows focus on the difference in understanding ratings after the initial break in illusion plus one week of time. Predictor variables remained the same as above except the average of T1 and T2 ratings for explained items per participant became the average of T2 and T3 ratings centered by its grand-mean. Only data for devices explained in Session 1 were used for this analysis and all predictor variables were centered as outlined in the previous analysis before being added to the model.

The resulting LMM can be seen in Table 6 under “Base Model” for predictors of relevance. Within the model, three predictors were found to be significant: Percent Completeness, Big Parts Inclusion and Small Parts Inclusion. Random slopes for each of the four predictor variables were also added to the model to see if they would improve it. Each random slope addition significantly improved the model based on the Wald statistic ($ps < .001$), and

decreased the model's AIC. Therefore, this was taken as the better-fitting model. Results for the significance of relevant predictors in the updated version of the model can be found in Table 6 under "Random Slopes Model". When random slopes were added, only Percent Complete was found to be significant. The model was then re-run including only the one significant predictor and its relevant average with the rest of the model kept the same (including the average T2 and T3 ratings).

Percent Complete was found to be significant in the new "Final Model" (Table 6). The negative estimate shows a negative relationship between Percent Complete and amount of change in understanding rating from T2 to T3. Figure 8 shows a visualization of how Percent Completeness relates to change in understanding scores for explained items of Experiment 2 Session 1 over the one-week delay period. As done in the previous analyses, the LMM results were used to calculate the degree of change in understanding ratings predicted for a participant whose values for each predictor were at one standard deviation below and above the predictor mean. Lower points on the y-axis of the graph show a greater decrease in understanding ratings one-week post-Session 1 (from T2 to T3). As shown in Figure 8, participants with higher Percent Complete ratings were less likely to have an increase in understanding ratings over time.

My predictions as to whether explanation quality would be predictive of retention of the illusion over time were based on my previous results showing that the quality of an explanation has a predictive effect on the amount of knowledge reassessment within an IOED paradigm (Wilson & Marsh, 2023). Specifically, I hypothesized a similar interaction between understanding ratings over time and what was shown for understanding ratings within a typical IOED paradigm - that lower percent completeness, completeness of big details, and number of

causal connections would all be significant positive predictors of the retention of the illusion across sessions (i.e., of T3 ratings being more similar to T2 ratings).

The results of this analysis partially support my prediction, as only percent completeness was found to have a significant relationship with retention of the illusion over time. As a reminder, in a typical IOED paradigm, understanding ratings tend to decrease from T1 to T2. Because change score was calculated by subtracting T1 from T2 in previous analyses, negative scores represented a decrease in ratings across timepoints while the absolute value of the score showed the magnitude of decrease over timepoints (i.e., higher negative scores represented more of a drop while lower negative scores represented less of a drop). A positive relationship was found between completeness and change score, such that participants who perceived their explanations to be more complete had a smaller negative number (i.e., less of a decrease) from T1 to T2. From T2 to T3, understanding ratings tended to increase over time, meaning that a positive change score (calculated by subtracting T2 from T3) would represent an increase in ratings across timepoints. In this analysis, a negative relationship was shown between percent completeness ratings and change score, meaning that participants who perceived their explanations to be more complete has a smaller positive number (i.e., less of an increase) from T2 to T3. Therefore, in both cases, higher perceived completeness led to less change in understanding ratings over time. This suggests that participants who perceived their explanations to be less complete were more susceptible to a decrease in their illusion during the original IOED paradigm and, at the same time, more susceptible to the illusion's return one week later. These results could be due to a floor/ceiling effect for participants with higher completeness ratings – i.e., because there was less of a decrease from T1 to T2, there is less room to increase at T3.

However, T3 ratings did not come close to raising back to the same level as T1 ratings, making this interpretation less likely.

Table 6

Experiment 2: Predictive Ability of Explanation Quality for Illusion Retention Over Time

Base Model							
Predictor	Estimate	SE	95% CI		df	<i>t</i>	<i>p</i>
			Lower	Upper			
% Complete	-0.011	0.002	-0.014	-0.007	3108.5	-5.52	< .001
Big Parts Inclusion	-0.054	0.021	-0.095	-0.014	3011.9	-2.63	.009
Small Parts Inclusion	-0.051	0.018	-0.087	-0.015	3014.3	-2.77	.006
Causal Links	-0.016	0.009	-0.035	0.002	1754.4	-1.73	.084
Random Slopes Model							
% Complete	-0.017	0.008	-0.032	-0.001	92.16	-2.08	.040
Big Parts Inclusion	-0.126	0.096	-0.317	0.065	94.87	-1.31	.194
Small Parts Inclusion	-0.084	0.121	-0.325	0.157	94.03	-0.69	.490
Causal Links	-0.013	0.030	-0.072	0.046	103.66	-0.45	.657
Final Model							
% Complete	-0.015	0.003	-0.021	-0.010	146.73	-5.35	< .001

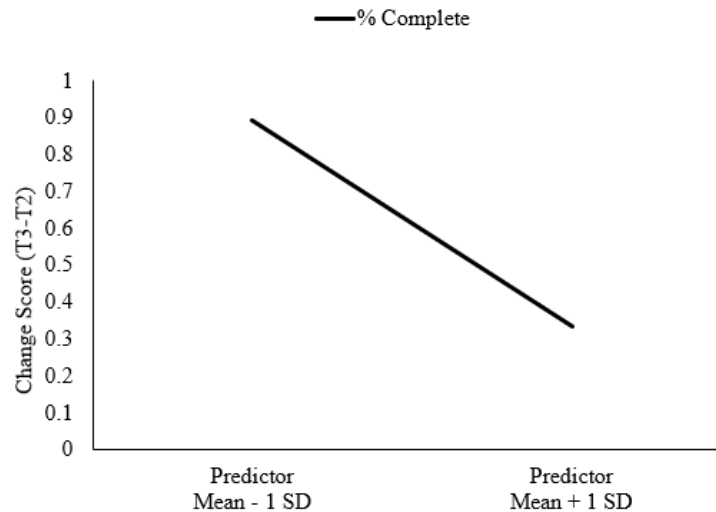


Figure 8: The effect of Percent Complete ratings on the difference in understanding ratings from the end of Session 1 (T2) to the beginning of Session 2 (T3) for explained devices in Experiment 2 Session 1.

General Discussion

Numerous studies have shown self-perceived knowledge to be unreliable (see e.g., Matute et al., 2015; Wilson & Keil, 1998; Rozenblit & Keil, 2002; Zeveney & Marsh, 2016) and causal thinking to be superficial (Sloman et al., 2021; Rabb et al., 2019). While this illusion of understanding has been studied extensively using the IOED paradigm, there are aspects about the boundaries of the IOED that have only been partially explored. Across 2 experiments, I investigated the strength of the IOED by determining its ability to transfer to other items and the persistence of the IOED over time. I also analyzed whether the robustness of the IOED, and the transferability of causal knowledge that it induces, are influenced by the quality of the participants' explanations.

The results of Experiment 1 showed that carryover to unexplained items does occur, replicating the findings of Meyers et al. (2023), but to a significantly larger degree than the decrease in understanding ratings seen for explained items. This suggests that this type of direct metacognitive feedback, as opposed to simply reflecting on one's ability to write an explanation, may be strong enough to induce a domain-overgeneralization. Explanation quality was not shown to have an influence on the transfer to unexplained items, although this question is difficult to investigate statistically and the technique of using average ratings results in the loss of significant variability. Future studies with a larger sample size would allow for a more sophisticated analysis that could account for this variability (as performed in Experiment 2).

In Experiment 2, the reassessment of understanding after the original IOED was shown to be largely retained over the one-week delay period. Understanding ratings after the time delay were significantly higher than the previous ratings, but did not come close to reaching the level of ratings before participants experienced the initial IOED paradigm. These results suggest that

generating a causal explanation can induce longer-lasting changes in metacognitive assessment.

In addition, there was evidence to support the idea that the degree of knowledge reassessment retained over time is transferable to unexplained items in the same domain, again, suggesting that this retained knowledge is domain-general. Lastly, a second IOED paradigm was successfully able to further break whatever illusion returned after the time delay, suggesting that an additional IOED paradigm may be beneficial in subduing any inflated causal knowledge that may return over time.

In relation to the influence of explanation quality on understanding ratings over time, results showed that participants who perceived their explanations to be less complete overall were more likely to increase their understanding ratings after the one-week delay period. Therefore, participants who felt that their explanations were more complete were more likely to retain their metacognitive assessment of knowledge about items in that domain over time.

Overall, these findings strengthen support for the idea of explanations acting as metacognitive judgments to help expose one's misperception of their own causal knowledge and help to integrate the accumulating body of literature on the IOED and explanations with the literature investigating metacognitive beliefs. My results also extend these bodies of literature by providing additional support for the domain-general nature of the knowledge reassessment that occurs during the IOED paradigm and that writing a causal explanation induces a broader update of metacognitive beliefs. In addition, these experiments provide the first evidence that explanation generation has lasting effects on these illusions of knowledge and the metacognitive monitoring system. Lastly, they strengthen previous research showing that perceived completeness of generated explanations is an important component influencing the degree of knowledge reassessment that occurs during the IOED paradigm, and provide evidence that this

component of explanation quality also determines whether the illusion of knowledge returns over time.

Johnson et al. (2016) investigated how asking people to internally reflect on their explanatory ability for certain devices impacted their initial assessment of their metacognitive ability to explain how that device works. They found that only internal reflection was not as powerful in diminishing the IOED as the act of attempting to write out a causal explanation. Preliminary findings from Roeder (2016) found similar results when they asked participants to write out explanations for different devices from which participants had initially rated. Meyers et al. (2023) tested the generalization of knowledge reassessment by using a within-subjects comparison, and found a significant decrease in understanding ratings for both explained and unexplained items. My results support the findings of Meyers et al. (2023), as opposed to Roeder (2016), suggesting a domain-general mechanism for making metacognitive judgments to improve metacognitive accuracy. They also extend this work by providing a window into the magnitude of influence that generating an explanation has on other IOEDs a person may hold. Specifically, that writing out explanations may cause an overgeneralization of this knowledge reassessment to similar beliefs that were not specifically addressed.

This work is the first to address the question of whether explanation generation has only momentary or lasting effects on these illusions of knowledge. Similar to previous findings on metacognitive judgments and the effect of metacognitive feedback over time (e.g. Barenberg & Dutke, 2019; Georgiades et al., 2006), my results support the retention of corrected metacognitive judgments over a period of one week. They suggest that explanation generation has lasting effects on metacognitive judgments, specific not only to previously explained items, but to any items within the same domain. This retention of metacognitive judgments over time

also refutes the idea of regression to the mean as the cause of the IOED, showing that even one-week later, metacognitive ratings remain significantly lower than initial judgments.

Finally, my results provide preliminary evidence of the influence of explanation quality on the retention of corrected metacognitive judgments over time. Previous work in this lab showed a negative correlation between the magnitude of knowledge reassessment that occurs during an IOED paradigm and the completeness of the explanations generated (Wilson & Marsh, 2023). Three aspects of the generated explanations were predictive of the impact of the IOED paradigm on that participant – perceived overall completeness, perceived completeness of important details, and objective completeness (determined by the number of causal links). The more complete the explanations, the less of a decrease in ratings T1 to T2. The results of Experiment 2 furthered this narrative by suggesting that higher levels of perceived completeness cause less recalibration over time, as well. Meaning, people who felt they gave more complete explanations during the IOED paradigm were more steadfast in their ratings over time. Conversely, those with low perceived completeness of their explanations had larger decreases during the IOED paradigm (T1 to T2) and larger increases over time (T2 to T3) – suggesting that more of a shock to their metacognitive ability led to less retention of that shock one week later.

Limitations and Future Directions

One limitation of both of my experiments was that I was unable to completely replicate the explanation quality results seen for explained items in Wilson and Marsh (2023). Because these previous results were the basis for the idea that explanation quality may influence the transfer of knowledge reassessment, whether across to unexplained items (Experiment 1) or over time (Experiment 2), the failure to entirely replicate previous results in these experiments leads me to caution the interpretation of their extensions. Future studies with larger sample sizes would

help to determine the reliability of the influence of different variables related to explanation quality on understanding ratings in the IOED paradigm.

A second limitation of both experiments is the range of complexities of the stimuli used. Johnson et al. (2016; Experiment 5) looked at the impact of item complexity on changes in understanding ratings in the IOED paradigm in both a mental reflection condition (where participants were asked to reflect on their ability to explain the device) and the typical explanation condition. They found that, while device complexity had little effect on the explanation condition, simply reflecting on a complex device (e.g., a vacuum cleaner) led to a significantly greater decrease in understanding ratings than reflecting on a simple device (e.g., Velcro). If the transfer from explained to unexplained items employs a similar mechanism to that of simple reflection, then the complexity of both the explained and unexplained items may impact the degree of transfer witnessed. In future work, I will be exploring the impact of item complexity on transfer to unexplained items of similar, greater, or less complexity.

An obvious follow-up to Experiment 1 has already been explored in the final experiment of Meyers et al. (2023). Specifically, Meyers et al. (2023) investigated knowledge transfer across domains (devices and natural phenomena) from explained to unexplained items and found supportive evidence of transfer. Future work is needed to look at the magnitude of this transfer – i.e., is it overgeneralized like it is within a domain? – and if a similar generalization is seen to other causal domains that are arguably less similar – i.e., mental disorders.

Future studies should also investigate whether the degree of knowledge retention shown in Experiment 2 continues to decrease with larger delay periods between sessions (i.e., 2 weeks, 3 weeks, etc.). While the knowledge reassessment seen after the first IOED paradigm in Experiment 2 was largely retained after the one-week delay period, understanding ratings for

some items did significantly increase. It is possible that, as the delay period is increased, understanding ratings will continue to increase back toward their T1 values. That is, the more time that passes from the original IOED paradigm, the more likely that the original IOED will have returned.

A final avenue for future research would be to determine what other types of knowledge can be recalibrated given a recalibration of explanatory knowledge. Rozenblit and Keil (2002; Experiments 7-10) have shown explanatory knowledge reassessment to be specific to causal knowledge as opposed to other forms of knowledge (i.e., knowledge of facts or procedures). While these other forms of knowledge show the same degree of initial perceived understanding (i.e., similar T1 ratings), the action of providing an explanation seems to be most influential in the induction of knowledge reassessment for causal knowledge. Still to be explored is whether this recalibration for specific causal knowledge can then be transferred to other forms of knowledge that do not require an explanation. For example, if participants' inflated causal knowledge had already been decreased, could this make them recalibrate their knowledge of certain procedures or facts? Is any transfer seen retained over time? These are important questions for understanding the boundaries of knowledge reassessment.

Conclusion

The IOED is a metacognitive error that is disturbingly prevalent in causal reasoning and dangerously influential on decision-making. The experimental results discussed in this manuscript help to assess the strength and persistence of the IOED – providing evidence that writing a causal explanation induces a broader update of metacognitive beliefs that are retained over time. While future work is needed to further pin-point the boundaries of the IOED, these

results offer necessary foundational information on which to build as we work toward overcoming our own ignorance.

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>
- Barenberg, J., & Dutke, S. (2019). Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory*, 27(3), 269–279. <https://doi.org/10.1080/09658211.2018.1506481>
- Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, 36(7), 1178–1203. <https://doi.org/10.1111/j.1551-6709.2012.01262.x>
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, 148(1), 51–64. <https://doi.org/10.1037/xge0000505>
- Fernbach, P. M., & Light, N. (2020). Knowledge is shared. *Psychological Inquiry*, 31(1), 26–28. <https://doi.org/10.1080/1047840X.2020.1722601>
- 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>
- Field, A (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications Limited.

- Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, 144(3), 674–687. <https://doi.org/10.1037/xge0000070>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Gaviria, C., & Corredor, J. A. (2021). Illusion of explanatory depth and social desirability of historical knowledge. *Metacognition and Learning*, 16(3), 801–832. <https://doi.org/10.1007/s11409-021-09267-7>
- Georghiades, P. (2000). Beyond conceptual change learning in science education: Focusing on transfer, durability and metacognition. *Educational Research*, 42(2), 119–139. <https://doi.org/10.1080/001318800363773>
- Georghiades, P. (2006). The role of metacognitive activities in the contextual use of primary pupils' conceptions of science. *Research in Science Education*, 36(1), 29–49. <https://doi.org/10.1007/s11165-004-3954-8>
- Johnson, D., Murphy, M., & Messer, R. (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.supp>
- Keil, F. (2003). Categorisation, causation, and the limits of understanding. *Language and Cognitive Processes*, 18(5–6), 663–692. <https://doi.org/10.1080/01690960344000062>
- Koriat, A., & Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder. *Consciousness and Cognition*, 9, 193–202. <https://doi.org/10.1006/ccog.2000.0436>

- Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2), 180–188.
<https://doi.org/10.1037/0022-3514.82.2.180>
- Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, 34(8), 1667–1675. <https://doi.org/10.3758/BF03195929>
- Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00888>
- Mazor, M., & Fleming, S. M. (2021). The Dunning-Kruger effect revisited. *Nature Human Behaviour*, 5(6), 677–678. <https://doi.org/10.1038/s41562-021-01101-z>
- 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>
- 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>
- Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, 46(2), 168–175.
<https://doi.org/10.1177/0098628319834381>
- Roeder, S. (2016). The disparity between what we know and how we communicate. [Doctoral dissertation, University of California, Berkeley] UC Berkeley Campus eScholarship.
<https://escholarship.org/uc/item/35j6b5dz>

- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562.
- Scharrer, L., Stadtler, M., & Bromme, R. (2014). You'd better ask an expert: Mitigating the comprehensibility effect on laypeople's decisions about science-based knowledge Claims. *Applied Cognitive Psychology*, 28(4), 465–471. <https://doi.org/10.1002/acp.3018>
- Schraw, G. (1996). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education*, 65(2), 135–146. <https://doi.org/10.1080/00220973.1997.9943788>
- Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, 14(4), 332–348. https://doi.org/10.1207/s15327663jcp1404_2
- Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, 7. <https://www.frontiersin.org/article/10.3389/fpsyg.2016.00218>
- Sloman, S. A., Patterson, R., & Barbey, A. K. (2021). Cognitive neuroscience meets the community of knowledge. *Frontiers in Systems Neuroscience*, 15, 1-13. <https://doi.org/10.3389/fnsys.2021.675127>
- Sloman, S. A., & Rabb, N. (2016). Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science*, 27(11), 1451–1460.
- Wilson, J., & Marsh, J. K. (2023). Perceptions of explanation completeness help decrease knowledge overestimation. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society* (pp. 717–723). Cognitive Science Society. <https://escholarship.org/uc/item/9gv611vd>

Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines*, 8, 137–159.

Zeveney, A., & Marsh, J. K. (2016). Illusion of understanding in a misunderstood field: the illusion of explanatory depth in mental disorders. In A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Zheng, M., Marsh, J. K., Nickerson, J. V., & Kleinberg, S. (2020). How causal information affects decisions. *Cognitive Research: Principles and Implications*, 5(1), 6.

<https://doi.org/10.1186/s41235-020-0206-z>

Appendix A: Experimental Stimuli**Table 1A***Device Stimuli for Experiments 1 and 2*

Device	Experiment 1 and Experiment 2 - Session 1		Experiment 2 - Session 2	
	Explained	Unexplained	Explained	Unexplained
Can opener	x		x	
Piano Keys	x		x	
Flush Toilet	x		x	
Zipper	x			x
Spray-bottle	x			x
Freezer	x			x
Ballpoint pen		x	x	
Lock and Key		x	x	
Toaster		x	x	
Printer		x		x
Water Faucet		x		x
Electric Blanket		x		x
Washing Machine			x	
Humidifier			x	
Bicycle			x	
Fireplace				x
Air Conditioner				x
Vacuum				x

Note. All devices took the form of "How a(n) X works," except for the following: "How piano keys make sounds," "How a flush toilet operates," "How a spray-bottle sprays liquids," "How a ballpoint pen writes," "How a key opens a lock," and "How a water faucet controls water flow."

Appendix B: IOED Initial Instructions

In the following questions, you will be asked to rate your understanding of various things on a scale from 1 to 7. Below, you will see examples of a level 7 understanding, a level 4 understanding and a level 1 understanding for a crossbow.

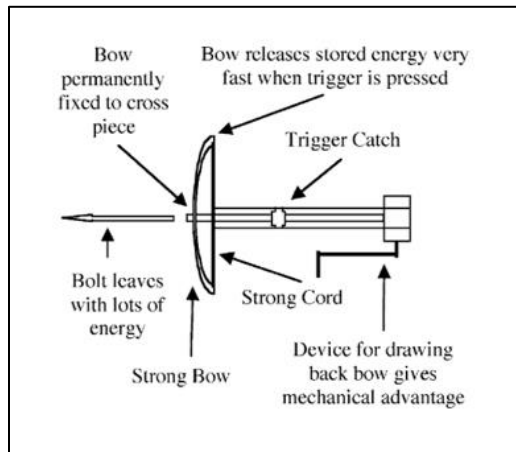
Some people might understand that a crossbow has a stiff, flexible piece of metal as a bow with a wire or strong line; that the bow is permanently mounted on a block of wood or metal; that the wire is pulled back by something that gives a mechanical advantage, either a lever, or small block and tackle, or by a crank wound around a spool that pulls a wire attached to the bow wire. The bow wire is then held back by a pin that is connected to a trigger, and an arrow is set in front of it. Often the pin is forked so the arrow can sit directly in the wire. The pin is directly connected to the trigger so that when you pull on the trigger, it causes it to pivot around a point such that the end that is the pin moves downwards and releases the bow wire. When the pin releases the string, the bow very quickly un-flexes, rapidly imparting all the energy stored in the flexed bow to the arrow. Someone with this level of understanding would give him or herself a rating of 7.

Some people know less detail. For example, someone might know only that the crossbow is a fixed bow and arrow arrangement; that it gets more power than a normal bow and arrow because it allows you to pull the string back extra hard and then trap it there rather than hold it, and that it is then released by a trigger. Someone with this level of understanding would give him or herself a rating of 4.

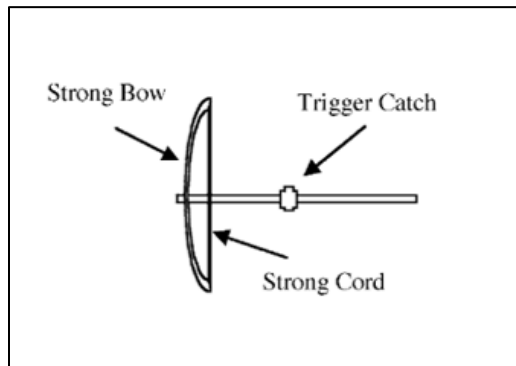
Some people might know even less. For example, someone might really only know what a crossbow looks like and what it does - shoots arrows. Someone with this level of understanding would give him or herself a rating of 1.

Below are diagrams that represent a level 7, level 4 and level 1 understanding of a crossbow.

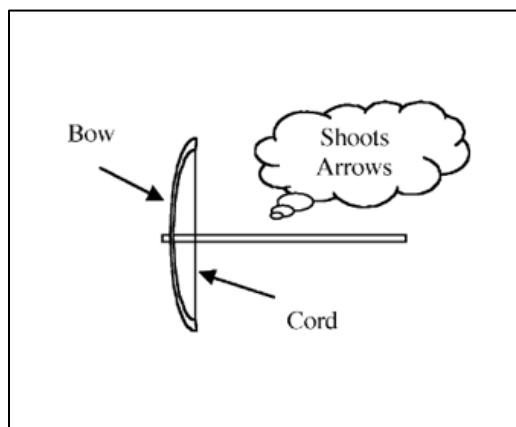
Level 7:



Level 4:



Level 1:



Note that one does not need to be an expert to have level 7 knowledge – an intelligent, educated layperson who has read and understood a good description of the phenomenon in an appropriate reference source probably has level 7 knowledge, as we define it.

We are trying to get a sense of how people feel about their understandings of various phenomena as part of a larger study of how people make sense of the world. We are going to present you with some items; we want you to rate how well you feel you understand each one. Remember, 7 means you have a very thorough understanding of a phenomenon, 1 means you have a very vague understanding of the phenomenon.

It's very important to give us your first impression. We find that taking too long really hurts people's answers. Please go through the list below as quickly as possible and select a number from 1 to 7 on the scale next to each phenomenon, telling us how well you feel you understand each item.

Appendix C: Qualitative Questions about Understanding Ratings

Experiment 1

Question 1

“During this experiment, we asked you to rate your understanding of different devices more than once. Please explain how you estimated your understanding for a particular device in the beginning of the experiment versus later in the experiment. Feel free to use as much detail and you would like.”

Question 2

“During this experiment, we asked you to explain some items but not others. However, you were asked to rate all items a second time. Please explain how you estimated your understanding the second time you rated:

1. A particular device you explained
2. A particular device you did not explain

Feel free to use as much detail as you would like.”

Experiment 2

Question 1

“During this experiment (sessions 1 and 2), we asked you to rate your understanding of different devices more than once. Please explain how you estimated your understanding for a particular device the first time you came into lab (session 1) versus this time you came to the lab (session 2). Feel free to use as much detail and you would like.”

Question 2

“During this lab session, we asked you to write an explanation for some items but not others. However, you were asked to rate your understanding for all items twice. Thinking only about the

current lab session, please explain how you estimated your understanding the second time you rated:

1. The devices you explained
2. The devices you did not explain

Feel free to use as much detail as you would like.”

Appendix D: Updated Rating Instructions for Experiment 2 Session 2

Updated Time 1 Instructions

“For each of the following, please rate your understanding using the 1 to 7 scale that you learned about. Although you may have already rated some of the same items the last time you were in the lab, we are NOT asking you to remember what you put before. Rather, we are asking you to rate how well you feel you understand each item right now.”

Updated Time 2 Instructions for Unexplained Items

“Now, you are going to rate some more items that you rated before. Although you may have already rated some of the same items the last time you were in the lab, we are NOT asking you to remember what you put before. Rather, we are asking you to rate how well you feel you understand each item right now.” This was followed by the typical T2 rating prompt of “Please rate how well you feel you understand X” with “X” being one of the devices that the participant was not asked to explain specifically in Session 2 of Experiment 2.

Updated Instructions for Completeness Ratings

“Think about everything a person could have produced in generating an explanation of X. What percent of the possible information do you think you produced during this lab session?

“Think about what you did produce in your explanation of X. Please select how much you agree or disagree with the following statements based on the completeness of your explanation of X during this lab session.” *with “X” being one of the devices that the participant was asked to write an explanation for in Session 2 of Experiment 2.*

Mechanism Look Up Between Sessions

“When you came in a couple of weeks ago to the lab, you made ratings on some of these devices. Some people get curious and look up how these things work. ‘Looking up’ could

include watching a YouTube video about how it works, reading a website, talking to a family member or friend, talking to an expert about how it works, etc. Since the last time you were in lab, did you look up any of these devices? (Please be as honest as possible. Your response will not affect your credit for this study in any way.)”

Updated Overall Look Up Instructions

“Now try to think back to before the first time you came into lab. Remember: ‘Looking up’ could include watching a YouTube video about how it works, reading a website, talking to a family member or friend, talking to an expert or any other place where you could get information about how a device works. Which of the following best fits the description of how often you have looked up information about the following items before your first session of this study? If you do not remember, please choose your best guess. (Please be as honest as possible. Your response will not affect your credit for this study in any way.)”

Appendix E: Experiment 2 – Further IOED Analyses

Carryover to Unexplained Items in Session 2

Because the results of Experiment 1 showed that the broken IOED transfers to unexplained items, I performed an analysis to explore whether the item carryover seen in Experiment 1 also occurred in a similar fashion during Session 2 of Experiment 2. (This was not done in Session 1 because unexplained items were not rated at T2 to keep it like the classic IOED paradigm.) I performed an LMM analysis with Time (T3 vs. T4) and Item Type from Session 2 (explained vs. unexplained) as factors and understanding ratings as the dependent variable. There was no improvement in the AIC for UN, so CS was used.

A main effect of Time was present, $F(1,2344) = 73.67, p < .001$, where T4 ratings ($M = 2.82, SE = 0.17$) were significantly lower than T3 ratings ($M = 3.01, SE = 0.17$). There was no main effect of Item Type from Session 2, $p = .577$, nor an interaction, $p = .602$. These results confirm the results of Experiment 1 in that the breaking of the participants' illusions of knowledge for explained items successfully transferred to unexplained items. However, unlike the results of Experiment 1, unexplained items were shown to decrease by the same amount as explained items.